Central European Journal of Economic Modelling and Econometrics

Modelling Recovery Rate for Incomplete Defaults Using Time Varying Predictors

Wojciech Starosta*

Submitted: 16.01.2020, Accepted: 10.04.2020

Abstract

The Internal Rating Based (IRB) approach requires that financial institutions estimate the Loss Given Default (LGD) parameter not only based on closed defaults but also considering partial recoveries from incomplete workouts. This is one of the key issues in preparing bias-free samples, as there is a need to estimate the remaining part of the recovery for incomplete defaults before including them in the modeling process. In this paper, a new approach is proposed, where parametric and non-parametric methods are presented to estimate the remaining part of the recovery for incomplete defaults, in predefined intervals concerning sample selection bias. Additionally it is shown that recoveries are driven by different set of characteristics when default is aging. As an example, a study of major Polish bank is presented, where regression tree outperforms other methods in the secured products segment, and fractional regression provides the best results for non-secured ones.

Keywords: LGD, workout approach, incomplete defaults, partial recovery rate

JEL Classification: C51, G32

*University of Lodz, Poland; e-mail: w.starosta@wp.pl; ORCID: 0000-0002-2306-0263

1 Introduction

Basel II regulations on the Advanced Internal Rating Based approach permit financial institutions calculate three risk parameters (Probability of Default - PD, Exposure at Default - EAD, and Loss Given Default - LGD) in-house. Simultaneously with this option, minimal technical standards and guidelines concerning estimation have been described (Basel Committee on Banking Supervision 2017). Among them, four methods of LGD calculation can be found. The first and, at the same time, the most popular practice is "workout approach" (Basel Committee on Banking Supervision 2005, p. 4), which is based on discounting cash flows up to the moment of default in reference to the amount of exposure from the same date. The second technique is the implied historical LGD, based on the experience of total losses and PD estimates. The third and fourth methods are market LGD, based on the prices of traded defaulted loans, and implied market LGD, which is derived from non-defaulted bond prices by means of an asset pricing model (Basel Committee on Banking Supervision 2005, p. 12).

Due to the quality of estimates, the workout approach is preferred both by supervisors and in the literature (Basel Committee on Banking Supervision 2017, p. 114 and Anolli, Becalli, Giordani 2013, p. 92). However, for the sake of a complicated way of defining and calculating the mentioned recovery amounts, as well as determining the exposure at the moment of default, the workout is governed by a non-standard number of guidelines. One of them states that it is essential to take into account all observed defaults from the selected period (Basel Committee on Banking Supervision 2017, p. 34). Such a period should cover as broad information as possible so that the financial institution can reflect the current debt collection process and policies in the LGD model. Taking into consideration that debt recovery can last for several years, in the selected sample there are cases where the process has started but not yet finished at the moment of model preparation (so called open or incomplete default). It leads to the state in which the value of a dependent variable is not known for part of the observations, which is a consequence of its definition, usually referred to as the recovery rate (RR):

$$RR = \frac{\sum_{t=1}^{n} CF_t/(1+d)^t}{EAD}, \tag{1}$$

wherein the nominator sum of discounted cash flows is located and the denominator contains Exposure at Default (Anolli, Becalli, Giordani 2013, p. 92). The need of taking all defaults from selected period is problematic in cases where the final value of the nominator is not known due to the open debt collection process. Even if regulatory issues (Article 181(1)(a) of the Capital Requirements Regulation (CRR)) were not present, including only completed workouts would be not representative for the modeled parameter, and also unjustified bias would be introduced connected to the omitted cases. As stated by Rapisarda and Echeverry (2013), profiles of closed and open defaults can result in different LGD, so properly reflecting such situation lead

to more reliable estimates. In particular using only resolved cases in building LGD model introduce downward bias as more short-lasted high-RR cases would be taken into sample. On the other hand using unresolved cases as-is, end with upward bias, as unresolved cases will on average have higher final RR as observed at the moment of model preparation. In this paper we present a method of inclusion the unresolved cases, using the estimate of the remaining part of RR, which will be realized in future, to the resolved part of the sample. This leads to non-biased sample, which produces non-biased LGD estimates. The more reliable are the results of partial RR estimation, the more precise the final LGD output is, as then it possess all the patterns observed during the historical period used in the model preparation.

Our first contribution is a time and collateral dependent sample preparation, which aims to reduce bias connected primarily with the occurrence of different recovery patterns in closed and open defaults samples. Direct estimation from closed cases may lead to downward estimation bias. This is due to the fact that, among completed cases, there are usually more relatively short ones which ended with full recovery. On the other hand, among open defaults, reverse dependence is possible, so cases that are in default status for a long time with a low recovery rate may prevail. To solve this issue, we separately estimate partial recovery rate models in pre-defined sub-samples to reflect inherent features of each. We split the sample by the time in default such that different variables drives the recovery of 3-month default opposite to 30-month default. What is more, we differentiate the state before and after collateral realization for secured credits to include the change in client recovery pattern, when tangible asset is lost. The second contribution is related to the potential superiority of non-parametric methods in estimating the partial recovery rate over parametric ones in terms of the precision of the estimates given. In Dermine and Neto (2006) or Bastos (2010) additive or multiplicative version of the Kaplan-Meier estimator was used to incorporate unresolved cases, as time in default and marginal recoveries were used to estimation. Our idea is to build a different parametric and non-parametric models, potentially using various predictors which should address the problem of non-linearity between dependent and independent variables. Such methods are widely used in the LGD modeling, but according to the author's knowledge will be used for the first time in the partial recoveries estimation. The conservativeness of the estimates could be easily obtained in each solution (which is one of the major assumptions concerning Basel II/III regulations), so the presented approach can be treated as part of the discussion about the upcoming adjustments in Basel IV.

The process of estimating the recovery rate is held via fractional regression, beta regression, regression trees and support vector machines, which are gaining more and more popularity in both academic and business applications, as an alternative to the canonical regression methods. All four approaches were previously used in the LGD estimation, so we adopt them to predict the partial recovery rates not coincidentally. The ultimate goal is to prepare a sample containing all defaults together with an estimation of the remaining part of the recovery rate for open cases. This leads to

more precise LGD estimates than those resulting from the estimation only based on the sample with closed cases. In addition, performance of the methods is checked on out-of-time data, which refers to defaults that are not closed at the moment of estimation, but their realized value is already known in the validation set.
The structure of the paper is as follows. First, a review of the existing literature both on the subject of overall LGD estimations, as well as those studies where the problem of open cases is raised is presented. The second section discusses the sample preparation method to take into account the problem of the different statuses of open cases. The third section contains a brief description of the methods used in the estimation. The fourth section demonstrates the conducted study on the training sample carried out in 2015 and check the effectiveness of the methods on the out-of-sample data from 2017. Finally, a summary of the results is presented along with a suggested direction for further research.

2 Literature review

With the appearance of the settlements enclosed in the New Basel Capital Accord (Basel II), the interest in modeling credit risk parameters both among practitioners and in the academic environment increased dramatically. Although the approach to each of them has been standardized since 2004, methods that are often a combination of techniques previously described, or the application in a particular area of solutions known from other fields, are still being developed. In the LGD parameter modeling as classical methods averaging in pools (Izzi, Oricchio, Vitale 2012), linear regression (Anolli, Becalli, Giordani 2013) and beta regression (Huang and Oosterlee 2011) can be considered. These are also the methods most preferred by supervisors as well recognized both in theoretical and interpretative terms. However, it is necessary to notice the shift to more complex or even non-parametric methods, often inadequately referred to as "black boxes". Some of the most interesting proposals have been included in works of Belotti and Crook (2007, decision trees), Luo and Shevchenko (2013, Markov Chains), Brown (2012, neural networks and two-staged models) and Siddiqi and Van Berkel (2012, scoring based methods usage).
However, the literature mentioned above in most cases does not discuss the subject of the inadequacy of the sample; the modeling process begins when the dependent variable is already completely prepared. The subject of open cases was initially discussed in the paper by Dermine and Neto (2006), where the actuarial-based mortality approach with the Kaplan-Meier estimator was used to determine the recovery rate. Initially, Marginal Recovery Rates (MRR) in period t were determined as cash flow paid at the end of period t divided by loan outstanding at time t. Secondly, PULB (Percentage Unpaid Loan Balance at the end of period t) was calculated as $1 - MRR$, and finally, Cumulative Recovery Rate T periods after the default was recognized as $1 - \prod_{t=1}^{T} PULB_t$. By using both completed and open cases, recovery rate curves and exposure-weighted recovery rate curves for each period t

were determined. A similar approach was used in Bastos (2010) with its explication in Rapisarda and Echeverry (2013), where a reformulation from the exposure-weighted Kaplan-Meier estimator to a default-weighted one was shown. This is viewed as more appropriate to ensure compliance with supervisor guidelines. A second change was the transition from the aggregation of recovery rates over time and then across exposure, to aggregation recovery rates across exposure and then over time. The difference is situated in the statement that *in the first case, ultimate recovery rates must be realizations of the same random variable whereas in the second recovery, profiles need to be realizations of the same stochastic process* (Rapisarda and Echeverry 2013, p.1). Finally, the authors show distributions of recovery rates over time, which leads to more precise LGD estimators than those based only on completed cases. An overview of methods like the use of external databases, time criteria or the extrapolation of future recoveries was described in Zięba (2017), where it was stated that extrapolation gives the best results, both in terms of increasing the sample size and the impact of the final LGD estimators. The most conservative approach has been presented in Baesens, Roesch and Scheule (2016), where one of the proposals is to take account of incomplete cases as if they were completed; however, it may lead to a revaluation of the final LGD values. On the regulatory side, precise assumptions regarding the treatment of open cases should appear together with the records of Basel IV (Nielsen and Roth 2017, p. 72).
One of the aims of this study is to extend the existing literature with further methods of estimating recovery rates for open cases, which is also in line with upcoming regulations. Additionally, an attempt is made to reduce bias coming from the possibility of differences in populations of open and closed defaults and the potential revaluation of recovery rates based only on closed cases. Finally, described approach is validated on out-of-time data.

3 A bias free sample design

This section provides an overview of the sample preparation process. The nature of the recovery rate imposes at least three states in which an exposure with the premise of default can be found.

Figure 1: Closed default. All recoveries were obtained before the reference date

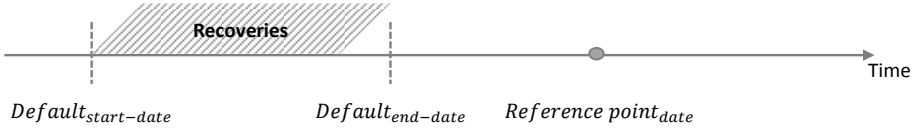


Figure 1 illustrates a standard example in which the final recovery rate is known, regardless of where the recoveries come from. Figures 2 and 3 demonstrates the

Figure 2: Incomplete default with collateral realization before the reference date. It is still possible to obtain recoveries from the client's own payments

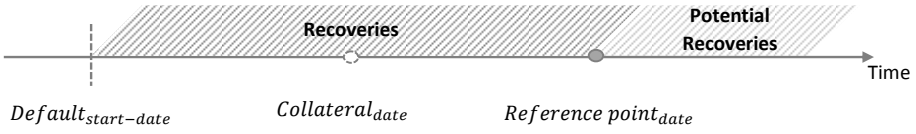
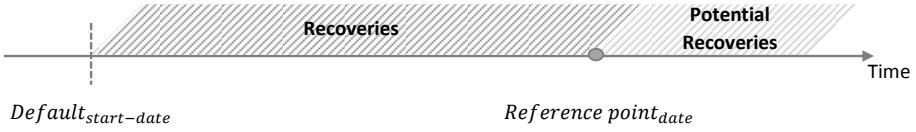


Figure 3: Incomplete default without collateral realization before the reference date. It is still possible to obtain recoveries both from the client's own payments and collateral realization in the case of a secured product



situations where the recovery process has not been finalized and it is necessary to estimate the remaining part of the recovery rate. At this stage, it seems essential to distinguish secured exposures (e.g. by mortgage or vehicle), for which the process differs radically before and after collateral realization. Before realization, the recovery rate consists of the consumer's own payments and a theoretically possible repayment from the collateral; after realization, only the consumer's own payments are possible, but their motivation is significantly different from before. The recovery rate, taking into account the above distinction, is calculated as follows:

$$RR = \begin{cases} RR_{pay} + RR_{coll}, & \text{for closed default (Fig. 1),} \\ RR_{pay} + RR_{coll} + \widehat{RR}_{pay}, & \text{for open default with collateral realization (Fig. 2),} \\ RR_{pay} + \widehat{RR}_{pay} + \widehat{RR}_{coll}, & \text{for open default without collateral realization (Fig. 3),} \end{cases} \tag{2}$$

where:
RR – recovery rate, as the dependent variable in the LGD model,
RR_{pay} – actual value of the recovery rate from the client's own payments,
RR_{coll} – actual value of the recovery rate from collateral realization,
$\widehat{RR}_{pay}$ – predicted value of the partial recovery rate from the client's own payments for the period from the reference point to the end of recovery process,

$\widehat{RR}_{coll}$ – predicted value of partial recovery rate from collateral realization for the period from the reference point to the end of recovery process.
The reference point is understood as the date from which the data originate. The actual values come from recoveries obtained before this date. The predicted values are values estimated for the period from the reference point till the end of the recovery process (it is not defined as a time period, rather any point in the future when the process will finish). And although the reference date is the same for all cases in the sample, for incomplete ones, the period from the moment of default till the reference date is different. This is key information in the recovery process, because the estimated recovery rate will be different for cases in which the default occurred a month before the reference date to the cases where the default occurred five years before the reference date. Therefore, for the needs of estimation, both closed and open cases should be divided into sub-periods in which the estimation of parameters will take place. The more granular the period selected, the more accurate the possible results will be; however, excessive fragmentation may lead to instability of estimates, as fewer and fewer observations will be involved in subsequent intervals.
Taking into consideration the remarks above, the recovery rate formula for open cases can be transformed in a manner that depends on the time in default and the collateral realization:

$$RR = \frac{\sum_{t=1}^{l} CFpay_t/(1+d)^t}{EAD} + \frac{\sum_{t=1}^{l} CFcoll_t/(1+d)^t}{EAD} + \widehat{RR}_{pay}^{l+1} + \widehat{RR}_{coll}^{l+1}, \quad (3)$$

where:
$CFpay_t$ – cash flows from own payments up to the reference date carried out in period t,
EAD – exposure at default,
l – the number of periods from the date of the default to the reference date,
$CFcoll_t$ – cash flows from the collateral realization up to the reference date carried out in period t,
$\widehat{RR}_{pay}^{l+1}$ – estimated value of the partial recovery rate from own payments from the moment $l+1$ until the end of recovery process,
$\widehat{RR}_{coll}^{l+1}$ – estimated value of the partial recovery rate from the collateral realization from the moment $l+1$ until the end of recovery process in cases where the collateral realization has not yet taken place.

This method of recovery rate construction is free from the bias caused by the selection of the sample, as it contains appropriate patterns both for complete and open cases.

At this point, it is possible to determine a way to estimate $\widehat{RR}_{pay}^{l+1}$ and $\widehat{RR}_{coll}^{l+1}$. At each time interval, the actual recovery rates are calculated from the start time of the interval (m) to the end of recovery process window (n) on the basis of complete cases.

$$RR_{pay}^{m} = \sum_{t=m}^{n} \frac{CFpay_t}{EAD(1+d)^t}. \quad (4)$$

The result of this equation is population divided into sub-samples consists of cases which lived long enough to be a part of each. Taking 6-months intervals as an example, we can see that all cases are used to determine $\widehat{RR}_{pay}^{l+1}$ for defaults being in interval from 0 till 6 month, but only defaults which lived at least till month 60 are used to estimate $\widehat{RR}_{pay}^{l+1}$ for open cases being in interval from 60 till 66 month. The recovery rate for $\widehat{RR}_{coll}^{m}$ is calculated analogously, where m is time interval for which the variable value is calculated. For example for 6-months periods, sum of recoveries from the beginning of default to the end of recovery process is determined first. The second period runs from the sixth month of recovery until the end of recovery process, and so on. Such a construction allows us to create a set, on the basis of which it is possible to estimate the partial recovery rate for each open case depending on: (i) time in default, (ii) hitherto recovery from own payments, and (iii) recovery from collateral realization.

4 Recovery rate estimation methods for open cases

The following section presents a brief summary of the methods used in recovery rates modeling, which in our study are used in the process of partial recovery rate estimation and is divided into two sub-sections corresponding to the groups convergent in terms of theoretical assumptions. The first category consists of parametric methods in which fractional regression and beta regression are presented. The second one contains regression trees and support vector machines.

4.1 Parametric methods

The first method discussed in this subsection is fractional regression (FR). Its use for LGD modeling was proven to give reasonable results, inter alia, in Belotti and Crook (2009) or Bastos (2010). Detailed assumptions about this type of regression can be found in Papke and Woolridge (1996). For the problem of estimating recovery rates, a lack of assumptions about the distribution is crucial; only the conditional mean must be correctly specified in order to obtain consistent estimators. Assuming that

$$E(y_i \mid \mathbf{x_i}) = G(\mathbf{x_i}\boldsymbol{\beta}) = 1/[1 + \exp(-\mathbf{x_i}\boldsymbol{\beta})] \tag{5}$$

the fractional logit model parameters $\hat{\boldsymbol{\beta}}$ can be estimated by maximizing the Bernoulli log-likelihood function (as in binary logistic regression) (Papke and Woolridge 1996, p. 621):

$$L(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} y_i \log[G(\mathbf{x_i}\hat{\boldsymbol{\beta}})] + (1 - y_i) \log[1 - G(\mathbf{x_i}\hat{\boldsymbol{\beta}})], \tag{6}$$

where $i = 1, ..., n$, n is a sample size and $\mathbf{x_i}$ is a vector of explanatory variables for case i. However, it should be noted that the explained variable must come from a

specific range ($0 \leq y_i \leq 1$), which is not always ensured in the case of recovery rate modeling (like where direct and indirect costs were added or collateral was sold at price higher than EAD). The solution is to apply a linear transformation in the form of classical unitarization:

$$\widetilde{RR_i} = \frac{RR_i - \min\limits_i\{RR_i\}}{\max\limits_i\{RR_i\} - \min\limits_i\{RR_i\}}. \tag{7}$$

As a result of the above-mentioned normalization formula, the obtained transformed recovery rates $\widetilde{RR_i}$ belong to the interval $[0; 1]$. Backward transformation is done during out-of-sample verification.

The second method, which is gaining more and more popularity in LGD estimation, is Beta Regression (BR). Besides the publications mentioned in Section 1, it can be found in Chalupka and Kopecsni (2008), Stoyanov (2009) or Tong, Mues, and Thomas (2013). What makes Beta Regression so popular is its flexibility in the case of modeling quantities constrained in the interval $(0; 1)$. Depending on the choice of parameters, the probability density function can be unimodal, U-shaped, J-shaped or uniform:

$$f(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \tag{8}$$

where $\Gamma(\cdot)$ denotes the gamma function. It is assumed that $\alpha > 0$ and $\beta > 0$. In such a formulation, α pushes the density toward 0 and β toward 1. Without loss of generality, these two parameters can be reformulated in terms of mean (μ) and dispersion (assuming $\varphi = \alpha + \beta$) in the following way (Huang and Oosterlee 2011):

$$\alpha = \mu\varphi, \qquad \beta = (1-\mu)\varphi. \tag{9}$$

Within the framework of Generalized Linear Models (GLM), both μ and φ can be modeled separately, with a location model for μ and a dispersion model for φ, using two different or identical sets of covariates (Liu and Xin 2014). The mean model can be expressed as:

$$g(\mu) = \gamma_0 + \sum_i \gamma_i a_i, \tag{10}$$

where a_i denotes explanatory variables, γ_i coefficients and g is the monotonic, differentiable link function. Since the expected mean μ is bounded by 0 and 1, logit can be used as the link function:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right). \tag{11}$$

Dispersion parameter φ can be treated as fixed or it can be modeled by another GLM (Huang, Oosterlee 2011):

$$h(\varphi) = \zeta_0 + \sum_i \zeta_i a_i, \tag{12}$$

where h is a link function and ζ_i are coefficients. The simplest way to achieve it is to use:

$$\varphi = e^{\zeta_0 + \sum \zeta_i a_i}. \tag{13}$$

4.2 Non-parametric methods

Tree-based methods (RT) recursively partition the original sample into smaller subsamples and then fit a model in each one. The concept is clear and easy to implement, yet the method is powerful and was adopted for LGD purposes inter alia in Qi and Zhao (2011) or Van Berkel and Siddiqi (2012). To build a tree, an algorithm is needed which, at each node t, evaluates the set of variable splits to find the best one, i.e., the split s that maximizes the decrease in impurity (im) (Brown 2012, p.51):

$$\Delta im(s,t) = im(t) - p_L im(t_L) - p_R im(t_R), \tag{14}$$

where p_L and p_R denote the proportion of observations associated with node t that are sent to the left child node t_L or to the right child node t_R. In the case of a continuous variable, like a recovery rate, regression trees are used and a standard criterion for this type of model is minimizing the sum of squares $\sum (y_i - \hat{y}_l)^2$, which leads to averaging recovery rate in region R_m as the value of each leaf:

$$\hat{c}_m = avg(y_i \mid x_i \in R_m). \tag{15}$$

Finding the best partition is quite straightforward. First, splitting variable j and split point s are selected, so a pair of half-planes can be defined:

$$R_1(j,s) = \{X \mid X_j \leq s\} \text{ and } R_2(j,s) = \{X \mid X_j > s\}. \tag{16}$$

The second splitting variable j and split point s are searched to solve:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]. \tag{17}$$

For the choice of j and s, the inner minimization is solved by:

$$\hat{c}_1 = avg(y_i \mid x_i \in R_1(j,s)) \text{ and } \hat{c}_2 = avg(y_i \mid x_i \in R_2(j,s)). \tag{18}$$

After the first split is determined, the procedure is repeated on all regions (Hastie, Tibshirani, and Friedman 2008, p. 307). The question arises when one should stop growing each tree. This is another advantage of the described approach, as there are many elegant methods to achieve this:

1. establishing a minimal impurity decrease,
2. fixing the maximal depth,

3. selecting the minimal number of observation in a leaf.

These are also the most common methods of solving the instability issue, which is often raised when tree-based models are used. The lack of estimates smoothness can be considered as another drawback, as it can deteriorate performance in the regression setting, where underlying function is expected to be smooth (Hastie, Tibshirani and Friedman 2008). However in the case of partial recovery rate estimation, it is not an issue, because we can define each region as different recovery pattern (specific scenario which leads to particular value of partial RR).
The Support Vector Machine (SVM) is another non-parametric technique for classification and regression problems used in LGD modeling more and more frequently (see Loterman et al., 2012 or Yao, Crook, and Andreeva 2017). It produces nonlinear boundaries by constructing a linear boundary in the transformed version of the feature space. Formally, an SVM constructs a hyperplane or set of hyperplanes in a potentially infinite dimensional space. The SVM finds this hyperplane using support vectors and margins (defined by support vectors). In a regression model (Hastie, Tibshirani, and Friedman 2008, p. 434):

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x), \tag{19}$$

where $h_m(x)$ is a set of basis functions (by which we denote a function that augments vector of $\mathbf{X}$ by additional variables via selected transformation, like $h_m(x) = x_j x_k$ or $h_m(x) = \log(x_j)$) and $m = 1, 2, ..., M$, the goal is to minimize:

$$L(\beta) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \|\beta\|^2 \tag{20}$$

for some general error measure $V(r)$. Regardless of $V(r)$ the solution of has the form:

$$\hat{f}(x) = \sum_{i}^{n} \hat{\alpha}_i K(x, x_i), \tag{21}$$

where $K(x, y) = \sum_{m=1}^{M} h_m(x) h_m(y)$ and it denotes specific kernel. This allows SVM to easily capture non-linear dependencies by using different kernel function. There are many possible kernels, but in this study, the radial one is used with squared Euclidean distance:

$$K(x, x') = e^{-\|x - x'\|^2 / 2\sigma^2}. \tag{22}$$

5 Empirical analysis of partial Recovery Rates

In this section, an attempt to estimate the partial recovery rate is made using data from one of the largest Polish banks applying the AIRB regime. A sample of completed defaults from 2003 to 2015 is used for models preparation. These models then predicts the recovery rate for open cases from the same time period, and finally, goodness of fit is checked on a part of the sample where the recovery process finished during the 2015 – 2017 period. The process of parameter estimation is conducted in 6-months intervals, so the first interval predicts the final recovery rate for cases whose default lasted from 0 to 5 months, the second from 6 to 11 months, etc. We assume that such a split is granular enough and allows to prepare stable models in each interval. In presented models part of the recovery connected with the collateral ($\widehat{RR}_{coll}$) is included via Loan To Value (LTV) variable calculated at each point after default. To reflect both possibilities drawn on Figure 2 and Figure 3, its construction is as follows:

$$LTV_l = \begin{cases} \dfrac{\text{Loan value}_l}{\text{Collateral value}_l}, & \text{if collateral was not sold in selected interval,} \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

Additionally, we benchmarked our models to simple Naïve Markov chain in the form of transition matrix (cf. Jarrow et al., 1997). We divided partial recoveries into classes, taking into account only months since default, and estimate the final class according to the initial class for each case. It is an equivalent of "mean prediction", frequently treated as a benchmark to more sophisticated methods or recovery rates estimation.

5.1 Sample description

As mentioned above, the sample is made up of default events which occurred between 2003 and 2015 and contains both secured (ML) and non-secured loans (NML). The predictors were obtained at the moment of default and then at each point respectively. This allows us to show the dynamic nature of characteristics during default window and access variables specific to this stage of the process (like DPD or due amounts). The proportion of completed and open cases is presented in Table 1. It clearly demonstrates that this specific portfolio suffers from a huge share of open defaults. Taking into account only the completed ones would lead to the removal of 46.53% of secured and 62.15% of unsecured contracts, so there is no doubt that data selection bias would be introduced. What is more, there is a significant difference in the distribution of explanatory variables, as shown in Tables 2 and 3. This may cause another problem with data representativeness.
The variables RR^m_{pay} and RR^m_{coll} are prepared according to formulas from Section 2 in 6-month time intervals, and consist of both principal and interest recoveries. So, each point in Figure 4 is a mean recovery rate from the beginning of the interval till the end of the recovery process.

Table 1: Proportion of closed cases in the sample by type of credit

	Closed	Open	All
Secured	53.47%	46.53%	6 953
Non-secured	37.85%	62.15%	122 353

Table 2: Descriptive statistics for secured credits by label

Variable	Label	Mean	5^{th} Pctl	25^{th} Pctl	50^{th} Pctl	75^{th} Pctl	95^{th} Pctl	Max
EAD	Closed	318k	38k	105k	207k	394k	942k	6.598k
	Open	421k	60k	161k	297k	509k	1.167k	9.505k
Interest rate	Closed	0.042	0.009	0.025	0.040	0.054	0.090	0.128
	Open	0.037	0.009	0.013	0.034	0.050	0.089	0.160
Days past due (DPD)	Closed	46.00	0	0	33	91	92	443
	Open	49.20	0	13	46	91	91	1681
Tenor	Closed	294	120	239	336	359	360	360
	Open	306	155	240	358	359	360	360
Requested amount	Closed	359k	54k	123k	236k	438k	1015k	7617k
	Open	465k	78k	184k	330k	561k	1281k	9977k
Months on book (MOB)	Closed	42.99	8	22	39	60	93	144
	Open	50.19	12	31	47	68	97	143
Due principal	Closed	3.4k	0	0	648	1843	7755	893k
	Open	4.5k	0	322	1254	3006	11k	722k
Due interest	Closed	1.6k	0	0	584	1592	6217	109k
	Open	2.8k	0	210	891	2192	8k	135k
Principal	Closed	321k	40k	106k	208k	340k	945k	7209k
	Open	427k	60k	162k	302k	514k	1189k	12354k
Interest	Closed	2.1k	23	347	868	2056	7482	130k
	Open	2.8k	70	490	1214	2806	9662	172k
Due amount	Closed	5.1k	0	0	1442	3709	14k	910k
	Open	6.7k	0	782	2476	5531	19k	722k
LTV	Closed	0.95	0.03	0.27	0.69	1.00	1.06	1.63
	Open	1.06	0.05	0.30	0.76	1.00	1.32	1.78
Foreign currency	Closed	0.77	0	1	1	1	1	1
	Open	0.77	0	1	1	1	1	1

It can be seen that recoveries from the consumer's own payments decrease over time, which seems reasonable, as client motivation to repay diminishes with duration of default and longer defaults are seen as more problematic (poor financial situation, difficulties with reaching the customer, client goes into litigation, etc.). Also, recoveries from secured loans are greater than non-secured ones, which indeed is consistent with the findings from the previous studies (see e.g., Gurtler and Hibbeln 2013), as clients care more about losing their home or car as a consequence of a default. Finally, the shape of the collateral RR curve results from the more discrete

Table 3: Descriptive statistics for non-secured credits by label

Variable	Label	Mean	5^{th} Pctl	25^{th} Pctl	50^{th} Pctl	75^{th} Pctl	95^{th} Pctl	Max
EAD	Closed	7.1k	85	2.4k	4k	5.8k	23k	808k
	Open	12k	910	3.2k	5.5k	11k	47k	243k
Interest rate	Closed	0.184	0.110	0.160	0.200	0.210	0.227	0.662
	Open	0.168	0.098	0.144	0.169	0.200	0.230	0.590
Days past due (DPD)	Closed	126.4	25	91	91	92	474	3132
	Open	75.68	0	43	88	91	123	2201
Tenor	Closed	16.06	12	12	12	12	60	120
	Open	24.6	12	12	12	36	60	156
Requested amount	Closed	8.1k	1.2k	3k	5k	6.9k	27k	1000k
	Open	13.7k	1.3k	3.5k	5.6k	13k	50k	2400k
Months on book (MOB)	Closed	22.36	5	9	16	30	58	138
	Open	29.18	5	12	23	40	74	154
Due principal	Closed	1.9k	1.92	431	623	1.1k	5.9k	800k
	Open	1.6k	0	138	523	961	4.2k	425k
Due interest	Closed	196	0.01	26	100	204	650	17k
	Open	297	0	39	128	304	1.2k	43k
Principal	Closed	6.9k	73	2.3k	3.8k	5.7k	23k	800k
	Open	12k	889	3.1k	5.2k	11k	46k	2399k
Interest	Closed	267	2.66	69	164	301	792	19k
	Open	394	19	90	204	414	1.4k	43k
Due amount	Closed	2.1k	26	526	746	1.4k	6.3k	808k
	Open	1.9k	1.33	327	682	1.3k	5.2k	452k

construction of the process where collateral realization can happen (in general) once in the default window, in contrast to RR from the client's own payments, where the client can repay the due amount using more than one transaction. The structure of the sample in division by payers, non-payers and partial payers is shown in Figure 5. Diminishing number of payers along with relatively stable share of non-payers, in great extend supports the conclusions drawn from analyzing the mean recovery rate curves.

5.2 Models

We estimate RR till month in default equal to 60, as a result of sharp observation number decrease after this interval. This effects with assigning values from the 60m interval to every observation with months in default greater than 60, but no greater than 96, when the values of average recoveries are set to zero. This is in line with the general assumption that after a certain time, financial institutions no longer expect any repayments (Basel Committee on Banking Supervision 2017, p. 34). Following Section 2, twelve models for each method are prepared based on static and dynamic variables presented in Tables 2 and 3. Point estimates are provided in the Appendix.

Figure 4: Mean of the recovery rate at each point till the end of recovery process for cases which lived in particular interval

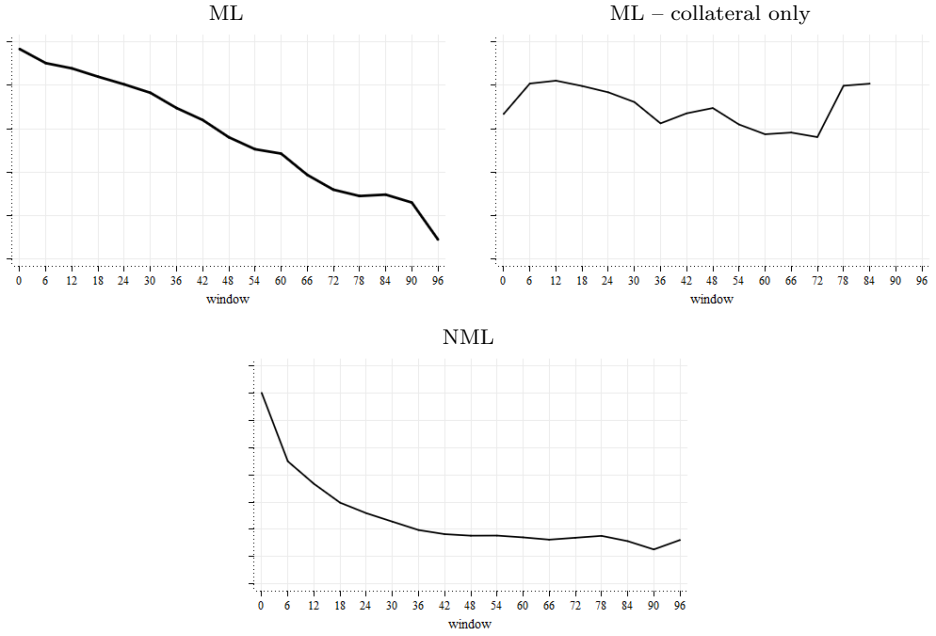
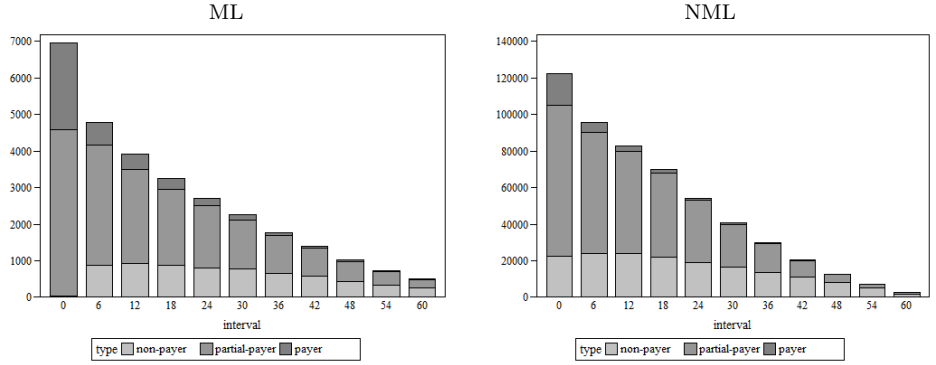


Figure 5: Number of observations in consecutive intervals in division by payers, non-payers and partial payers



Each column shows the information from the beginning of the interval till the end of recovery process

Fractional regression

Due to highly correlated variables in our data set, we use L1 criterion for regularization scheme to select the best set of predictors. Tables 4 and 5 summarizes the results in the case of variables used, RMSE and the selected correlation measures (Pearson and Spearman coefficients).

Table 4: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Fractional Regression – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√										
DPD	√	√	√	√	√	√	√	√	√	√	√
TENOR											
REQ. AMOUNT											
MOB	√	√	√	√				√			
DUE PRINCIPAL											
DUE INTEREST											
PRINCIPAL											
INTEREST											
DUE AMOUNT											
LTV								√	√	√	√
FOREIGN CURRENCY											
RMSE	.1329	.1795	.1879	.1962	.2265	.2419	.2612	.2592	.2998	.2649	.2673
PEARSON	.2388	.3207	.3988	.4692	.4418	.4761	.5418	.6097	.5281	.5740	.6532
SPEARMAN	.1956	.1084	.1547	.2157	.1758	.2052	.3490	.4653	.4411	.5023	.5680

First conclusion, that we can draw from Table 4 and Table 5, consist in recovery pattern changes observable across time in default. The only variable significantly important in all regressions is DPD, which means that along with the increase of days-past-due partial recovery rate is decreasing over time. But as time in default rise, we can see switch from contract based variables (interest rate, months on book) to LTV. This is definitely something worth to examine. When client goes into default, at the beginning his recoveries consist mainly of own payments (RR_{pay}) and majority of them deal with debt with their own strengths. But if default last for more than 42

Table 5: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Fractional Regression – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√		√	√	√	√	√	√	√		
DPD				√	√					√	√
TENOR	√	√									
REQ. AMOUNT											
MOB	√	√	√	√	√	√	√	√	√	√	
DUE PRINCIPAL											
DUE INTEREST											
PRINCIPAL							√	√	√	√	
INTEREST											
DUE AMOUNT											
RMSE	.2995	.2883	.2682	.1750	.2345	.2215	.2139	.2120	.2266	.2422	.2385
PEARSON	.3054	.3098	.2664	.2563	.2341	.2089	.2026	.2157	.2092	.1981	.2432
SPEARMAN	.2303	.2824	.2321	.1521	.1154	.1046	.1079	.1613	.2076	.3064	.3783

months, then capability to repay worsens and the collateral is used more frequently to compensate the remaining part of the debt. Collateral realization can introduce non-linearity into the model, which could not be easily captured by fractional regression and can be the reason for RMSE rising in latter intervals. It is also an unique characteristic of secured credits, as being in default for more and more months usually leads to involving court, bailiff or consumer bankruptcy. This events are not efficiently modeled by contract characteristics only or even if so, then non-linear approach to each case should be handled by different method, which is able to produce more robust estimates for this intervals.

Non-secured credits behave similarly when it comes to recovery pattern change. At the beginning we can see that static variables, like interest rate or tenor, are used more frequently to estimate partial recovery rate. But in closing intervals DPD, months on book and principal are of greater importance. We can conclude that at the beginning it is difficult to predict partial recovery rate as similar contracts are driven by the same characteristics to different RR levels, which is confirmed by higher RMSE in initial intervals. After that stronger patterns appear derived by DPD and MOB mainly and RMSE decreases (opposite to secured loans). The reason for this could also be situated in specific collection department policy, which could be the part of the process from some point (after DPD threshold exceeded for example).

Beta regression

Due to highly correlated variables in our data set, we use L1 criterion for regularization scheme to select the best set of predictors. Tables 6 and 7 summarizes the results in the case of variables used, RMSE and the correlation parameters.

Table 6: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Beta Regression – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√									√	
DPD	√	√	√	√	√	√	√	√	√	√	√
TENOR										√	
REQ. AMOUNT								√	√		
MOB	√	√	√	√	√	√	√				
DUE PRINCIPAL											
DUE INTEREST											
PRINCIPAL											
INTEREST											√
DUE AMOUNT											
LTV			√					√	√		
FOREIGN CURRENCY								√	√		
RMSE	.1415	.1797	.1875	.1983	.2268	.2442	.2683	.2733	.3003	.2817	.2793
PEARSON	.2150	.3231	.4082	.4680	.4485	.4862	.5247	.5779	.5591	.5089	.6331
SPEARMAN	.2139	.1062	.1685	.2083	.2238	.2840	.4086	.5998	.5241	.4856	.5383

Beta regression is able to find more relationships with predictors than fractional regression but it did not translate into better results on average. What is interesting is a fact, that in BR collateral is important in one of the first stages, then this importance is lost for a while, but finally like in FR it is main driver along with DPD when it comes to secured loans. RMSE rises with time, which can be the result of high complexity of defaults lasting years in default (like in FR).
An opposite arises with NML loans, where the biggest errors are observed again at the beginning of the default. Here, motivation to repay is significantly different from secured loans, so finding proper patterns in the data seems to be harder for the first

Table 7: Variables used in particular regression for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Beta Regression – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD											
INTEREST RATE	√	√	√	√	√	√	√	√	√	√	√
DPD	√				√					√	
TENOR	√	√	√							√	√
REQ. AMOUNT									√	√	√
MOB	√	√	√	√	√	√	√	√	√		
DUE PRINCIPAL											
DUE INTEREST		√	√								
PRINCIPAL						√	√	√	√	√	√
INTEREST	√										
DUE AMOUNT											
RMSE	.3048	.3048	.2865	.1842	.2540	.2411	.2361	.2356	.2507	.2634	.2592
PEARSON	.3118	.3171	.2796	.2401	.2186	.2051	.1986	.2123	.2287	.2381	.2262
SPEARMAN	.2736	.2955	.2405	.1635	.1396	.1094	.0990	.1599	.2996	.3582	.4856

year in default. Then there is a meaningful decrease in error for month 18, which may suggest that the debt collection policy might result in write-offs or termination at that time, and the model captures it. For month 24 and later, the RMSE is quite stable, mainly due to the fact that there is no factor of collateral, so only customers' own payments are modeled.

Regression Trees

As stated in Section 4.2, parametrization needs to be made to build a tree. For the sake of the results comparison, we decide to select the same parameters for every tree, which are ANOVA as the splitting selection method (applicable for continuous variables), complexity parameter selected based on a 10-fold cross-validation, 10 as the maximum depth (selected arbitrarily, but this constraint is not binding as no tree grew so deep) and 30 as the minimum observations in a leaf (to get the statistical significance of the mean). Such parametrization allows to avoid overfitting with building precise tree at the same time.

On the sample selected for model building, it can be seen that regression trees give a lower RMSE for latter intervals (compared to fractional and beta regression), which suggests strong non-linearity between the recovery rate and the explanatory variables. Initial intervals are comparable when it comes to error measure, although regression tree is not limited by not correlated variable selection, which can be easily seen when

Table 8: Variables used in particular tree for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Regression Tree – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD	√	√	√	√	√	√			√		
INTEREST RATE	√		√	√	√	√	√	√	√		
DPD	√	√	√	√					√		
TENOR	√	√						√		√	
REQ. AMOUNT	√	√	√	√	√			√			
MOB	√	√	√	√	√	√	√				
DUE PRINCIPAL	√	√			√			√	√		
DUE INTEREST	√	√			√		√	√	√	√	
PRINCIPAL	√		√	√	√			√	√		
INTEREST		√	√	√		√					
DUE AMOUNT	√					√					√
LTV	√	√	√	√				√	√		
FOREIGN CURRENCY											
RMSE	.1268	.1726	.1790	.1870	.2078	.2168	.2373	.1923	.2103	.2202	.2023
PEARSON	.3811	.4846	.5296	.5755	.5676	.6157	.6492	.8087	.8005	.7310	.8021
SPEARMAN	.3199	.2409	.3670	.4567	.4599	.4643	.5586	.7402	.7652	.7114	.7706

one collate Table 8 with Table 6 or Table 9 with Table 7. Regression tree like BR find collateral, expressed in terms of LTV, significant at the beginning and at the end of recovery process.
For non-secured products EAD, interest rate, DPD and MOB are the main drivers, significant in almost all intervals. However there are also variables which differentiate RR at the initial stages of default, like tenor, requested amount or interest. This is also coherent with statement, that each interval's inherent features should be taken into account, when partial RR are estimated.

Support Vector Machines

The final method also requires parametrization; however, in the plain version, only variable classification (continuous) and the kernel (radial) need to be specified. Tables 10 and 11 summarize the results based on these assumptions.
It is particularly clear, that according to SVM, for secured products partial RR is mainly driven by delinquencies (due principal, due interest, due amount). At the

Table 9: Variables used in particular tree for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (Regression Tree – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD	√	√	√	√	√	√	√	√	√	√	
INTEREST RATE	√	√	√	√	√	√	√	√	√		
DPD	√	√	√	√	√	√	√	√	√	√	√
TENOR	√	√	√	√	√						
REQ. AMOUNT	√	√									
MOB	√	√	√	√	√	√	√	√	√	√	
DUE PRINCIPAL	√	√	√	√	√	√	√	√			
DUE INTEREST	√	√	√								
PRINCIPAL	√	√	√	√	√	√					√
INTEREST	√	√	√								
DUE AMOUNT							√	√			
RMSE	.3245	.3003	.2783	.2548	.2407	.2304	.2171	.2039	.2116	.2123	.1885
PEARSON	.5004	.5563	.5417	.4953	.4696	.4347	.4281	.5203	.4781	.5113	.6412
SPEARMAN	.3815	.5114	.4845	.4090	.4152	.3981	.4235	.5429	.5763	.6511	.6721

beginning more importance is found in EAD and principal, but finally DPD and requested amount took its place. It seems that when client goes into default the main drivers consist in how much he owe at this point and how much of this exposure is past due. But after some time not going back to performing portfolio, his repaying pattern is more dependent on number of days past due and initial amount as higher amounts are generally harder to repay. RMSE for SVM shows lower values than other methods on average, especially for latter intervals, which may be a good reason to consider ensemble of models (but it is beyond this paper).
The most stable results, when it comes to variable selection, are made by SVM for non-secured products. Due principal find its place in 11 out of 12 intervals, EAD in 10/12, interest and due amount in 9/12. Months on book and principal seems to be more important after some time in default, but there is no clear evidence that some variable is particularly meaningful only at the beginning stages. This can support the fact that finding different patterns for the group of close intervals is crucial, as RMSE for SVM achieve higher levels than for the other methods.

5.3 Out-of-sample verification

Using the data from 2017, compiled from cases marked as open in 2015 and closed in 2017, Table 12 show the results of four considered estimation methods. For each case, the time in default is calculated so that assignment to the proper interval could

Table 10: Variables used in particular model for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (SVM – secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD	√	√	√								
INTEREST RATE											
DPD					√		√	√		√	√
TENOR	√										
REQ. AMOUNT						√	√	√	√		
MOB		√	√	√					√	√	
DUE PRINCIPAL	√	√		√	√	√	√	√	√	√	√
DUE INTEREST			√	√	√	√	√		√	√	√
PRINCIPAL	√	√	√								
INTEREST			√	√	√	√			√		√
DUE AMOUNT	√	√		√	√	√	√	√		√	√
LTV								√			
FOREIGN CURRENCY											
RMSE	.1370	.1886	.1869	.1720	.1885	.1920	.2010	.1859	.2059	.1780	.2059
PEARSON	.2541	.4700	.5873	.6761	.6732	.7183	.7662	.8232	.8102	.8399	.8000
SPEARMAN	.3727	.6027	.6493	.6208	.6520	.6861	.7093	.7720	.7552	.8054	.7532

As SVM uses combination of all variables in each interval, five strongest are selected to show meaningful results

be made. Then, the final estimated recovery rate is computed as:

$$\widehat{RR} = \min(RR^l + \widehat{RR}_{pay}^{l+1} + \widehat{RR}_{coll}^{l+1}, 1), \tag{24}$$

where RR^l denotes the recovery rate obtained till the moment of the reference point, which is fixed as 02.2015. We limit the estimated value to 1, to avoid the recovery being higher than the value of the clients' obligations. Table 12 shows the values of RMSE for each method with confidence intervals computed on 100 bootstrapped samples.

The out-of-time predictions shows that for secured credits regression trees seems to capture non-linearity in partial RR modeling with the highest accuracy, but SVM also performs well. Regression trees are supported by its stability, when it comes to comparing RMSE on whole sample and bootstrapped. Fractional regression and beta regression performs significantly worse, as even confidence interval are not overlapping

Table 11: Variables used in particular model for consecutive intervals with RMSE and correlation coefficients between realized and predicted partial RR (SVM – non-secured products)

Variable	0	6	12	18	24	30	36	42	48	54	60
EAD		√	√	√	√		√	√	√	√	√
INTEREST RATE	√										
DPD	√						√				
TENOR											
REQ. AMOUNT	√										
MOB						√	√	√	√		√
DUE PRINCIPAL	√	√	√	√	√	√	√	√	√	√	
DUE INTEREST		√	√	√	√	√					√
PRINCIPAL							√	√	√	√	√
INTEREST	√	√	√	√	√	√				√	√
DUE AMOUNT		√	√	√	√	√		√	√	√	
RMSE	.3574	.3180	.2961	.2713	.2574	.2449	.2296	.2258	.2330	.2346	.2334
PEARSON	.4156	.5270	.5129	.4667	.4377	.4220	.4236	.4537	.4460	.4835	.5106
SPEARMAN	.3588	.4891	.4688	.4445	.4973	.5560	.5817	.6158	.6285	.6760	.6993

As SVM uses combination of all variables in each interval, five strongest are selected to show meaningful results

Table 12: RMSE level for consecutive methods on the out-of-sample set. ML denotes secured loans, and NML non-secured. Best measure is underlined

	Method	RMSE	LCLM	RMSE Bootstrap	UCLM
ML	Fractional Regression	0.2361	0.2336	0.2355	0.2375
	Beta Regression	0.2413	0.2389	0.2410	0.2430
	Regression Trees	<u>0.2168</u>	<u>0.2149</u>	<u>0.2168</u>	<u>0.2187</u>
	Support Vector Machines	0.2197	0.2162	0.2179	0.2197
	Naïve Markov Chain	0.2499	0.2477	0.2492	0.2507
NML	Fractional Regression	<u>0.2871</u>	<u>0.2871</u>	<u>0.2875</u>	<u>0.2879</u>
	Beta Regression	0.3068	0.3064	0.3068	0.3071
	Regression Trees	0.2891	0.2890	0.2895	0.2900
	Support Vector Machines	0.3001	0.3000	0.3006	0.3012
	Naïve Markov Chain	0.4336	0.4331	0.3236	0.4341

non-parametric methods values. The crucial thing here is the presence of collateral, which can be realized in almost any point in time and paths leading to this scenario are not captured well by parametric methods. On particular it can be a result of collection strategy performed by the financial institution or the real estate market

liquidity (or combination of both).
For non-secured loans fractional regression outperforms all other methods both in case of whole sample RMSE, like in non-overlapping confidence intervals. As an alternative regression trees can be viewed. Beta regression and SVMs give significantly worse results, so here the choice is straightforward. The recovery process for credits without collateral seems to be more linear, as it consists only of own payments made by the client during default window. Such patterns can be described mainly by due amounts and DPD, which in fact is done by every method used. And because relationship between RR and these variables can be well described by distribution underlying parametric method, these advantage is moved on out-of-sample data, where non-parametric methods are slightly worse (regression tree) or significantly worse (SVM). Comparing the results to the Naïve Markov Chain, it can be clearly seen, that the rise in quality is relevant. Even the worse method in each segment is not comparable to the selected benchmark (in terms of not overlapping confidence levels), which shows material upgrade of presented approach.
The tasks for future research on partial RR estimation are as follows. Another parametrization of Regression Trees and SVM, like choosing a different splitting method or kernel, should be studied. The trees built in this paper are relatively small, to prevent overfitting, but it looks like there is room to make it more complex to obtain better estimates. Techniques like Random Forest or Gradient Boosting, used, inter alia, in Papouskova and Hajek (2019), can lead to an improvement in performance at the expense of interpretability. SVMs can also be reparametrized with another kernel, like Sigmoid or Hyperbolic, which may reflect the pattern more accurately. Partial Dependency Plots along with Individual Conditional Expectation plots could be added to compare the results with our study. Secondly, interval range is selected arbitrarily, so what is good for one institution, will not always work well for another. Next studies can be broadened by interval selection basing on the recovery patterns specific to the collection process. Thirdly, a reference data set containing information about other risk drivers (such as credit bureau data or detailed collateral characteristics) should be studied to find additional relevant dependencies for consecutive intervals in partial recovery rates estimation.

6 Conclusions

This paper consider a method of estimating partial recovery rates for open cases, basing on modeling recoveries in intervals, where explained variable consist of all cash flows observed from the beginning of the interval till the end of recovery process window. Two parametric and two non-parametric methods are applied on a sample from a Polish commercial bank using AIRB regime to calculate LGD. The selection of the methods was dictated by their robustness confirmed in previous studies (compare with Bastos, 2010 and Yao, Crook, and Andreeva, 2017). Models are built on data from 2003-2015 and validated on defaults closed during the 2015-2017 period. This

study shows that different features drives recoveries when time in default progresses. Recovery patterns are changing and reflect them properly can lead to producing more precise estimates, which finally leads to bias reduction in LGD model, which is in line with Rapisarda and Echeverry (2013) findings. In addition, it is confirmed which method is more suitable to model partial recovery rate. We find that when secured loans are considered, non-parametric methods are able to capture non-linearity, mostly coming from collateral inclusion. Superiority of non-parametric methods was also confirmed in other studies regarding LGD estimation, mention the Loterman et al. (2012) or Tobback et al. (2014). Opposite is true for non-secured loans, where fractional regression gave the best result, but regression trees are only slightly worse. This finding is in opposition to some of the newest studies, but has support in Belotti and Crook (2009), who shows superiority of OLS over selected non-parametric methods. Our solution can be adopted as part of the planned Basel IV framework. The Basel IV will require extended treatment of incomplete defaults compared to previous regulations and consequently leads to more appropriate risk quantification both for setting capital buffers and provision level in the current regulatory and economic environment.

Acknowledgments

The author reports no conflicts of interest. The author alone is responsible for the content and the writing of the paper. We gratefully acknowledge the insightful comments provided by Paweł Baranowski.

References

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

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

[3] Basel Committee on Banking Supervision (2005), Studies on the validation of Internal Rating System, available at: `https://www.bis.org/publ/bcbs_wp14.htm`.

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

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

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

[7] Belotti T., Crook J., (2009), Loss Given Default models for UK retail credit cards, *CRC Working Paper* 09/1.

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

[9] Chalupka R., Kopecsni J., (2008), Modelling Bank Loan LGD of Corporate and SME Segments, *IES Working Paper*.

[10] Dermine J., Neto de Carvahlo C., (2006), Bank Loan Losses-Given-Default: a Case Study, *Journal of Banking and Finance* 30(4), 1219–1243.

[11] Gurtler M., Hibbeln M., (2013), Improvements in loss given default forecasts for bank loans, *Journal of Banking and Finance* 37, 2354-2366, DOI: 10.2139/ssrn.1757714.

[12] Hastie T., Tibshirani R., Friedman J., (2008), *The Elements of Statistical Learning*, Springer, DOI: 10.1007/978-0-387-84858-7.

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

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

[15] Jarrow R., Lando D., Turnbull S., (1997), Markov model for the term structure of credit risk spreads, *Review of Financial Studies* 10, 481–523.

[16] Liu W., Xin J., (2014), Modeling Fractional Outcomes with SAS, *SAS Paper* 1304–2014.

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

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

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

[20] Papke L., Woolridge J., (1996), Econometric method for fractional response variable with an application to 401(K) plan participation rates, *Journal of Applied Econometrics*, DOI: 10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1.

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

[22] Qi M., Zhao X., (2011), Comparison of modeling methods for Loss Given Default, *Journal of Banking and Finance* 35(11), 2842–2855, DOI: 10.1016/j.jbankfin.2011.03.011.

[23] Rapisarda G., Echeverry D., (2013), A Non-parametric Approach to Incorporating Incomplete Workouts Into Loss Given Default Estimates, *Journal of Credit Risk* 9(2), DOI: 10.21314/JCR.2013.159

[24] Regulation (EU) No 575/2013 of the European Parliament and of the council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012.

[25] Stoyanov S., (2009), Application LGD Model Development, *Credit Scoring and Credit Control XI Conference*, available at: `https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2017/03/Application-LGD-Model-Development-Nistico-and-Stoyanov.pdf.`

[26] Tobback E., Martens D., Van Gestel T., Baesens B., (2014), Forecasting Loss Given Default models: impact of account characteristics and the macroeconomic state, *Journal of the Operational Research Society* 65(3), DOI: 10.1057/jors.2013.158.

[27] Tong E., Mues C., Thomas L., (2013), A zero-adjusted gamma model for mortgage loss given default, *International Journal of Forecasting* 29(4), 548–562, DOI: 10.1016/j.ijforecast.2013.03.003.

[28] Van Berkel A., Siddiqi N., (2012), Building Loss Given Default Scorecard Using Weight of Evidence Bins, *SAS Global Forum*, available at: `https://support.sas.com/resources/papers/proceedings12/141-2012.pdf.`

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

[30] Zięba P., (2017), Methods of Extension of Databases Used to Estimate LGD Parameter, *Studia i Prace Kolegium Zarzadzania i Finansów* 150, 31–55.

Table 13: Point estimates for ML products (fractional regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	LTV	Foreign currency	AIC
0		−1.2397	−5.3147			4.7684								856.2
6			−2.7503			2.2711								922.8
12			−1.8843			1.9214								684.6
18			−1.9907			1.0843								562.2
24			−1.9232											470.0
30			−1.9091											368.6
36			−2.1006											281.2
42			−1.4796			1.4982						2.3506		214.9
48			−1.5652									2.7410		172.6
54			−1.5384									2.4625		111.5
60			−2.3784									0.9706		83.6

Table 14: Point estimates for NML products (fractional regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	AIC
0		−6.0128		2.4939		3.5752						49264
6				1.4825		3.1668						42487
12		−1.1924				2.4163						37379
18		−1.0850	−0.6197			1.1666						31277
24		−1.1985	−0.4834			1.3169						21616
30		−0.7032				1.4386						14170
36		−0.3831				1.4426			1.8663			8549.0
42		−0.4507				1.3873			1.6683			4500.7
48		−0.8006				0.9894			2.0679			1905.6
54			1.0519			1.1111			1.4902			897.8
60			1.7044									535.2

Table 15: Point estimates for ML products (beta regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	LTV	Foreign currency	AIC
0		−1.4276	−8.9717			4.5474								−36986
6			−2.5036			1.9032								−8814
12			−1.8090			1.8197						−2.1738		−5333
18			−1.7366			0.6983								−2925
24			−1.4090			0.8710								−1707
30			−1.3399			1.2221								−964.3
36			−1.0940			1.4032								−481.5
42			−1.8745		−1.5825							1.7143	−0.9733	−288.3
48			−1.8366		−1.3086							1.4967	−0.8371	−161.0
54		2.7190	−2.0339	1.5944										−65.4
60			−2.5486							0.9476				−70.0

Table 16: Point estimates for NML products (beta regression)

Interval	EAD	Interest Rate	DPD	Tenor	Req. amount	MOB	Due principal	Due interest	Principal	Interest	Due amount	AIC
0		−3.5808	−3.7917	1.1895		2.8365				−4.3850		−196000
6		−0.7727		0.5931		2.1413		−2.0596				−27684
12		−1.1711		0.2217		1.8590		−1.2612				−9219
18		−1.3018				1.3188						−7715
24		−1.6910	0.2743			1.4502						−3537
30		−1.1769				1.4664			3.2828			−2736
36		−0.9329				1.7543			3.7141			−2917
42		−0.8839				1.4680			2.3343			−1644
48		−0.8798			−8.7441	0.3529			10.2564			−1438
54		−0.9045	0.4408	1.2967	−7.6474				5.6963			−2668
60		−0.8611		1.8029	−11.1726				8.0010			−1772