Hibbeln, Martin; Gürtler, Marc

Working Paper

Pitfalls in modeling loss given default of bank loans

Working papers // Institut für Finanzwirtschaft, Technische Universität Braunschweig, No. IF35V1

Provided in Cooperation with:
Technische Universität Braunschweig, Institute of Finance

Suggested Citation: Hibbeln, Martin; Gürtler, Marc (2011) : Pitfalls in modeling loss given default of bank loans, Working papers // Institut für Finanzwirtschaft, Technische Universität Braunschweig, No. IF35V1,
http://dx.doi.org/10.2139/ssrn.1757714

This Version is available at:
http://hdl.handle.net/10419/55246

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 data set of 71,463 defaulted 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. 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 LGDs. 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), Acharya et al. (2007), and Bade et al. (2011). 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 of LGDs 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 relevant pitfalls in modeling workout LGDs of bank loans. Moreover, we derive recommendations for action in order to avoid these problems and demonstrate the proposed methods on a data set consisting of 71,463 defaulted loans of a German bank. In the following, we characterize these pitfalls within the typical steps of the modeling process. After collecting all payments during the workout processes of historical defaults, the realized workout LGDs have to be calculated.[1] Within the calculation of LGDs, we observe that the empirical literature on LGDs ignores the effect that samples of historical LGDs are usually biased, which is due to differences in the length of the workout process (pitfall 1). Two types of default end can be distinguished: contracts that can be recovered and contracts that have to be written off. 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.

On the basis of calculated workout LGDs, prediction models for non-defaulted loans can be developed. This is mostly done with a direct regression on LGDs. However, due to the different characteristics of recovered loans and write-offs, the estimation of LGDs with a single model performs poorly (pitfall 2). 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

[1] For retail loans, a default is usually assigned on contract level. Contrary, for corporate loans a default is generally determined on firm level so that several contracts default simultaneously. This has to be considered in the calculation of LGDs, too.

recovered loans as well as the LGD of write-offs is predicted separately. These predictions are combined into the total LGD forecast.

The existing literature on LGD modeling only concentrates on non-defaulted loans. Though, also for defaulted loans with active default status, estimates of LGDs are required, e.g. for regulatory and economic capital calculations. 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 (pitfall 3). 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, to our best knowledge, these pitfalls 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, which have up to now widely been ignored in the LGD literature.

The remainder of this paper is structured as follows. Section 2 contains a description of the data and describes the calculation of LGDs. In this context, we give attention to the first pitfall. In Section 3, we discuss LGD modeling for non-defaulted loans including pitfall 2. Section 4 deals with LGD modeling for defaulted loans, which covers pitfall 3. 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. Let S be a set of loans and $i \in S$ an individual loan. The workout LGD of loan i is typically expressed as follows:

$$LGD_i = 1 - \frac{RCF_i - C_i}{EAD_i}, \tag{1}$$

where RCF_i stands for the sum of discounted recovery cash flows of loan i, C_i represents the sum of discounted direct and indirect costs of loan i, and EAD_i is the exposure at default of

loan *i*. [2] 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.

- Recoveries (RCs): In the case of a recovery, the default reason is no longer existent, e.g. the obligor paid the amount that he was in arrears with payments or a new payment plan has been arranged. Thus, the contract is thenceforward handled as a normal non-defaulted loan.
- Write-offs (WOs): If the chance of recovering additional money from the obligor or the realization of collateral is considered to be small, the contract will be written off. Thus, there are generally no further payments for this contract.

While equation (1) is correct for write-offs, we additionally have to consider the exposure at recovery (EARC) for the case of RCs. At the time of recovery, there is still a significant exposure resulting from installments after the time of recovery. However, since the EARC reduces the economic loss resulting from a default but the EARC is not included in the cash flows, we have to add the (discounted) exposure at recovery $EARC_i$ of loan i to the corresponding (discounted) recovery cash flows:

$$LGD_i = 1 - \frac{RCF_i - C_i + EARC_i}{EAD_i}. \tag{2}$$

If the type of default end is a write-off, we can set the value of $EARC_i$ to zero.

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.[3] The loans correspond to several subportfolios of the bank, which can be divided into private and commercial clients meeting the criteria of retail portfolios,[4] as well as secured and unsecured loans. The description of the data set can be found in Table 1.

- Table 1 about here -

[2] We used the effective interest rate to discount the cash flows since this method has been favored by the national banking supervisor. For details regarding appropriate discount rates see Basel Committee on Banking Supervision (2005a) and Maclachlan (2005).

[3] 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.

[4] See e.g. Basel Committee on Banking Supervision (2005b), §70.

With a total of 59,442 contracts, the major part of the data consists of secured loans to private clients. The LGD frequency distribution corresponding to this subportfolio 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 of secured loans to private clients, 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% and a small peak can be found for an LGD of 100%. 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 LGD distributions of the other subportfolios show some minor differences to Figure 1. For secured loans of commercial clients, the distribution is very similar but the small peak at $LGD = 1$ is missing. This might be a result of higher effort that is made to recover a part of the exposure in connection with a better cost-benefit ratio due to higher loan amounts. If the loans are unsecured, the LGDs are on average significantly higher for both private and commercial clients. However, for all subportfolios there is a large amount of contracts with LGDs close to zero. While these observations mainly consist 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 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 LGDs and the duration of the workout process are not stochastically independent, not only the average duration of the workout process but also average LGD is biased if this effect is ignored. 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.[5] However, we want to determine the LGDs of censored data and not the duration, so that we cannot apply these models. In Proposition 1, we show that the censored data lead to an underestimation of LGDs. Furthermore, we propose to restrict the data set in order to get unbiased results.

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

Proposition 1[6]

Let $i \in S$ be a loan, $\tilde{\tau}_i$ is the point in time of default of loan i, and $\tilde{T}_i$ is the duration of the workout process for loan i.[7] Assume $\tilde{\tau}_i$ to be independent of $\widetilde{LGD}_i$ and of $\tilde{T}_i$. In addition, there exists a barrier $T_{\max}$ with $\tilde{T}_i \leq T_{\max}$. Furthermore, for all $t_1 \geq t_2$ the (conditional) random variable $\widetilde{LGD}_i \mid \tilde{T}_i > t_1$ is assumed to have strict first-order stochastic dominance over $\widetilde{LGD}_i \mid \tilde{T}_i = t_2$. Finally, $\underline{\tau}$ and $\overline{\tau}$ with $\underline{\tau} < \overline{\tau}$ are two points in time with $T_{\max} < \overline{\tau} - \underline{\tau}$. Then the following statements hold:

(I) $\widetilde{LGD}_i$ has strict first-order stochastic dominance over the conditional random variable $\widetilde{LGD}_i \mid \underline{\tau} \leq \tilde{\tau}_i < \tilde{\tau}_i + \tilde{T}_i \leq \overline{\tau}$. Particularly, $E(\widetilde{LGD}_i) > E(\widetilde{LGD}_i \mid \underline{\tau} \leq \tilde{\tau}_i < \tilde{\tau}_i + \tilde{T}_i \leq \overline{\tau})$.

(II) The random variables $\widetilde{LGD}_i$, $\widetilde{LGD}_i \mid \underline{\tau} \leq \tilde{\tau}_i \leq \overline{\tau} - T_{\max}$, and $\widetilde{LGD}_i \mid \underline{\tau} + T_{\max} \leq \tilde{\tau}_i + \tilde{T}_i \leq \overline{\tau}$ are identically distributed, which implies $E(\widetilde{LGD}_i) = E(\widetilde{LGD}_i \mid \underline{\tau} \leq \tilde{\tau}_i \leq \overline{\tau} - T_{\max}) = E(\widetilde{LGD}_i \mid T_{\max} \leq \tilde{\tau}_i + \tilde{T}_i \leq \overline{\tau})$.

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

[6] The proof of the proposition is presented in Appendix A.

[7] Random variables are denoted by a tilde "~".

If we model LGDs on the basis of defaults with completed workout process, the data set consists of observations where the default occurs after the begin of the observation period, i.e. $\tilde{\tau}_i \geq \underline{\tau}$, and the point in time of the default end is $\tilde{\tau}_i + \tilde{T}_i \leq \overline{\tau}$. Thus, an estimation of LGDs on the basis of the complete sample leads to an underestimation of LGDs due to Proposition 1(I). 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.

In order to analyze the relationship between LGDs and default lengths further, we present the length of the default period separately for recovered loans and write-offs. As can be seen in Figure 4, the workout process is typically significantly shorter for loans that can be recovered than for write-offs. Since recoveries usually have significantly smaller LGDs than write-offs, as already demonstrated in Figure 2, we have an essential reason for the finding that defaults with a short default length typically have small LGDs.

- Figure 4 about here -

As can also be seen in Figure 4, almost all workout processes of the presented data are completed after 450 days. Hence, we set $T_{\max}$ = 450 and restrict the data set according to Proposition 1(II). This means that we do not consider all available default data but only those that could have been recovered or written off within 450 days, in order to avoid the systematical underestimation of LGDs. There are two ways of assuring this.

First, we can apply the condition $\underline{\tau} \leq \tilde{\tau}_i \leq \overline{\tau} - T_{\max}$, so that we 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 apply the condition $\underline{\tau} + T_{\max} \leq \tilde{\tau}_i + \tilde{T}_i \leq \overline{\tau}$, so that we 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 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. As a consequence of this restriction, the relative increase of LGD is 8.3%. This is the amount that LGDs would have been underestimated if pitfall 1 has been ignored. Thus, pitfall 1 can indeed lead to a significant bias.

Nevertheless, in existing empirical studies on LGDs 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 few studies report out-of-sample tests of the specified models.[8] This is surprising since it is essential for banks that the models deliver a high accuracy of LGD estimates for unobserved data. 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. Especially, the characteristics that lead to the binary event recovery vs. write-off are often different from the characteristics influencing the LGD within the group of write-offs. For example, it is obvious that the LGD of write-offs is low if the value of collateral is high.

[8] This is also noticed by Bastos (2010).

Contrary, a high value of collateral does not necessarily reduce the *probability* of a write-off. As noticed before, reasons for a recovery can be that the obligor paid the amount that he was in arrears with payments or a new payment plan has been arranged. However, there is no obvious reason that the probability of these events should be influenced by the value of collateral. 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 $i = 1,...,n$, these estimates can be combined into an LGD forecast, which is given by

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

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

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

where $\tilde{1}_{\{\text{WO}\},i}$ is an indicator variable, which equals one if credit i is written-off and zero otherwise. The variables $x_{1,i}$, …, $x_{k,i}$ 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:

$$\widehat{LGD}_{\text{WO},i} = \gamma_0 + \sum_{j=1}^{m} \gamma_j \cdot y_{j,i}, \tag{5}$$

where $y_{1,i}$, ..., $y_{m,i}$ are m different variables, which can also be 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:

$$\widehat{LGD}_{\mathrm{RC},i} = \sum_{j=1}^{N} w_j \cdot LGD_{\mathrm{RC},j},, \tag{6}$$

with $w_j := EAD_j / \sum_{n=1}^{N} EAD_n$. Our methodology is related to the modeling approach of Belotti and Crook (2007). They apply the following two-step approach: In the first step, it is determined whether $LGD = 0$, $LGD = 1$, or $0 < LGD < 1$.[9] 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 *LGD*, the event that a loan can be recovered and the outcome $LGD = 0$ coincide only for a part of the data. Moreover, we did not find different characteristics for defaults with $LGD = 1$. Consequently, we get more reasonable results if the target variable is the type of default end (recovery or write-off).

The predictive power of the model can be evaluated at different stages. First, we evaluate the performance of the logit-model on the basis of the adjusted R^2 and the receiver operating characteristic (ROC). 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 value for the area under the ROC curve is abbreviated as AUC. Second, the linear model is evaluated using the coefficient of determination R^2. Finally, in order to assess the total performance of the model, we combine the predictions of the two-step model according to (3) and compute the R^2 for the combined forecast. However, the statistic expressing the predictive power can be overestimated when calculated in-sample. Against this background, we evaluate the models on the basis of the out-of-sample statistic. The out-of-sample statistic R^2_{OS} is computed as

$$R^2_{\mathrm{OS}} = 1 - \frac{\sum_{i=1}^{M} \left(LGD_i - \widehat{LGD}_i \right)^2}{\sum_{i=1}^{M} \left(LGD_i - \overline{LGD}_{\mathrm{IS}} \right)^2}, \tag{7}$$

where $\overline{LGD}_{\mathrm{IS}}$ is the average LGD of the in-sample data, $\widehat{LGD}_i$ (with $i = 1, \ldots, M$) are the forecasted LGDs calculated out-of-sample (applying the model which is based on the in-sample data), and LGD_{i} are the realized LGDs of the out-of-sample data.[10] This statistic

[9] The authors model recovery rates and not LGDs, but due to LGD = 1 – RR this distinction does not matter.

[10] The out-of-sample R^2 statistic is proposed by Campbell/Thompson (2008) in context of equity premium prediction.

measures the reduction of the mean square prediction error relative to the average LGD of the in-sample data. If $R^2_{\text{OS}} > 0$, the forecasts are better than the in-sample average.

3.2 Comparison of the two-step model and the direct regression by simulation

The following statement reveals that the two-step model is superior to a direct LGD regression. We formulate the statement as a hypothesis that has to be tested since an explicit proof is not possible.

Hypothesis

The out-of-sample coefficient of determination $R^2_{\text{OS, two-step}}$ of the two-step model (formulas (3)-(6)) is higher than $R^2_{\text{OS, direct}}$ of a direct LGD regression.

Test of the Hypothesis by simulation

We analyze the performance of the proposed two-step model in comparison to a direct regression on LGDs on the basis of a simulation study. First, we simulate LGDs for a portfolio of 1000 defaulted loans. When generating LGDs, we use a structure which incorporates differences between write-offs and recovered loans, consistent to our argument and empirical findings. However, we choose a model structure which differs from (4) and (5) to induce some model error. We generate the event of a write off if some observable or unobservable influence factors x_i, y_i ,ε_i lead to an excess of the barrier δ:

$$\tilde{1}_{\{\text{WO}\},i} = 1: \;\; \Phi\left(\sqrt{\rho^2_{x,1}} \cdot \tilde{x}_i + \sqrt{\rho^2_y} \cdot \tilde{y}_i + \sqrt{1-\rho^2_x-\rho^2_y} \cdot \tilde{\varepsilon}_i\right) > \delta, \tag{8}$$

with $\tilde{x}_i, \tilde{y}_i, \tilde{\varepsilon}_i \sim \mathcal{N}(0,1)$ and Φ is the standard normal CDF. Since the argument of Φ is standard normally distributed, the result $\Phi(\cdot)$ is uniformly distributed with $\Phi(\cdot) \sim \mathcal{U}(0,1)$. In our simulation, we set δ = 0.8, leading to a 20% probability of a write-off. Similarly, we generate the LGDs within the group of write-offs by

$$\widetilde{LGD}_{\text{WO},i} = \Phi\left(\sqrt{\rho^2_{x,2}} \cdot \tilde{x}_i + \sqrt{\rho^2_z} \cdot \tilde{z}_i + \sqrt{1-\rho^2_x-\rho^2_z} \cdot \tilde{\xi}_i\right), \tag{9}$$

with $\tilde{x}_i, \tilde{z}_i, \tilde{\xi}_i \sim \mathcal{N}(0,1)$. Thus, the LGD is bound between zero and one. Altogether, the outcome of LGD is calculated as

$$\widetilde{LGD}_i = \tilde{1}_{\{\text{WO}\},i} \cdot \widetilde{LGD}_{\text{WO},i}, \tag{10}$$

which implies that the LGD of recoveries is set to zero.

According to our argument above, the event of a write-off and the LGD within the group of write-offs can be influenced by different variables. However, some variables can be relevant for both equations. Against this background, $\tilde{x}_i$ influences both dependent variables but the coefficients can be different. Contrary, $\tilde{y}_i$ and $\tilde{z}_i$ each affect only one of the dependent variables. Moreover, we assume that $\tilde{x}_i, \tilde{y}_i, \tilde{z}_i$ are observable whereas $\tilde{\varepsilon}_i$ and $\tilde{\xi}_i$ are unobservable random variables. Thus, only $\tilde{x}_i, \tilde{y}_i$, and $\tilde{z}_i$ are input variables for the regressions which are applied subsequently.

In order to compare the performance of both modeling approaches, we perform a direct regression with target variable LGD on the one hand and apply the two-step model on the other hand. As stated above, we combine the predictions of the two-step model according to (3) and compare the out-of-sample R^2 of both modeling approaches with formula (7). For the out-of-sample analysis, we generate 10,000 additional LGDs using formula (8)-(10).[11]

The simulation procedure from above is performed for a broad range of parameter combinations. The coefficients $\rho_{x,1}^2$ and $\rho_{x,2}^2$ are independently set to (0.1, 0.2, …, 0.9) and the coefficients ρ_y^2 and ρ_z^2 are set to (0.1, …, $1-\rho_{x,1}^2$) and (0.1, …, $1-\rho_{x,2}^2$), respectively. This leads to a total number of 1,936 different parameter combinations. For each parameter combination, we repeat the simulation procedure 1,000 times and compare the average in- and out-of-sample R^2 of both models. The mean R_{OS}^2 of the two-step model is 52.2% whereas the mean R_{OS}^2 of the direct regression is only 32.5%, as can be seen in Table 2. Moreover, the difference $\Delta R_{\text{OS}}^2 = R_{\text{OS, two-step}}^2 - R_{\text{OS, direct}}^2$ is positive for each individual parameter combination, which confirms our hypothesis. Thus, the two-step model impressively outperforms the direct regression.

- Table 2 about here -

□

The application of our two-step approach to real data is presented subsequently.

[11] Due to the known LGD generating process, we can create an arbitrary number of LGDs for testing the models out-of-sample. With an increasing number of LGDs the measured predictive power converges towards the true value.

3.3 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 $\hat{\lambda}_{WO}$ 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 (in-sample) and 30% validation data (out-of-sample). 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 on the basis of the receiver operating characteristic (ROC) for the validation data.[12] 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 are $AUC_{Train} = 73.5\%$ and $AUC_{Validate} = 71.3\%$. 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 $AUC_{All} = 73.0\%$. 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 and the collateralization level).[13] Interestingly, for unsecured loans it was possible to develop a model where the explanatory power is significantly higher, with $AUC_{Train} = 81.6\%$ and $AUC_{Validate} = 82.2\%$ (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

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

[13] The publication of the concrete model including the coefficients is prohibited by the bank.

models is mainly evaluated with the coefficient of determination for the validation data R^2_{Validate} applying formula (7). For secured loans to private customers, the coefficients of determination for the selected model are $R^2_{\text{Train}} = 19.9\%$ and $R^2_{\text{Validate}} = 17.6\%$.[14] The final coefficients are calibrated on the complete data set leading to $R^2_{\text{All}} = 19.3\%$. Again, the explanatory variables can be classified into borrower characteristics (e.g. the occupational category for private customers), collateral characteristics (e.g. type and value of collateral), 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.

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.

[14] After transforming the LGD estimates using $\widehat{Loss}_i = (1 - \widehat{LGD}_i) \cdot EAD_i$, it is also possible to evaluate the predictive power with respect to absolute instead of relative losses. This leads to coefficients of determination of 52.23% and 57.27%, respectively.

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.[15] 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 maximized.[16] 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 the workout process. Contrary, when applying the model to active defaults, we only know the current default length, which is obviously smaller than the total length $\tilde{T}$. In Proposition 2, we show that ignoring the difference between the information sets would lead to a significant underestimation of the LGD. Furthermore, we present a consistent estimator using the information of the current default length.

Pitfall 3: Underestimation of LGDs when using the total length of the default period as explanatory variable

Proposition 2[17]

Let the assumptions of Proposition1 be fulfilled and let CDL_i denote the current default length of loan *i*. Furthermore, consider a sequence of loans denoted by j = 1, 2, …, whereby $(\widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\})_{j\in\mathbb{N}}$ is a sequence of independently and identically distributed random variables, each member of the sequence with expectation value $E(\widetilde{LGD}_i \cdot I\{\tilde{T}_i > t\})$.[18] Furthermore, $(I\{\tilde{T}_j > t\})_{j\in\mathbb{N}}$ is a sequence of independently and identically distributed random variables, each member of the sequence with expectation value $E(I\{\tilde{T}_i > t\})$ Finally, the

[15] 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.

[16] For details see Breiman (1984).

[17] The proof of the proposition is presented in Appendix B.

[18] $I\{T_j > t\}$ takes the value one if the argument is true and zero otherwise.

corresponding exposures at default EAD_1, EAD_2, … are assumed to be deterministic and to fulfill the following conditions:

(a) $\sum_{j=1}^{N} EAD_j \underset{N\to\infty}{\to} \infty$, (b), $\sum_{j=1}^{\infty} \frac{Var(\widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\})}{\left(\sum_{k=1}^{j} EAD_k\right)^2} < \infty$, and (c) $\sum_{j=1}^{\infty} \frac{Var(I\{\tilde{T}_j > t\})}{\left(\sum_{k=1}^{j} EAD_k\right)^2} < \infty$.

Then the following statements hold:

(I) $P(\widetilde{LGD}_i \le x \mid CDL_i = t) \le P(\widetilde{LGD}_i \le x \mid \tilde{T}_i = t)$.

(II) $\frac{\sum_{j=1}^{N} EAD_j \cdot \widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\}}{\sum_{j=1}^{N} EAD_j \cdot I\{\tilde{T}_j > t\}} \xrightarrow[N\to\infty]{a.s.} E\left(\widetilde{LGD}_i \mid \tilde{T}_i > t\right)$.

If we model LGDs using the default length as explanatory variable and ignore the different information sets of the default length for the modeling and scoring data, the LGDs are underestimated as shown in Proposition 2(I). However, since the length of the default period has a high explanatory power for LGDs, we intend to use the known information set. The information that the current default length equals t for the scoring data is identical to the information that the total length of the default period T is larger than t. Though, for the modeling data we can calculate the (EAD-weighted) average LGDs for all contracts with $T > t$. If we proceed so for every value of $t \in [0, T_{\max}]$, we can assign LGDs to every defaulted loan using the information of the current default length and, as shown in Proposition 2(II), get consistent LGDs when we apply the model. Since these LGDs are calculated on the basis of modeling data with a minimum default length (MDL) of t, we call the corresponding values $LGD(MDL = t)$. Though, we want to include additional influence factors, i.e. the mentioned segmentation variables and the default reason. Against this background, we first partition our modeling data into classes which are homogeneous regarding these variables and calculate $LGD(MDL = t)$ for every class. Under consideration of $LGD_{\text{Default},i} := E(\widetilde{LGD}_i \mid CDL = t) = E(\widetilde{LGD}_i \mid \tilde{T}_i > t)$ and due to Proposition 2(II), we are able to define an estimator of $LGD_{\text{Default},i}$ as follows:

$$\widehat{LGD}_{\text{Default},i} = \frac{\sum_{j=1}^{N} EAD_j \cdot LGD_j \cdot I\{T_j > t\}}{\sum_{j=1}^{N} EAD_j \cdot I\{T_j > t\}} =: LGD\left(MDL = t\right), \tag{11}$$

where $N \in \mathbb{N}$ and $j = 1, \ldots, N$ stands for all contracts of our modeling data within a class. 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 (11), 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.[19] 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 *MDL* = 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 relevant pitfalls in modeling workout LGDs which can easily lead to inaccurate LGD forecasts. Furthermore, we propose methods how to deal with these pitfalls and apply these methods to a data set of 71,463 defaulted loans of a German bank. First, the LGDs within the modeling data can be significantly biased downwards if all available defaults with completed workout process are considered. This is mainly due to length-biased sampling in connection with a different default length of recovered loans and

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

write-offs. We show how the modeling data could be chosen in order to get unbiased LGD estimates. Second, we propose a two-step approach for modeling LGDs of non-defaulted loans. With this approach, we could achieve better predictions than with other approaches proposed in the literature, since different influencing factors of recoveries and write-offs can be considered. We demonstrate the potential of this approach on the basis of a simulation study and apply the model to the data set. Third, we propose a model to forecast LGDs of defaulted loans on the basis of regression trees. 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 ignored. However, neglecting this influence factor leads to considerable worse predictions. Against this background, we have constructed the variable "minimum default length" for the modeling data, which contains the same information set as the current default length of the scoring data, leading to consistent LGD estimates.

Another interesting finding is that the predictive power for estimating the probability of a recovery or a write-off is higher for unsecured than for secured loans. Contrary, for the predictions of LGDs conditional on the type of default end the opposite is true. However, it would be interesting to verify that this observation is generally valid and not specific to the used data set. Moreover, while we mainly focused on retail loans, our models could also be beneficial for the prediction of LGDs of 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 77(7), 11–23.

Bade, B., Rösch, D., Scheule, H., 2011. Default and recovery risk dependencies in a simple credit risk model. European Financial Management 17, 120–144.

Basel Committee on Banking Supervision, 2005a. Guidance on paragraph 468 of the framework document, Bank for International Settlements.

Basel Committee on Banking Supervision, 2005b. International convergence of capital measurement and capital standards – a revised framework, Bank for International Settlements.

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 Out of 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.

Gordy, M.B., 2003. A Risk-Factor Model Foundation for Rating-Based Capital Rules. Journal of Financial Intermediation 12(3), 199–232.

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. Journal of Economic Literature 26, 649–679.

Maclachlan, I., 2005. Choosing the discount factor for estimating economic LGD. In: Altman, E., Resti, A., Sironi, A. (Eds.), Recovery Risk: The Next Challenge in Credit Risk Management. Risk Books: London.

Petrov, V., 1996. Limit Theorems of Probability Theory: Sequences of Independent Random Variables. Oxford University Press: Clarendon.

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

Frequency distribution of loss given default of secured loans of private clients

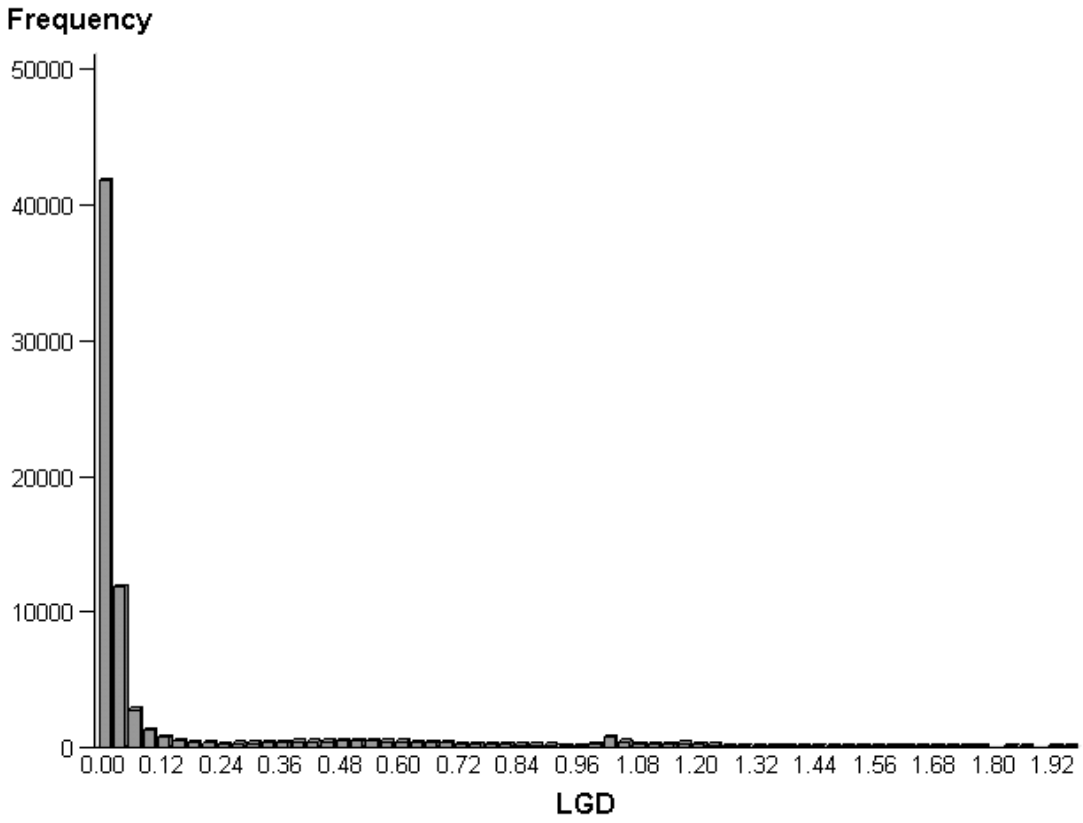
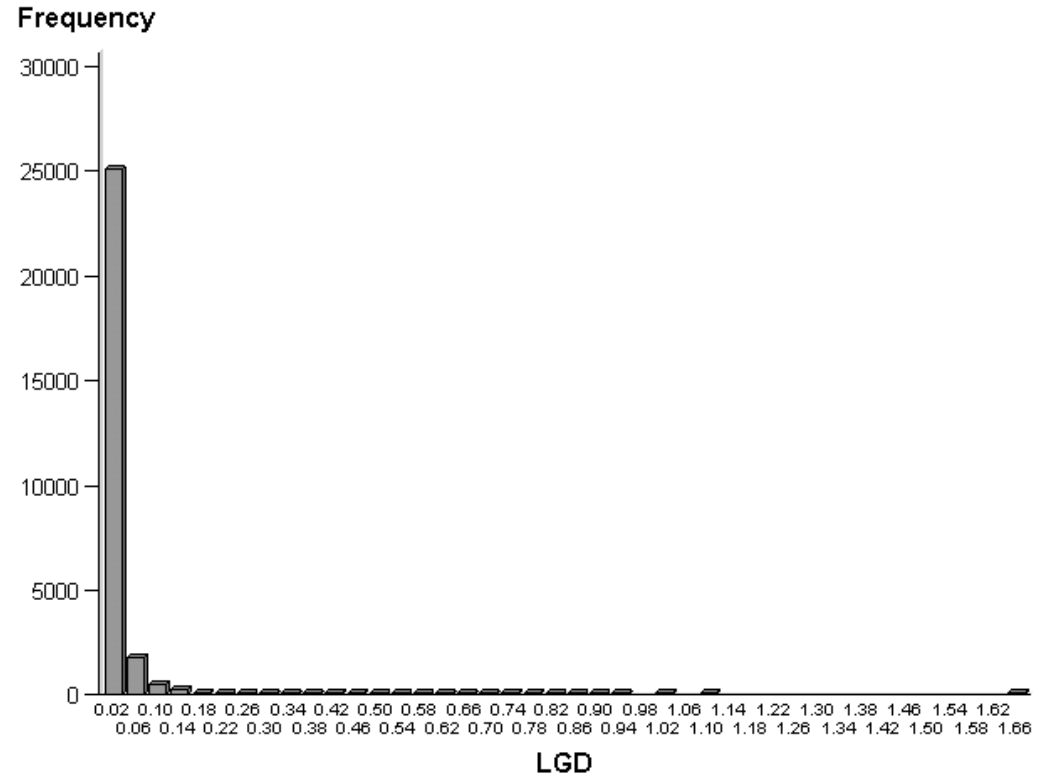


Figure 2

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



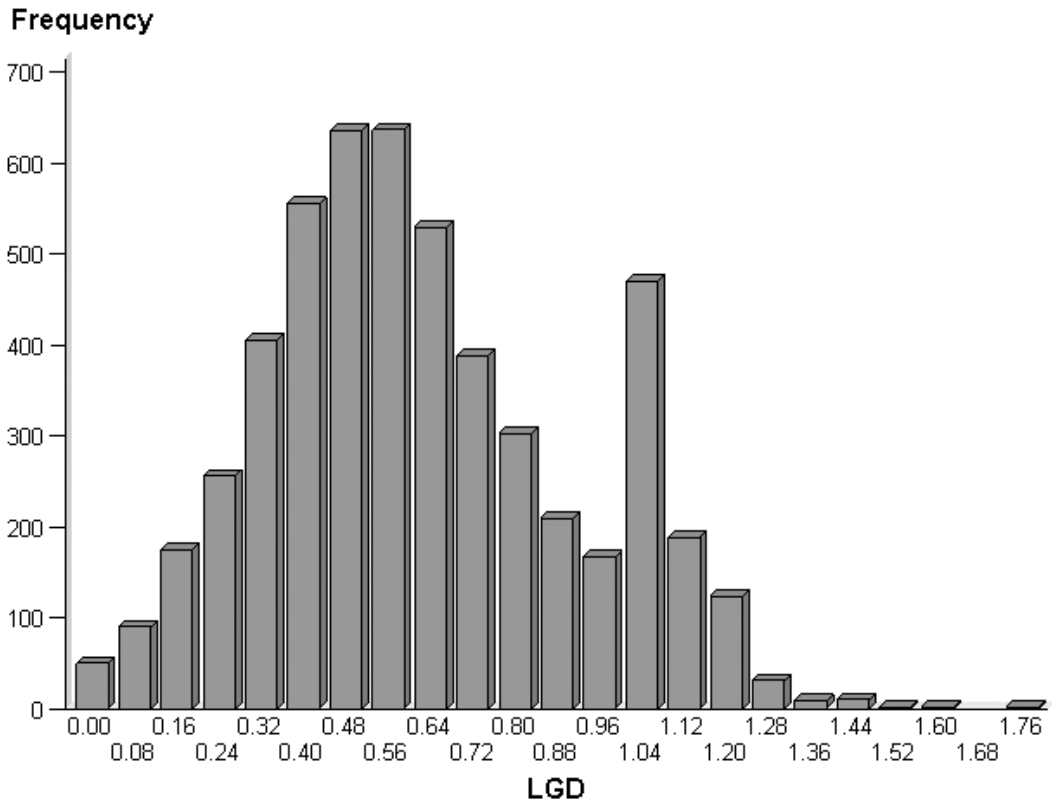


Figure 3

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

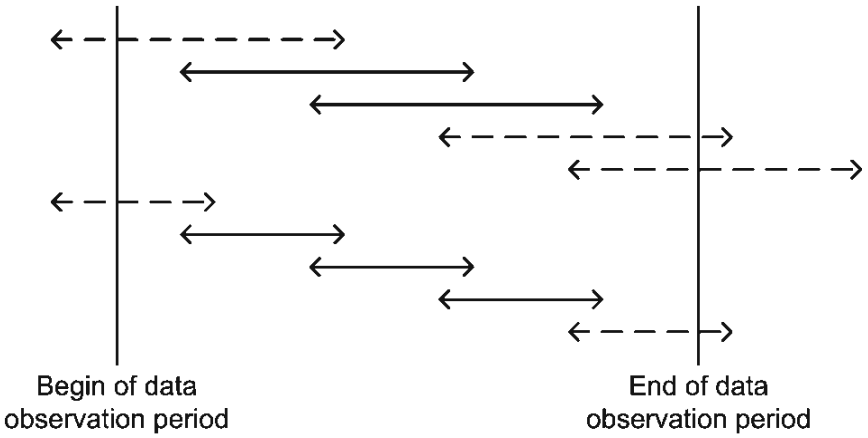
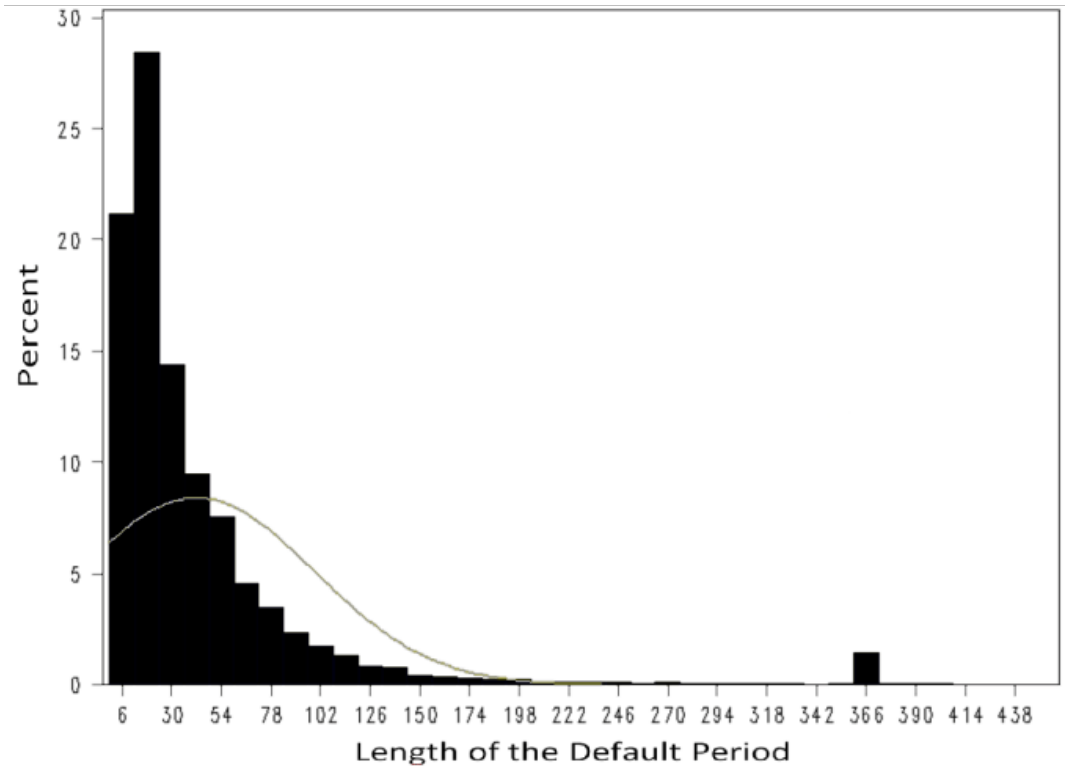


Figure 4

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



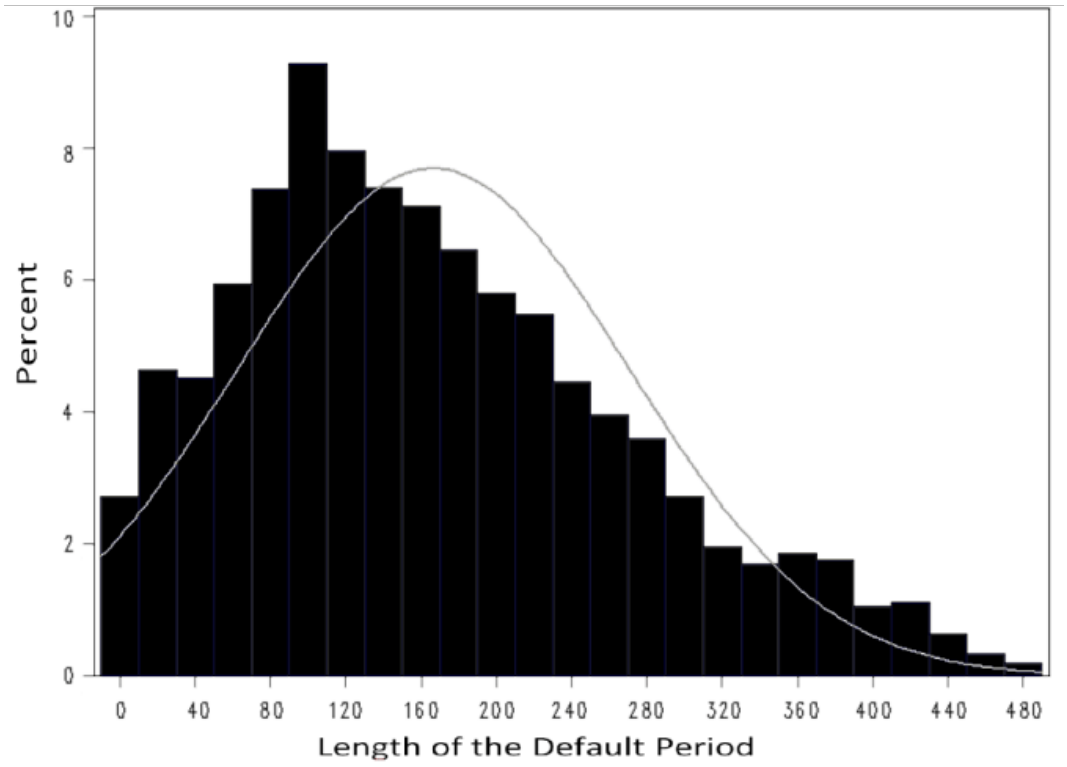


Figure 5

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

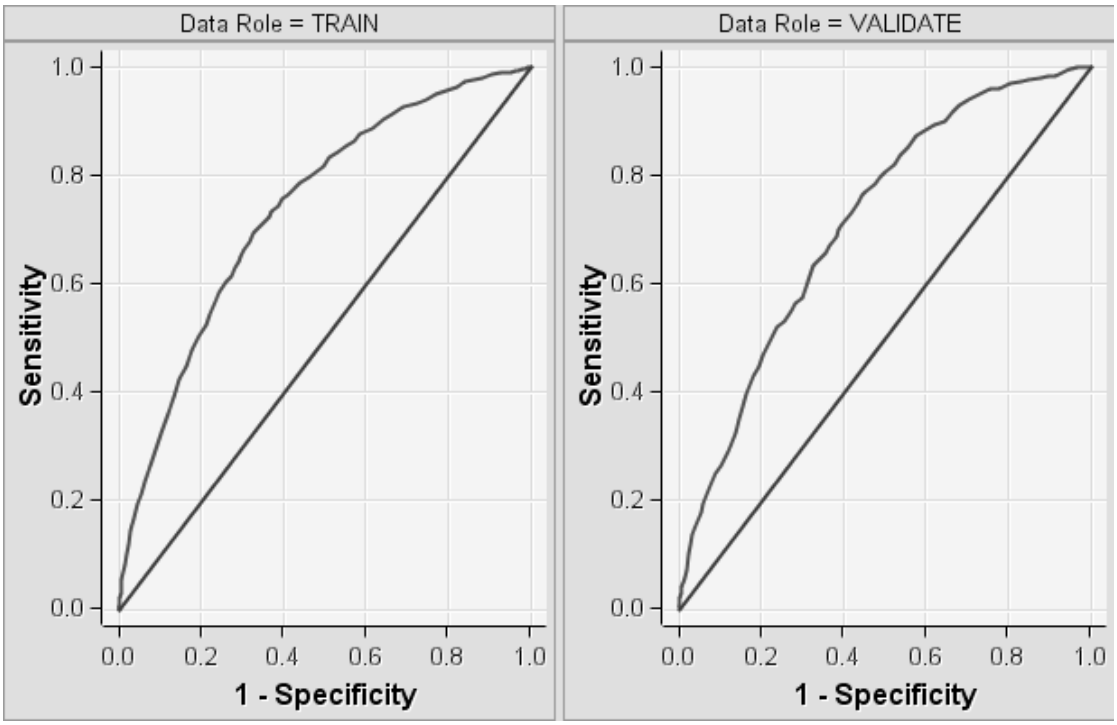


Figure 6

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

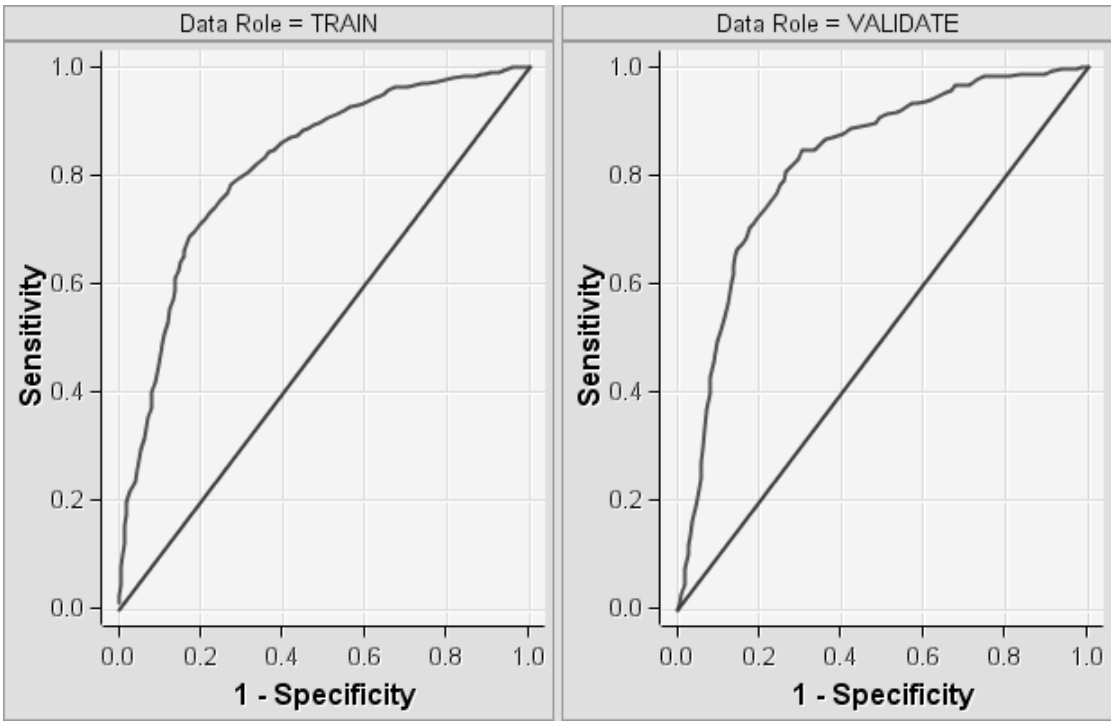


Figure 7

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)

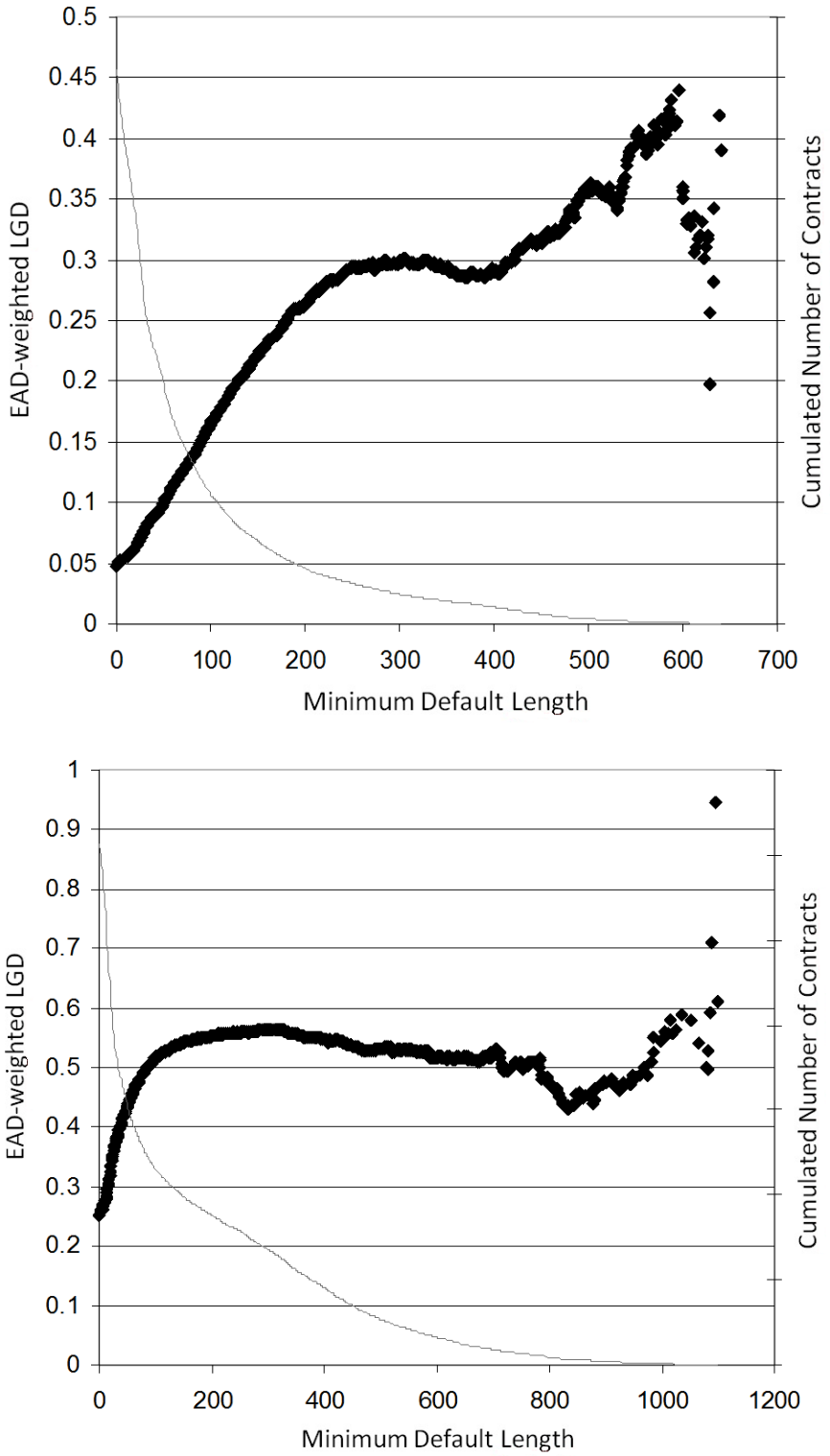


Figure 7 (continued)

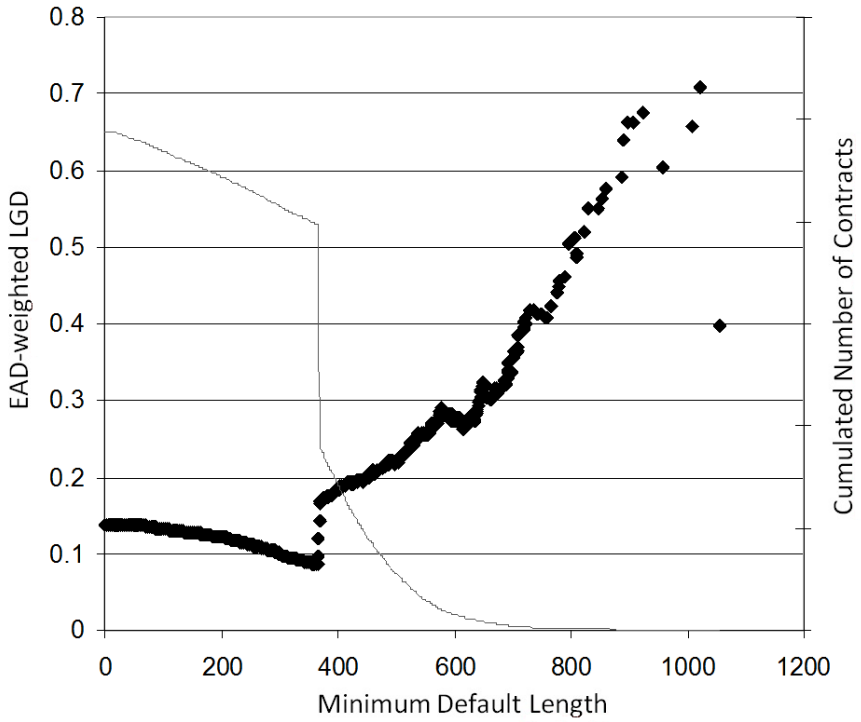


Table 1

Summary statistics

	Number of defaults
Private clients	61,860
Commercial clients	8,125
Secured loans	67,410
Unsecured loans	2,575

	Mean	Std. Dev.	Median
Exposure at default (€)	9,329.34	7,563.85	7,571.52
Collateralization level of secured loans	1.04	1.15	0.68

Table 2

The table shows statistics for the R^2 on the basis of 1,000 simulation runs for each 1,936 different parameter combinations. The in- and out-of-sample R^2 is calculated for the two-step model and the direct regression.

	Obs.	Mean	Std. Dev.	Min.	Max.
$R^2_{\text{IS,two-step}}$	1,936	0.590	0.213	0.122	0.997
$R^2_{\text{IS,direct}}$	1,936	0.346	0.077	0.107	0.506
$R^2_{\text{OS,two-step}}$	1,936	0.584	0.211	0.117	0.991
$R^2_{\text{OS,direct}}$	1,936	0.342	0.078	0.102	0.504
$\Delta R^2_{\text{IS}} = R^2_{\text{IS,two-step}} - R^2_{\text{IS,direct}}$	1,936	0.244	0.168	0.015	0.807
$\Delta R^2_{\text{OS}} = R^2_{\text{OS,two-step}} - R^2_{\text{OS,direct}}$	1,936	0.242	0.166	0.015	0.772

Appendix A. Proof of Proposition 1

Ad (I):

First of all, the random variable $\widetilde{LGD}_i \mid \tilde{T}_i > t$ has strict first-order stochastic dominance over $\widetilde{LGD}_i \mid \tilde{T}_i \leq t$ for all *t* since

$$P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i \leq t) = E_\theta[\underbrace{P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i = \tilde{\theta})}_{>P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i > t)} \mid \tilde{\theta} \leq t] > P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i > t). \tag{12}$$

On this basis we get

$$\begin{aligned} P(\widetilde{LGD}_i \leq x) &= P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i \leq \overline{\tau} - \tilde{\tau}_i) \cdot P(\tilde{T}_i \leq \overline{\tau} - \tilde{\tau}_i) \\ &\quad + P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i > \overline{\tau} - \tilde{\tau}_i) \cdot P(\tilde{T}_i > \overline{\tau} - \tilde{\tau}_i) \\ &\leq P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i \leq \overline{\tau} - \tilde{\tau}_i) \\ &= P(\widetilde{LGD}_i \leq x \mid \underline{\tau} \leq \tilde{\tau}_i \wedge \tilde{T}_i \leq \overline{\tau} - \tilde{\tau}_i). \end{aligned} \tag{13}$$

The inequality results from the statement that $\widetilde{LGD}_i \mid \tilde{T}_i > \overline{\tau} - \tilde{\tau}_i$ strictly dominates $\widetilde{LGD}_i \mid \tilde{T}_i \leq \overline{\tau} - \tilde{\tau}_i$ according to first order stochastic dominance, and the latter equality results from the stochastic independence of $\tilde{\tau}_i$ to $\widetilde{LGD}_i$ and $\tilde{T}_i$.

Ad (II):

Since $\tilde{\tau}_i$ is independent of $\widetilde{LGD}_i$, and $T_{\max} < \overline{\tau} - \underline{\tau}$, it immediately follows that

$$P(\widetilde{LGD}_i \leq x) = P(\widetilde{LGD}_i \leq x \mid \underline{\tau} \leq \tilde{\tau}_i \leq \overline{\tau} - T_{\max}). \tag{14}$$

Furthermore, since $\tilde{\tau}_i$ is additionally independent of $\tilde{T}_i$, and $\tilde{T}_i \leq T_{\max}$, we have

$$\begin{aligned} P(\widetilde{LGD}_i \leq x) &= P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i \leq T_{\max}) \\ &= P(\widetilde{LGD}_i \leq x \mid \tilde{T}_i \leq T_{\max} \wedge \underline{\tau} + T_{\max} - \tilde{T}_i \leq \tilde{\tau}_i \leq \overline{\tau} - \tilde{T}_i). \end{aligned} \tag{15}$$

□

Appendix B. Proof of Proposition 2

Ad (I):

For all t the (conditional) random variable $\widetilde{LGD}_i \mid \tilde{T}_i > t$ is assumed to have strict first-order stochastic dominance over $\widetilde{LGD}_i \mid \tilde{T}_i = t$ (cf. section 2). Thus, it immediately follows:

$$P(\widetilde{LGD}_i \le x \mid CDL_i = t) = P(\widetilde{LGD}_i \le x \mid \tilde{T}_i > t) \le P(\widetilde{LGD}_i \le x \mid \tilde{T}_i = t). \tag{16}$$

Ad (II):

By definition we have

$$E\left(\widetilde{LGD}_i \mid \tilde{T}_i > t\right) = \frac{E\left(\widetilde{LGD}_i \cdot I\{\tilde{T}_i > t\}\right)}{E(I\{\tilde{T}_i > t\})}. \tag{17}$$

Furthermore, under consideration of the assumptions with regard to the sequences $(\widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\})_{j\in\mathbb{N}}$ and $(I\{\tilde{T}_j > t\})_{j\in\mathbb{N}}$, we are able to apply the "strong law" for weighted averages as presented in Petrov (1996), Theorem 6.7,[20] according to which

$$\frac{1}{\sum_{k=1}^{N} EAD_k} \cdot \left(\sum_{j=1}^{N} EAD_j \cdot \widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\} - \sum_{j=1}^{N} EAD_j \cdot E\left(\widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\}\right) \right) \xrightarrow[N\to\infty]{a.s.} 0 \tag{18}$$

and

$$\frac{1}{\sum_{k=1}^{N} EAD_k} \cdot \left(\sum_{j=1}^{N} EAD_j \cdot I\{\tilde{T}_j > t\} - \sum_{j=1}^{N} EAD_j \cdot E\left(I\{\tilde{T}_j > t\}\right) \right) \xrightarrow[N\to\infty]{a.s.} 0. \tag{19}$$

Since $E(\widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\}) = E(\widetilde{LGD}_i \cdot I\{\tilde{T}_i > t\})$ and $E(I\{\tilde{T}_j > t\}) = E(I\{\tilde{T}_i > t\})$ for all j, the almost sure convergences in (18) and (19) lead to

$$\frac{1}{\sum_{k=1}^{N} EAD_k} \cdot \left(\sum_{j=1}^{N} EAD_j \cdot \widetilde{LGD}_j \cdot I\{\tilde{T}_j > t\} \right) \xrightarrow[N\to\infty]{a.s.} E\left(\widetilde{LGD}_i \cdot I\{\tilde{T}_i > t\}\right) \tag{20}$$

and

$$\frac{1}{\sum_{k=1}^{N} EAD_k} \cdot \left(\sum_{j=1}^{N} EAD_j \cdot I\{\tilde{T}_j > t\} \right) \xrightarrow[N\to\infty]{a.s.} E\left(I\{\tilde{T}_i > t\}\right). \tag{21}$$

(20) and (21) together with (17) immediately imply the statement of part (II). □

[20] See Gordy 2003, p. 223, for a similar application of the Theorem.