Forecasting bank loans loss-given-default

João A. Bastos*

CEMAPRE, School of Economics and Management (ISEG)

Technical University of Lisbon, 1200-781 Lisboa, Portugal

Abstract

With the advent of the new Basel Capital Accord, banking organizations are invited to estimate credit risk capital requirements using an internal ratings based approach. In order to be compliant with this approach, institutions must estimate the expected loss-given-default, the fraction of the credit exposure that is lost if the borrower defaults. This study evaluates the ability of a parametric fractional response regression and a nonparametric regression tree model to forecast bank loan credit losses. The out-of-sample predictive ability of these models is evaluated at several recovery horizons after the default event. The out-of-time predictive ability is also measured for a recovery horizon of one year. The performance of the models is benchmarked against recovery estimates given by a naïve model in which predicted recoveries are given by historical averages.

JEL Classifications: G17; G21

1 Introduction

The new Basel Accord (Basel Committee on Banking Supervision, 2004) allows banking organizations to estimate credit risk capital requirements according to two distinct approaches: a standardize approach relying on ratings generated by external agencies for risk-weighting assets and an internal ratings based (IRB) approach which allows institutions to implement their own internal models to calculate credit risk capital, subject to supervisory review. In order to be compliant with the IRB approach, institutions must estimate the key parameters that determine the credit risk of a financial asset: (i) the probability of default over a one-year horizon, (ii) the expected loss-given-default, and (iii) the expected exposure at default. Loss-given-default is the credit that is lost by a financial institution when a borrower defaults, expressed as a fraction of the exposure at default. In the academic and practitioner literature, loss-given-default is frequently expressed by its complement, the recovery rate. Accurate estimates of potential losses are essential for an efficient allocation of regulatory and economic capital and for pricing the credit risk of debt instruments. Banks can gain competitive advantage by improving their internal loss-given-default estimates.

*jbastos@iseg.utl.pt

While the modeling of the probability of default has been the subject of many studies during the past few decades, a thriving literature on recovery rates only emerged recently, with the advent of the new Basel Capital Accord. Most of the published research pertains to recoveries of corporate bonds rather than loans (for a review, see e.g. Altman (2006)). Because loans are private instruments, few data is publicly available to researchers. Credit recoveries on bank loans are usually larger than those on corporate bonds. This difference may be attributable to the typically high seniority of loans with respect to bonds and the active supervision of the financial health of loan debtors pursued by banks (Schuermann, 2004). Empirical studies on bank loan losses report distributions and basic statistics, such as means and quartiles, examine the determinants of recoveries and the relationship between recoveries and the probability of default, and analyze the behavior of recoveries across business cycles. The geographic origins of studies on bank loan recoveries include the U.S. (Asarnow and Edwards (1995), Carty and Lieberman (1996), Carty et al. (1998), Hamilton and Carty (1999), Gupton et al. (2000), Frye (2003), Acharya et al. (2007)), Western Europe (Araten et al. (2004), Franks et al. (2004), Dermine and Neto de Carvalho (2006), Grunert and Weber (2009)) and Latin America (Hurt and Felsovalyi (1998)). Reported mean recoveries range from about 50 to 85% and the dispersion in recovery rates is generally high.

Although empirical work on credit losses has progressed, few studies have focused on forecasting recoveries. In fact, studies that attempt to model recovery rates rarely report the out-of-sample or out-of-time predictive accuracy of their models. However, for investors and lenders it is the performance of the models on unobserved data that is relevant. One of the few studies specifically developed to forecast recoveries can be found in the technical report of Moody's KMV LossCalc version 2.0 for dynamic prediction of loss-given-default (Gupton and Stein, 2005). This model was developed using 2,026 observations of defaulted loans, bonds and preferred stock occurring between 1981 and 2003. Recoveries were obtained from market prices of these instruments approximately one month after default. Recovery rates were modeled with a linear combination of the explanatory variables and the beta distribution. Model validation was performed out-of-time: the model was fit using data from one time period and tested on a subsequent period. Bellotti and Crook (2007) evaluated the performance of a variety of regression techniques in predicting recoveries on a large sample of credit card loans which were in default. The performance of the alternative models was evaluated out-of-sample using k-fold cross-validation. They showed that the fractional logit regression gives the best predictive accuracy in terms of mean absolute error. Gatti et al. (2008) examined 11,649 distressed loans to households and small and medium size companies. Loss given default was estimated from cash-flows recovered after the default event. Several models were tested in which loss-given-default is expressed as a linear combination of the explanatory variables. The estimation of model coefficients was achieved by ordinary least squares regression. The forecasting accuracy of the resulting regressions was evaluated out-of-time.

This paper builds on the work by Dermine and Neto de Carvalho (2006) to forecast recovery rates on a sample of bank loans using a parametric fractional response regression and a nonparametric and nonlinear regression trees model. Unlike ordinary least squares regression, the fractional response regression is particularly appropriate for modeling variables bounded to the interval $[0, 1]$, such as recovery rates, since the predictions are guaranteed to lie in the unit interval (Papke and Wooldridge, 1996). Regression trees are a powerful yet simple regression technique in which the predicted values of the target variable are obtained through a series of sequential logical if-then conditions. This sequence of binary splits divides bank loan observations into several partitions according to the loan characteristics. The objective of the splitting procedure is to divide the data into groups in which the recovery rate is as homogeneous as possible. The predicted recovery rate in a given partition is equal to the average recovery for the set of observations that lie in the partition. Regression trees resemble the "look-up" tables containing historical recovery averages that are commonly used by financial institutions for predicting credit losses. However, while the cells of "look-up" tables are defined rather arbitrarily by the analyst, the cells in a regression tree are defined by the data itself. It will be shown that the two models capture different aspects of the data. To the best of the author's knowledge this is the first study modeling recoveries using a nonparametric approach. Recovery rates are estimated using the full profile of cash-flows received by the bank after the loans became non-performing. The out-of-sample predictive accuracy of the models is evaluated using a k-fold cross-validation procedure. In contrast to previous studies, the out-of-sample accuracy is evaluated at several recovery horizons after the default event: 12, 24, 36 and 48 months. The out-of-time accuracy is also evaluated for a 12 months recovery horizon. The performance of the models is benchmarked against recovery estimates given by a naïve model in which the predicted recoveries are given by historical averages.

The remainder of this paper is structured as follows. The next section describes of the database of individual bank loans and the methodology that was employed to compute recovery rates. Section 3 presents the recovery rate models obtained with fractional response regressions for several recovery horizons. The determinants of the recovery rates are also discussed. Section 4 introduces regression trees and analyzes the models obtained with this technique. In Section 5, a comparison of the out-of-sample predictive accuracy of these models is presented. The out-of-time predictive accuracy for a 12 months recovery horizon is reported in Section 6. Finally, section 7 concludes the paper.

2 Data

In this section a brief characterization of the data employed in this study is given.[1] The database was provided by the largest private bank in Portugal, Banco Comercial Português. The sample contains 374 defaulted loans granted to small and medium size enterprises (SMEs) over the period June 1995–December 2000. All firms have turnovers greater that 2.5 million Euros. Table 1 shows information on the number of observations, loan amount and mean loan amount in each year. One can notice that the number of defaults is distributed evenly over the years. About half of the cases correspond to debt amounts of less than 50,000 Euros. Borrowing firms are classified into four broad groups according to their business sector: (i) real (activities with real assets, such as land, equipment or real estate), (ii) manufacturing, (iii) trade and (iv) services. To each individual loan is

[1] A comprehensive description of the data can be found in Dermine and Neto de Carvalho (2006).

attributed a rating by the bank's internal rating system. The rating reflects not only the probability of default of the loan but also the guarantees and collateral that support the operation. Good collateral and guarantees improve the operation's rating, resulting in lower contractual lending rates. There are seven classes of rating: A (the best), B, C1, C2, C3, D and E (the worst). These alpha-numeric rating notches were transformed into numeric values by an ordinal encoding that assigns the value 1 to rating A, the value 2 to rating B, and so on. Nearly half of the loans had no rating attributed. In order to avoid the exclusion of loans with missing rating class, which would reduce significantly the number of available observations, a surrogate rating equal to the mean value of the ratings in the sample was given to unrated loans. Fifty eight per cent of the loans are covered by personal guarantees. These are written promises that grant to the bank the right to collect the debt against personal assets pledged by the obligor. Fifteen per cent of the loans are covered by several varieties of collateral. These include real estate, inventories, bank deposits, bonds and shares. Thirty six per cent of loans are not covered by personal guarantees or any form of collateral. Other characteristics of individual loans that are included in the database are the contractual lending rate, frequency of default in the industry sector, the age of the borrowing firm and the number of years of relationship with the bank. The mean age of the firms is 17 years while the mean age of relationship with the bank is 6 years.

	Nr. of defaults	**Amount (Euros)**	**Mean amount (Euros)**
All years	374	52,687,032	140,874
1995*	65	5,950,097	91,540
1996	89	15,945,725	179,165
1997	59	12,479,614	211,519
1998	57	5,463,170	95,845
1999	47	7,013,532	149,224
2000	57	5,834,894	102,367

* second semester

Table 1: Number of defaulted loans, debt amount and mean debt amount, organized by year of default.

Two methodologies are usually found in the literature to calculate recovery rates. The first methodology considers the market price of the loan at the time of emergence from default. This approach is common in studies concerning recoveries on corporate bond defaults, but it can also be observed in studies of bank loan recoveries when a secondary market for defaulted loans is available (e.g., Gupton et al. (2000)) The second methodology considers the discounted value of cash or securities recovered during the bankruptcy's resolution process. These approaches may give significant differences in estimated recovery rates, that may be attributable to substantial risk premia demanded by the market of defaulted loans (Carty et al., 1998). Because there are no market prices of distressed loans in Portugal, the recovery rates were estimated using cash-flows recovered after default. The database includes the monthly history of cash-flows received by the bank after the loans became non-performing. For some loans, those of June 1995, a long recovery history of 66 months is available. As the default date approaches the end of year 2000, the recovery history is shortened. The discount rate that is used to compute the present value of the post-default cash-flows is the loan-specific contractual lending rate. Using the lending rate for discounting cash-flows one guarantees that the original loan balance and the present value of foregone interest and principal are equal. Recoveries were computed using Altman's mortality based approach (Altman, 1989). Denoting by $C_i(t)$ the cash flow recovered at the end of period t with respect to loan i, and by $L_i(t)$ the loan outstanding at time t, which includes principal and interest, the marginal recovery rate $R_i(t)$ in period t is simply given by

$$R_i(t) = \frac{C_i(t)}{L_i(t)}. \tag{1}$$

The cumulative recovery rate after time T is given by

$$r_i(T) = 1 - \prod_{t=1}^{T} \left[1 - R_i(t)\right]. \tag{2}$$

The cumulative recovery rate $r_i(T)$ represents the proportion of the debt outstanding at default that has been recovered, in present value terms, T months after default. Figure 1 shows the distributions of cumulative recovery rates for horizons of $T = 12$, 24, 36 and 48 months after the default event. The distributions are bimodal with many observations with low recovery and many with complete recovery. Bimodal distributions of bank loan recoveries are also found in Asarnow and Edwards (1995), Hurt and Felsovalyi (1998) and Franks et al. (2004). The frequency of complete recoveries increases with time as more cash-flows are collected in the work-out process.

Table 2 provides some basic statistics of the cumulative recovery rate distributions. The first row shows the number of observations with recovery history available for the respective horizon. For a recovery horizon of 12 months all defaults that occurred in the last available year (i.e., year 2000) are not considered, and likewise for other horizons. As expected the mean recovery rate increases with the recovery horizon as more cash-flows are collected. The median recovery rate exhibits a large jump when the horizon increases from 12 months to 24 months, as a consequence of the large increment in the number of near complete recoveries and the large reduction in the number of low recoveries. As Figure 1 shows and Table 2 confirms, there are no substantial differences between the recovery distributions for horizons of 24, 36 and 48 months. This is attributable to a marginal recovery rate that decreases with the horizon. In fact, a substantial part of the outstanding debt is recovered in the first year after default. One can also observe a reduction in the standard deviation of the recovery distributions as the horizon increases.

3 Fractional response regression

In this section, a parametric regression of the cumulative recovery rate as a function of the loan characteristics referred above is attempted. As Figure 1 shows, the recovery rate is restricted to the interval between 0 and 1. Due to the bounded nature of the dependent

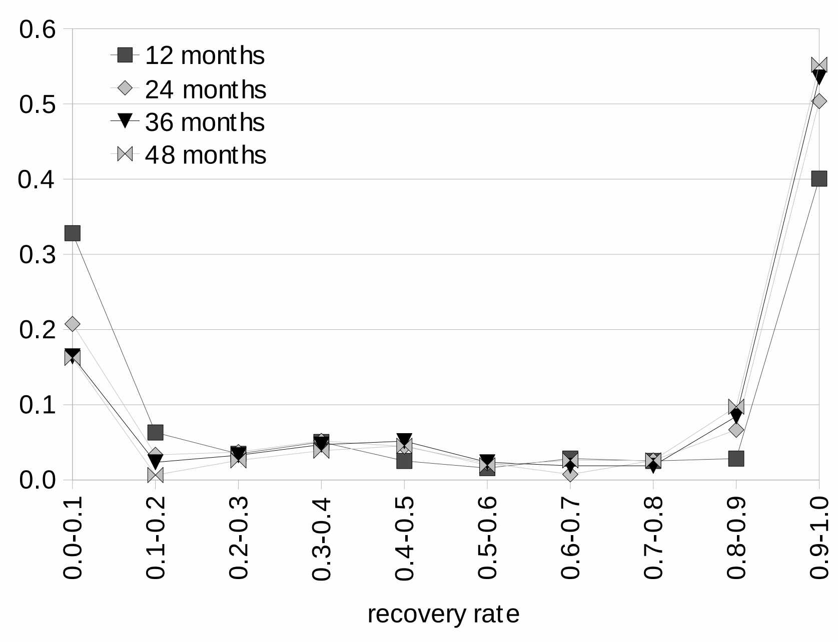

Figure 1: Distributions of the cumulative recovery rates for recovery horizons of 12, 24, 36 and 48 months after the default event.

	12 months	**24 months**	**36 months**	**48 months**
Number of observations	317	270	213	154
Mean	0.503	0.646	0.694	0.714
Median	0.493	0.907	0.946	0.950
Standard deviation	0.437	0.411	0.385	0.375

Table 2: Number of observations and mean, median and standard deviation of the recovery rate distributions for recovery horizons of 12, 24, 36 and 48 months.

variable one cannot implement an ordinary least squares (OLS) regression,

$$E(r|\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k = \mathbf{x}\beta, \tag{3}$$

since the predicted values from the OLS regression can never be guaranteed to lie in the unit interval. An alternative specification to equation (3) is

$$E(r|\mathbf{x}) = G(\mathbf{x}\beta), \tag{4}$$

where $G(.)$ satisfies $0 < G(z) < 1$ for all $z \in \mathbb{R}$. This guarantees that the predicted recovery rates lie in the unit interval. The most common functional forms for $G(.)$ are the cumulative normal distribution, the logistic function

$$G(\mathbf{x}\beta) = \frac{1}{1 + \exp(-\mathbf{x}\beta)}, \tag{5}$$

and the log-log function

$$G(\mathbf{x}\beta) = \exp(-\exp(-\mathbf{x}\beta)). \tag{6}$$

Because the recovery rate distributions in Figure 1 are slightly asymmetric, with more entries near complete recoveries than at low recoveries, the asymmetric log-log function may be more appropriate to this data than the symmetric cumulative normal distribution and logistic function. The non-linear estimation procedure consists of the maximization of the Bernoulli log-likelihood function (Papke and Wooldridge, 1996)

$$l_i(\hat{\beta}) \equiv r_i \log[G(\mathbf{x}_i\hat{\beta})] + (1 - r_i)\log[1 - G(\mathbf{x}_i\hat{\beta})]. \tag{7}$$

The quasi-maximum likelihood estimator of β is consistent and asymptotically normal regardless of the distribution of the recovery rate r_i conditional on $\mathbf{x}_i$ (Gourieroux et al., 1984).

Variable	**12 months**	**24 months**	**36 months**	**48 months**
Constant	0.870 (0.088)	1.590 (0.014)	1.872 (0.027)	1.578 (0.252)
Loan size	-0.610 (0.000)	-0.760 (0.000)	-0.836 (0.000)	-0.796 (0.001)
Collateral	0.183 (0.400)	0.625 (0.056)	0.699 (0.074)	1.708 (0.000)
Personal guarantee	-0.254 (0.107)	-0.399 (0.043)	-0.214 (0.363)	-0.425 (0.147)
Manufacturing sector	-0.233 (0.313)	-0.280 (0.298)	-0.361 (0.312)	-1.233 (0.017)
Trade sector	-0.216 (0.295)	-0.206 (0.408)	-0.668 (0.030)	-1.237 (0.007)
Services sector	-0.104 (0.676)	0.048 (0.882)	0.206 (0.633)	-0.276 (0.641)
Industry def. freq.	0.098 (0.101)	0.010 (0.895)	0.079 (0.343)	0.081 (0.433)
Lending rate	0.006 (0.714)	-0.007 (0.760)	-0.023 (0.411)	-0.007 (0.867)
Age of firm	0.001 (0.086)	0.001 (0.058)	0.001 (0.095)	0.001 (0.058)
Rating	-0.227 (0.002)	-0.212 (0.014)	-0.171 (0.104)	-0.048 (0.790)
Years of relationship	0.004 (0.127)	0.007 (0.017)	0.003 (0.335)	0.006 (0.250)
Wald test (χ^2)	57.91 (0.000)	104.1 (0.000)	119.7 (0.000)	116.4 (0.000)

Table 3: Model coefficients given by the fractional response regression for recovery horizons of 12, 24, 36 and 48 months after default. A log-log functional form was used.

In Table 3 one can find the model coefficients that were obtained for recovery horizons of 12, 24, 36 and 48 months after the default event. The p-values are shown in parenthesis. These results refer to the log-log functional form. The last row of Table 3 shows the Wald test for the hypothesis that the set of coefficients are jointly zero. This test shows that these models are statistically significant. Given the non-linearity of the functional forms $G(\mathbf{x}\beta)$, the partial effects of the explanatory variables on recovery rates are not constant. The partial effect of variable x_j on the recovery rate is given by:

$$\frac{\partial E(r|\mathbf{x})}{\partial x_j} = \frac{dG(\mathbf{x}\beta)}{d(\mathbf{x}\beta)}\beta_j. \tag{8}$$

Because $G(\mathbf{x}\beta)$ is strictly monotonic the sign of the coefficient gives the direction of the partial effects.

The first remark is that the size of loan has a statistically significant negative effect on recovery rates at all recovery horizons. This behavior of the recovery rates is also observed in the study of Hurt and Felsovalyi (1998). On the other hand, Asarnow and Edwards

(1995) and Carty and Lieberman (1996) find no relation between recovery rates and the size of loan. The negative effect of debt size on recovery rates might be explained by the reluctance of banks to foreclose large loans, hoping to preserve the option value of a future relationship (Dermine and Neto de Carvalho, 2006). The collateral dummy variable is statistically significant at 10% level for recovery horizons of 24, 36 and 48 months and, as expected, has a positive impact on recoveries. On the other hand, personal guarantees have a statistically significant, at 10% level, negative impact on recoveries for 12 and 24 months horizons. A tentative explanation for this counterintuitive result is that good clients may be exempt from providing personal guarantees. The business sector dummies are mostly not significant meaning that the regression cannot capture any industry-level variability. The industry default frequency and the contractual lending rate are also not significant. The age of firm has a statistically significant positive effect on recovery rates at 10% level. Older firms exhibit better recoveries. The rating is significant for 12 and 24 months recovery horizons at 5% level. The coefficient are in the direction one would expect: poor ratings have negative effect on recoveries. Finally, the age of the obligor's relationship with the bank is only significant for a 24 months horizon.

4 Regression trees

Regression trees (Breiman et al., 1984) are nonparametric and nonlinear predictive models in which the original dataset is recursively partitioned into smaller mutually exclusive subsets using a greedy search algorithm. A regression tree model is represented by a series of logical if-then conditions. Suppose one has a set of observations (i.e., bank loans) described by several attributes or characteristics and a target variable (i.e., recoveries). The algorithm begins with a root node containing all observations. Then it searches over all possible binary splits of all attributes for the one which will minimize the intra-subset variation of the target variable in the newly created daughter nodes. That is, in each daughter node the target variable will be more homogeneous than in the parent node. This procedure is repeated recursively for new daughter nodes until no further reduction in the variation of the target variable is achievable. Unsplit terminal nodes are denoted by "leaves" and are depicted by rectangles in the tree graphs shown below. The reduction of the target variation is measured by the "standard deviation reduction",

$$SDR = \sigma(T) - \frac{m(T_1)}{m(T)}\sigma(T_1) - \frac{m(T_2)}{m(T)}\sigma(T_2), \tag{9}$$

where T is the set of observations in the parent node and T_1 and T_2 are the set of observations in the daughter nodes that result from splitting the parent node according to the optimal attribute. The operators $m(.)$ and $\sigma(.)$ represent the mean and standard deviation of the target variable in the set. Starting from the root node, all observations flow down the tree and terminate their path in a leaf. The predictions are given by the mean value of the target variable for the set of observations in each leaf. Because the predictions are given by mean values of recoveries, they are inevitably bounded between 0 and 1, and, thus, regression trees are particularly appropriate for modeling recovery rates.

Very frequently the resulting trees are quite large and will overfit the data, giving good predictive accuracies on the data employed in the growth process but poor accuracies on new data. Improved accuracies on unobserved data can be obtained by "pruning" the tree after the basic growth process. The pruning procedure examines each node of the tree, starting near the bottom. An estimate of the expected variance of the target variable that will be experienced at each node for unobserved data is evaluated. If the variance of a subtree is greater than the variance of the parent node, then the parent node is pruned to a leaf. This process is repeated until pruning no longer improves the error.[2] Regression trees are not affected by the presence of outliers, since these are isolated into a node and have no further effect on the splitting. Furthermore, the resulting trees are invariant under monotone transformation of the explanatory variables. Regression trees are particularly suited to higher dimensionality problems and, as we shall see below, they can produce accurate results using only a few important explanatory variables.

The interpretation of regression trees is clear and easily understood. In the upper left corner of Figure 2 one can find the regression tree obtained for the recovery horizon of 12 months after default. This tree may be interpreted in the following way. First, it is asked if the loan amount (Debt) is greater than 56,000 Euros. If the answer is "yes", then the expected recovery rate is 0.3677 and the branch ends there. If the answer is "no" it is subsequently asked if the age of the client's relationship with the bank (AoR) is smaller than 65.5 months. If the answer is "yes" and the rating of the operation is A, B, C1 or C2, then the expected recovery is 0.6916; if the rating is C3, D or E then the expected recovery is 0.5073. If, on the other hand, the age of the client's relationship with the bank is greater than 65.5 months and the contractual lending rate is smaller than 12.4% then the expected recovery is 0.6031; if the lending rate is greater than 12.4% then the expected recovery is 0.8032.

Therefore, the following conclusions may be drawn. Smaller recoveries are expected for larger loan amounts since the recovery in the branch corresponding to loan amounts greater than 56,000 Euros (0.3677) is smaller than any of the remaining predicted recoveries present in the tree. This result is in agreement with the observation made above for the fractional response regression. For smaller loan amounts, one may expect greater recoveries if the age of the client's relationship with the bank is greater (0.8032 and 0.6031 for AoR $>$ 65.5 months versus 0.6916 and 0.5073 for AoR $<$ 65.5 months). If the loan amount is small and the age of relationship is large, the recovery will be larger if the contractual lending rate is higher. On the other hand, if the loan amount is small and the age of relationship is also small the recovery will be larger if the operation rating is good. Note that regression trees can capture effects on recovery rates due to explanatory variables that were statistically insignificant in the fractional response regression. Furthermore, a regression tree is conceptually similar to a "look-up" table, since recovery predictions are given by average recoveries. However, in a regression tree model the cells are defined by the data itself with no intervention by the analyst.

The tree for a recovery horizon of 24 months after default is simpler and is show in the lower left corner of Figure 2. Again, larger debt amounts result in smaller recoveries, but the threshold value (76,000 Euros) is larger than the respective value in the 12 months recovery horizon tree. If the loan amount is smaller than 76,000 Euros, than the expected recovery is higher if the age of the client's relationship is greater than 65.5 months. The

[2] Details of the specific implementation of regression trees employed in this work can be found in Wang and Witten (1996).

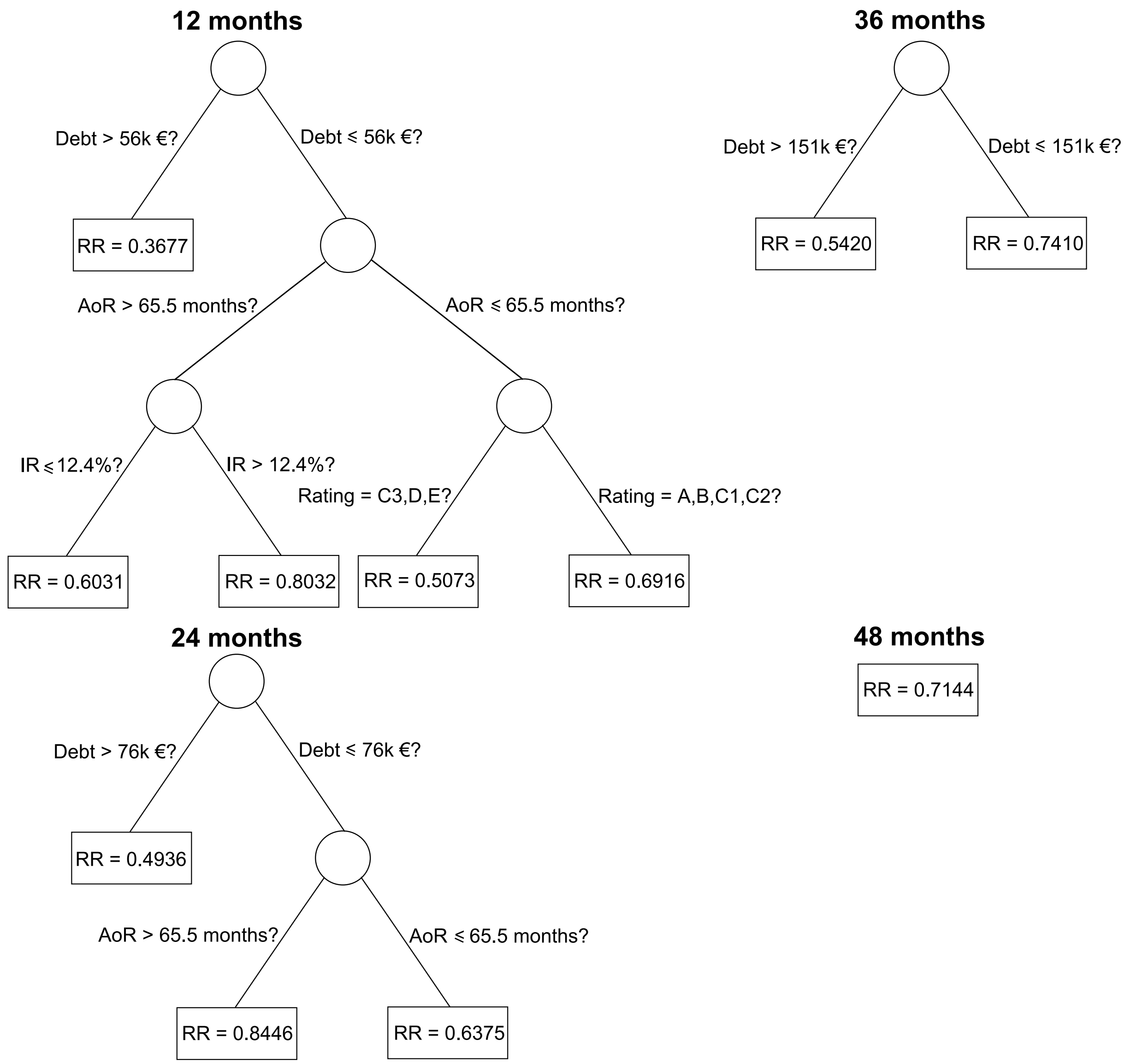

Figure 2: Regression trees for recovery horizons of 12, 24, 36 and 48 months.

tree for a 36 months recovery horizon is shown in the upper right corner of Figure 2 and is composed of a single binary split. Once more, larger debt amounts result in smaller predicted recoveries. For a recovery horizon of 48 months the tree is reduced to the root node and the expected recovery is equal to the mean value of the recoveries in the sample. That is, the tree model is equivalent to a naïve model in which the expected recovery is equal to the historical average. As the recovery horizon increases, tree structures gets simpler as a consequence of the increasing homogeneity of the recovery rates. As we shall see below, the predictive accuracy of regression trees with respect to alternative models is also deteriorated for longer horizons.

5 Out-of-sample predictive accuracy

The predictive accuracy of these models is assessed using two performance measures: the root mean squared error (RMSE) and the mean absolute error (MAE). The RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i}(r_i - \hat{r}_i)^2}, \tag{10}$$

where r_i and $\hat{r}_i$ are the actual and predicted recovery rates, respectively, on loan i and n is the number of loans in the sample. The MAE is defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i}\left|r_i - \hat{r}_i\right|. \tag{11}$$

Models with lower RMSE and MAE have smaller differences between the actual and predicted recoveries and predict actual recoveries more accurately. Note that a model chosen on the basis of the lowest RMSE, tends to give more conservative predictions of the recovery rate than a model that minimizes MAE, since large differences between actual and predicted recovery rates inflate the estimated RMSE.

Because the generated models may overfit the data, resulting in over-optimistic estimates of the predictive accuracy, the RMSE and MAE must be assessed on a sample which is independent from that used in the development of the models. A common technique to avoid over-fitting the data consists of randomly dividing the data sample in two sets: one set is used to fit the model and the other set is used to test its accuracy. This approach wastes valuable data since a large test sample is needed to evaluate accurately the prediction error. In order to develop models with a large fraction of the available data and evaluate the predictive accuracy with the complete dataset a 10-fold cross-validation was implemented. In this approach, the original sample is partitioned into 10 subsamples of approximately equal size. Of the 10 subsamples, a single subsample is retained for testing the model and the remaining 9 subsamples are used for building the model. The cross-validation process is repeated 10 times, with each of the 10 subsamples used exactly once as test data. The results from the 10 folds are then combined to produce a single estimate of the prediction error using the complete dataset.

To reduce variability, the cross-validation procedure was repeated 1,000 times using different randomly generated 10-fold partitions. Table 4 shows the average RMSE and MAE, over the 1,000 rounds, of the recovery rate predictions given by the fractional response regressions and the regression trees. The standard deviations are shown in parenthesis. For comparison, the last row of Table 4 shows the RMSE and MAE for a naïve model in which predicted recovery rates are obtained from simple historical averages. For the fractional response regression it is shown the results obtained using both the logistic (Equation 5) and log-log (Equation 6) functional forms. It can be observed that these functional forms do not exhibit substantial differences in forecasting performance. In terms of RMSE, the fractional response regressions yield poor results compared to historical averages. On the other hand, in terms of MAE, the performance of the fractional response regressions is superior to that of historical averages. Therefore, the choice of the best model depends on the measure that is used to evaluate the forecasting accuracy. This result corroborates the findings of Bellotti and Crook (2007) in their study on recoveries

	12 months		24 months		36 months		48 months	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Log-log	0.432	0.398	0.405	0.358	0.383	0.330	0.367	0.298
	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)	(0.005)	(0.004)
Logit	0.430	0.395	0.404	0.357	0.383	0.327	0.366	0.298
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)	(0.005)	(0.004)
Trees	0.408	0.363	0.389	0.338	0.388	0.337	0.374	0.324
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)		
Naïve	0.437	0.416	0.410	0.380	0.384	0.344	0.374	0.324

Table 4: Root mean squared errors (RMSE) and mean absolute errors (MAE) of the recovery rate estimates given by 10-fold cross-validation for recovery horizons of 12, 24, 36 and 48 months, and for the fractional log-log regression, the fractional logit regression, the regression trees and the naïve model. These numbers refer to average values over 1,000 random 10-fold partitions. The standard deviations are shown in parenthesis.

using a large set containing 55,000 observations of credit card loans. The introduction of interaction terms and of a shape parameter s, such that the original functional form $G(\mathbf{x}\beta)$ becomes $[G(\mathbf{x}\beta)]^s$, did not improve the performance of the fractional response regressions.

For recovery horizons of 12 and 24 months, the predictive accuracy of regression trees is better than those of the fractional response regressions and the historical averages. However, the performance of regression trees is degraded for larger horizons and the best model is the fractional response regression. Of course, for the horizon of 48 months the regression tree is equivalent to the naïve model, since the regression tree is reduced to a single leaf containing all observations.

6 Out-of-time predictive accuracy

In the previous section, the complete dataset was used to estimate the forecasting accuracy of the models via 10-fold cross-validation. However, this technique precludes the assessment of model accuracies over different periods of time. This section reports the forecasting accuracy of the models on an out-of-time sample. In this approach, the models are fit using data from one time period and the predictive accuracy is measured on a subsequent period. Given the constraint in the number of observations, only recovery rates with a 12 months recovery horizon are considered. The data was divided in two sets. The first set contains the loans that defaulted in years 1995, 1996 and 1997. This set comprises 213 observations and is employed in fitting the models. The second set corresponds to defaults that occurred in years 1998 and 1999. This set contains 104 observations and is used for testing the performance of the models.

Table 5 reports the RMSE and MAE of predicted recovery rates in the out-of-time sample, given by the fractional response regressions and the regression tree. Again, it is also shown the results for a naïve model in which the predicted recoveries for defaults occurring in years 1998 and 1999 are equal to the historical recoveries for defaults in the

Model	RMSE	MAE
Fractional log-log regression	0.502	0.460
Fractional logit regression	0.506	0.460
Regression tree	0.468	0.427
Naïve model	0.477	0.460

Table 5: Root mean squared errors (RMSE) and mean absolute errors (MAE) of the predicted recovery rates in the out-of-time sample (years 1998 and 1999) for a recovery horizon of 12 months, and for the fractional log-log regression, the fractional logit regression, the regression tree and the naïve model.

period 1995-97. The overall accuracy of the models is slightly worse than the respective accuracy obtained by 10-fold cross-validation for a recovery horizon of 12 months. This may be explained by the smaller amount of data that was employed to fit the models. Here, only approximately two-thirds of the data is used for fitting the models, while in the 10-fold cross-validation procedure 90% of the data is employed in each fold. The fractional response regressions give poor results and do not outperform the naïve model, both in terms of RMSE and MAE. On the other hand, the out-of-time predictive accuracy of the regression tree is better than those of the fractional response regression and the historical averages.

7 Conclusions

This study evaluated the performance of a parametric fractional response regression and a nonparametric regression tree model to forecast bank loan credit losses. Recovery rate models for several recovery horizons were implemented and analyzed. The regression tree models captured effects on recovery rates due to explanatory variables that were statistically insignificant in the fractional response regression. The forecasting accuracy of these models was evaluated using two different approaches: an out-of-sample estimation using the complete dataset with the help of a 10-fold cross-validation, and an out-of-time estimation in which the models are fit using data from one period and the accuracy is measured on a subsequent period. The performance of the models was benchmarked against predicted recoveries given by historical averages. When out-of-sample estimation was considered, the regression trees gave better results for shorter recovery horizons of 12 and 24 months, while the fractional response regression gave better results for longer horizons. For out-of-time estimation, the regression tree gave the best results. On the other hand, the fractional response regression did not outperform the naïve model based on historical averages.

The reader should note that this study is based on data from an individual bank in a single country and, therefore, these results will certainly capture some of the bank's intrinsic characteristics. In fact, Franks et al. (2004) reported that recovery rate distributions exhibit not only country-based differences, reflecting different insolvency regimes, but vary as well across banks within the same country. Nevertheless, the results shown here capture several empirical regularities found in previous studies on recovery rates.

The models tested here are parsimonious, give results that are easily interpretable, and are readily implemented. In particular, regression trees are an interesting alternative to the traditional statistical techniques commonly found in empirical studies on recovery rates.

Acknowledgments

I am very grateful to Cristina Neto de Carvalho for supplying the data for this study and for her invaluable help in the analysis. I would also like to thank João Santos Silva for his helpful comments and suggestions. This study was supported by Fundação para a Ciência e Tecnologia under program "Ciência 2007".

References

V.V. Acharya, S.T. Bharath, and A. Srinivasan. Does industry-wide distress affect defaulted loans? evidence from creditor recoveries. *Journal of Financial Economics*, 85: 787–821, 2007.

E.I. Altman. Default recovery rates and LGD in credit risk modeling and practice: An updated review of the literature and empirical evidence. *Working paper*, 2006.

E.I. Altman. Measuring corporate bond mortality and performance. *Journal of Finance*, 44(4):909–922, 1989.

M. Araten, M. Jacobs Jr., and P. Varshney. Measuring LGD on commercial loans: An 18-year internal study. *The RMA journal*, 4:96–103, 2004.

E. Asarnow and D. Edwards. Measuring loss on defaulted bank loans: A 24-year study. *Journal of Commercial Lending*, 77(7):11–23, 1995.

Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards. *Basel*, 2004.

T. Bellotti and J. Crook. Modelling and predicting loss given default for credit cards. *Working Paper, Quantitative Financial Risk Management Centre*, 2007.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees*. Wadworth International Group, Belmont, California, 1984.

L.V. Carty and D. Lieberman. Defaulted bank loan recoveries. *Moody's Investors Service*, 1996.

L.V. Carty, D.T. Hamilton, S.C. Keenan, A. Moss, M. Mulvaney, T. Marshella, and M.G. Subhas. Bankrupt bank loan recoveries. *Moody's Investors Service*, 1998.

J. Dermine and C. Neto de Carvalho. Bank loan losses-given-default: A case study. *Journal of Banking and Finance*, 30:1291–1243, 2006.

J. Franks, A. de Servigny, and S. Davydenko. A comparative analysis of the recovery process and recovery rates for private companies in the UK, france and germany. *Standard and Poor's report*, 2004.

J. Frye. A false sense of security. *Risk*, 16(8):63–67, 2003.

S. Gatti, S. Caselli, and F. Querci. The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Services Research*, 34(1):1–34, 2008.

C. Gourieroux, A. Monfort, and A. Trognon. Pseudo-maximum likelihood methods: theory. *Econometrica*, 52:681–700, 1984.

J. Grunert and M. Weber. Recovery rates of commercial lending: Empirical evidence for german companies. *Journal of Banking and Finance*, 33:505–513, 2009.

G.M. Gupton and R.M. Stein. LossCalc V2: Dinamic prediction of LGD. *Moody's Investors Service*, 2005.

G.M. Gupton, D. Gates, and L.V. Carty. Bank loan loss given default. *Moody's Investors Service*, 2000.

D.T. Hamilton and L.V. Carty. Debt recoveries for corporate bankruptcies. *Moody's Investors Service*, 1999.

L. Hurt and A. Felsovalyi. Measuring loss on latin american defaulted bank loans: A 27-year study of 27 countries. *Journal of Lending and Credit Risk Management*, 80: 41–46, 1998.

L.E. Papke and J.M. Wooldridge. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11:619–632, 1996.

T. Schuermann. What do we know about loss given default. *Wharton Financial Institutions Center Working Paper No. 04-01*, 2004.

Y. Wang and I.H. Witten. Induction of model trees for predicting continuous classes. *Working paper, Department of Computer Science, University of Waikato*, 1996.