A Proposed Framework for Backtesting Loss Given Default Models

Gert Loterman[a], Michiel Debruyne[b], Karlien Vanden Branden[b]
Tony Van Gestel[b], Christophe Mues[c]

[a] *Ghent University, Management Information and Operations Management, Tweekerkenstraat 2, 9000 Ghent, Belgium*
gert.loterman@ugent.be
[b] *Dexia, Credit Risk Modelling, Marsveldplein 5, 1050 Brussels, Belgium*
{michiel.debruyne,karlien.vandenbranden,tony.vangestel}@dexia.com
[c] *University of Southampton, Management School, Southampton S0171BJ, United Kingdom*
c.mues@soton.ac.uk

Abstract

The Basel accords require financial institutions to regularly validate their LGD (Loss Given Default) models. This is crucial so banks are not misestimating the minimum required capital to protect them against the risks they are facing through their lending activities. The validation of an LGD model typically includes backtesting which involves the process of evaluating to which degree the internal model estimates still correspond with the realized observations. Reported backtesting examples have typically been limited to simply measuring the similarity between model predictions and realized observations. It is however not straightforward to determine acceptable performance based on these measurements alone. Although recent research led to advanced backtesting methods for PD models, the literature on similar backtesting methods for LGD models is much scarcer. This study addresses this literature gap by proposing a backtesting framework using statistical hypothesis tests to support the validation of LGD models. The proposed statistical hypothesis tests implicitly define reliable reference values to determine acceptable performance and take into account the number of LGD observations as a small sample may

affect the quality of the backtesting procedure. This workbench of tests is applied to an LGD model fitted to real-life data and evaluated through a statistical power analysis.

1 Introduction

Banks are required to validate the internal estimation process and their internal models so as to prove their soundness to the national regulator [38]. The validation of the estimation process involves issues like data quality, reporting and problem handling and how the predictive models are used by the bank; it is mainly qualitative in nature, although quantitative methods are useful for the examination of data quality. The validation of the models on the other hand includes both the examination of the model design and the predictions that each such model produces for the key risk parameter it is modeling: Probability of Default (PD), Exposure at Default (EAD), or Loss Given Default (LGD), i.e. the percentage of the loan that the bank will not be able to recover in the event of a default. The evaluation of the model design consists of a qualitative review of the statistical techniques and the relevance of the data used to build the model. The assessment of a model's predictions typically includes quantitative methods such as benchmarking and backtesting.

While benchmarking methods evaluate the internal model estimates against (where available) external model estimates [35], backtesting methods evaluate the internal model estimates against the actual realized observations. The purpose of backtesting is to evaluate the predictive performance of a model and how this evolves over time, in order to detect model deterioration in a timely manner. An LGD model can experience reduced predictive performance when current loan loss behavior no longer reflects the previous loan loss behavior on which the model was originally built. This may lead to an overestimation or underestimation of a bank's required minimum capital so that its operations can become less profitable or more risky, respectively. Although banks are required to regularly validate their models in order to be Basel-compliant, the accord does not mention how to perform this validation [38]. In addition, recent research has largely focused on advanced methods for backtesting PD models [21, 23, 26] but literature on comparable methods for backtesting LGD models is virtually non-existing.

Current LGD performance evaluation practices found in the literature have so far been usually limited to comparing internal LGD predictions and realized LGD observations using error-based metrics, correlation-based metrics or even classification-based metrics [35]. It is however not straightforward to determine acceptable performance solely based on these metrics. A single value has little meaning without an appropriate reference value indicating acceptable accuracy. Additionally, these metrics do not take into account the number of LGD observations. When the portfolio lacks sufficient observations, a few extreme observations can distort the accuracy result and thus undermine its reliability. This study therefore proposes a backtesting framework in which the model performance on an out-of-time test data set is evaluated against earlier model performance, e.g. on the training data, using a number of suggested statistical hypothesis tests. Hence, an appropriate reference value is introduced for each metric of interest that takes into account the number of observations.

The remainder of this paper is organized as follows. First, a literature review is conducted on empirical LGD studies that focus on the evaluation of the predictive performance of LGD models. Second, the key idea of the proposed backtesting procedure is explained together with our workbench of available statistical hypothesis tests to evaluate LGD models. Third, the experimental set-up to apply and evaluate the backtesting framework is described. This involves information about the employed real-life LGD data, the design of a predictive LGD model based on this data, a statistical significance analysis of the measured predictive model performance and a statistical power analysis of the proposed tests based on these performance metrics. Fourth, the results of the application and the evaluation of the backtesting procedure are reported and discussed.

2 Literature review

The Basel accords require banks to backtest their internal models but do not further specify how this needs to be performed [38]. Current backtesting practices in the empirical LGD literature are usually limited to comparing internal LGD predictions and realized LGD observations with error-based metrics

(e.g. MAE, RMSE), correlation-based metrics (e.g. Pearson's r, Kendall's τ, Spearman's ρ, coefficient of determination R^2) or even classification-based metrics (e.g. AUROC) [35]. Each of these metrics has its own method of quantifying the degree of similarity between LGD model predictions and the actual realized observations. This section describes the workings of these metrics more in detail and explains how they are used to assess the predictive performance of LGD models. It will conclude by identifying several problems when using these metrics for the purpose of backtesting LGD.

Error-based metrics quantify the error or difference between predicted and observed values. One of the most often used error-based metrics is the Mean Squared Error (MSE) [12, 17, 31]. The MSE is defined as the average of the squared differences between loan-level LGD predictions and actual observed values. Since errors are squared, this metric heavily weights outliers. The metric is bound between the maximum squared error and zero (perfect prediction). The Root MSE (RMSE) is also often used as a metric in the literature [9, 11, 16]. The RMSE is merely the square root of the MSE but offers the additional advantage that it has the same unit scale as the dependent variable being predicted, unlike MSE. Another error-based metric used in the literature is the Mean Absolute Error (MAE) [9, 11, 17]. The MAE is given by the averaged absolute difference between predicted and observed values. Just like the RMSE, the MAE has the same unit scale as the dependent variable being predicted, but MAE is not as sensitive to outliers. The metric is bound between the maximum absolute error and zero (perfect prediction).

Correlation-based metrics quantify the degree of some statistical relationship between predicted and observed values. A very popular correlation-based metric seems to be the R^2 [12, 20, 24, 30]. The R^2 can be defined as one minus the fraction of the sum of squared errors to the variance of the observations. Since the second term in the formula can be seen as the fraction of unexplained variance, the R^2 can be interpreted as the fraction of explained variance. Although R^2 is usually a number on a scale from zero to one, R^2 can yield negative values when the model predictions are worse than using the mean $\overline{y}$ from the training set as prediction. Other correlation-based metrics include Pearson's r [31], Spearman's ρ [35] and Kendall's τ [22]. Pearson's r measures the degree of linear relationship between predictions and observations. Spearman's ρ is defined as Pearson's r applied to the rank-

ings of predicted and observed values. Kendall's τ measures the degree of correspondence between predictions and observations. All three correlation coefficients can take values between minus one (perfect negative correlation) and one (perfect positive correlation) with zero meaning no correlation at all.

Although not considered to be a metric to assess the performance of a regression model, a typical binary classification-based metric such as the Area Under the Receiver Operating Characteristic curve (AUROC) [27] is also used in the LGD literature [22, 31, 30]. It is employed in an LGD context to measure how good an LGD regression model is able to distinguish between high and low losses. In order for the curve to be produced, the observed values are first dichotomized into a high and a low class using e.g. the mean $\overline{y}$ of the training set as the cut-point. The area under the ROC curve is an estimate for the discriminatory power of a model. The metric varies from 0.5 (random classification) to one (perfect classification). Another similar metric is the Area Over the Regression Error Characteristic curve (AOREC) [14]. It can be seen as either a generalization of an error-based metric or a regression equivalent for the AUROC. The AOC curve plots the error tolerance on the x-axis onto the percentage of points predicted within that tolerance (or accuracy) on the y-axis. The resulting curve represents the cumulative distribution function of the squared error. The area over the REC curve (AOC) is an estimate of the prediction error by the model. The metric is bound between zero (perfect prediction) and the maximum squared error.

The evaluation scheme used to assess the predictive performance of a LGD model varies in the literature. For prediction it is important that the model performance is evaluated on unseen cases which is what it will also encounter in real-life. These evaluation schemes are called out-of-sample. In an out-of-sample schema [9, 11, 22, 20, 30], the LGD dataset is split into a random training set (e.g. two-thirds of the total dataset) and a test set (remaining one-third of the total dataset). The training set is used to build the model; the test set is used to evaluate the model. In order to enhance the reliability of the assessment, multiple hold-out validations can be considered [9, 11]. Alternatively, rather than using a simple out-of-sample test set, one can also opt for an out-of-time scheme. In an out-of-time scheme [9, 12, 16, 17, 31], the model itself is built on data from a specific time period and is evaluated on data collected after this time period. While an average of multiple hold-out validations is most applicable to assess how well a technique fits a model

to a dataset, out-of-time validation adds additional insights into real-life predictive model performance as the model is strictly built using historical data and strictly evaluated on future data. Backtesting always comes down to an out-of-time evaluation.

The use of the above-described metrics for backtesting an LGD model may present problems. First, it is not straightforward to determine acceptable model performance solely based on these metrics. A single value has little meaning without an appropriate reference value indicating acceptable performance. For example, a loan level R^2 of 50% may look poor on paper since a perfect LGD model should in theory yield an R^2 of 100%. However, comparing this performance with other real-life LGD benchmarking results where the average R^2 ranges from 4% to 43% [35], this may sound very good. Similarly, for error-based metrics, it is useful to employ, for example, the relative squared error or relative absolute error which measures the model's predictive performance compared to the predictive performance of historical average LGDs which may be seen as a reference model (null model) [10]. Second, the above-described metrics do not take into account the number of LGD observations. When the portfolio lacks sufficient observations, a small number of extreme observations can distort the accuracy results and thus affect its reliability. For example, when assessing LGD model performance in a specific year with only ten defaults in the portfolio, one or two large loan-level prediction errors may cause a disproportionately low performance.

3 Proposed backtesting framework

The proposed approach for backtesting the predictive performance of an LGD model is to evaluate model performance on the most recent out-of-time validation set collected (referred to as 'test performance' from here on) against model performance on the original training dataset[1] (referred to as 'training performance' in what follows); to conduct this comparison, we will suggest a series of alternative statistical hypothesis tests. By comparing test perfor-

[1]Additionally, one could also decide to backtest against a validation sample from another past time period (e.g. using predicted and observed data collected as part of a previous backtesting exercise as a reference); the statistical tests used would be similar though.

mance against training performance, a reference value is introduced, tailored to the respective model. Model deterioration is thus defined as a decrease of model performance compared to the performance during model building (or some other reference period). Note that this is in contrast to the process of benchmarking where the performance of multiple models is compared to each other. By applying statistical hypothesis tests, model deterioration can be statistically detected at a pre-defined significance level (e.g. 5%). In addition, statistical hypothesis tests implicitly take into account any insufficient number of observations (i.e. sample size) to prevent incorrect judgements.

In what follows, the proposed statistical hypothesis tests to decide upon acceptable model performance are explained. These tests typically start with the formulation of a null hypothesis, H_0, which assumes no model deterioration and an alternative hypothesis, H_a, which indicates model deterioration. Then, some test statistic is identified in order to assess H_0. A decision whether or not to reject H_0 can be made by calculating this test statistic for the sample at hand and comparing it to the critical value corresponding to a significance level of 5%. If the resulting test statistic falls in the rejection region (e.g. is greater than the critical value), H_0 may be rejected in favor of H_a; i.e., one would accept there is sufficient evidence to support model deterioration.

3.1 Central tendency error tests

The most basic model performance aspect is the *central tendency* of the error; the corresponding metrics are useful in assessing so-called model *calibration*, i.e. whether the model tends to under- or over-estimate the true LGD of loans. The error E for loan defaults in the test set is defined here as the difference between observed LGD, Y, and predicted LGD, $\hat{Y}$; thus, $E = Y - \hat{Y}$. Two well-known statistical hypothesis tests from the literature may be used in this context: the T test and the Wilcoxon signed rank test. Both tests allow one to evaluate whether the central tendency of the error equals zero, which serves as the reference value. In other words, it is assumed that the central tendency of the training error, E_t, of a well-aligned model equals zero. Whereas the T test compares the mean error to zero, the Wilcoxon signed rank test compares the median error to zero.

Note that one-tailed tests will be used instead of two-tailed tests because the former provide more power to detect whether model predictions are too low on average by not looking for systematic misestimations on either side. Although overestimating LGDs may needlessly increase a bank's capital requirements, detecting any systematic underestimation of losses is considered more important since the primary regulatory concern is that the bank would not have set aside sufficient capital. Nonetheless, a bank may become less profitable compared to other banks when their capital requirements are significantly overestimated so, where this would be a key concern, it may be beneficial to consider two-tailed versions of the proposed tests instead.

The T test can be used to make inferences about whether the mean of the test-set error μ_E equals zero or (provided that one opts for a one-tailed test) is positive:

$$H_0 : \mu_E = 0,\ H_a : \mu_E > 0$$

A test statistic T can be derived from the property that the sample mean of a normally distributed variable is normal or, in the absence of normality, approximately normal for a large enough sample (cf. the Central Limit Theorem). Hence, given that H_0 is true, and with the true variance of the error being unknown, the following test statistic follows a t-distribution:

$$T = \frac{\bar{e}}{\frac{s_e}{\sqrt{n}}} \sim t_{n-1}$$

with n the number of loss observations available for backtesting. Note that as n becomes larger (e.g. starting from $n > 30$), a t-distribution converges to a normal distribution. Hence, in that case, performing a Z test and comparing the test statistic against a normal distribution table would be an appropriate alternative.

The one-sample Wilcoxon signed rank test [42] on the other hand can be used for making inferences about whether the median of the test-set error, η_E, equals zero:

$$H_0 : \eta_E = 0,\ H_a : \eta_E > 0$$

A test statistic can now be derived by ranking the absolute values of non-zero errors in ascending order. The smallest error is ranked 1, the second

smallest is ranked 2, etc. Tied cases (i.e. absolute LGD prediction errors of the same magnitude) are assigned the average of their ranks. The test statistic is then given by the sum of the ranks of the positive errors, r_+, and is approximately normal under H_0 and for a large enough sample; hence, one can use the following standardized test statistic:

$$Z = \frac{r_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)^2$$

Compared to the T test which draws conclusions based on the actual value of the mean in the test sample and its assumed distribution, the Wilcoxon test statistic is a non-parametric alternative which looks solely at the ranking of the errors. Nonetheless, in order to be able to easily quantify and compare the central tendency error over the test years, we propose an additional metric that will be referred to in this paper as the Wilcoxon metric w_r. It is defined as the ratio of the rank sum of negative errors (r_-) to the total rank sum of positive and negative errors ($r_+ + r_-$). It is bound between zero (i.e. all errors are underestimations) and one (all are overestimations) with 0.5 indicating there is no upward or downward bias.

3.2 Error dispersion tests

Another model performance characteristic that complements the central tendency of the error is the *dispersion* of the error, or, inversely, model *precision*. Whereas the central tendency tests proposed in the previous section look for systematic under- (or over-)estimations of LGD, dispersion tests are meant to detect whether this error distribution is getting wider; i.e. loan-level predictions are becoming less precise. Two existing statistical hypothesis tests may be used for this purpose: the F test (a well-known parametric test) and the Ansari-Bradley test (a non-parametric alternative). Both tests allow one to evaluate whether the dispersion of the error in the most recently collected test set differs from the dispersion of the training error, which serves as a reference. While the F test compares the variance of the test set error to the variance of the training error, the Ansari-Bradley test compares the spread of both distributions by using a ranking procedure rather than relying on

[2]Because r_+ takes only integer values, a continuity correction may be applied.

the numerical error values directly. Note that, similarly to before, one-tailed tests are proposed in order to enhance the statistical power to detect when the dispersion of the test error is larger than the dispersion of the training error. A larger error dispersion implies larger prediction errors; this loss of predictive power might impact the model's ability to correctly identify the LGD risk of individual loans or to produce sufficiently homogeneous LGD risk grades.

The F test [43] can be used to determine whether the variance of the test set error σ_E^2 is equal to the variance of the training error $\sigma_{E_t}^2$, i.e. for a one-tailed test:

$$H_0 : \sigma_E^2 = \sigma_{E_t}^2,\ H_a :\ \sigma_E^2 > \sigma_{E_t}^2$$

A test statistic is produced by inspecting the ratio of observed error variance in the test sample, s_e^2, over observed error variance in the training sample, $s_{e_t}^2$. Assuming the error terms are sampled from an underlying normal distribution, $(n-1)\frac{s_e^2}{\sigma_E^2}$ and $(n_t-1)\frac{s_{e_t}^2}{\sigma_{E_t}^2}$ follow a χ^2-distribution, with $n-1$ and n_t-1 degrees of freedom, respectively (where n is the number of loan defaults to backtest and n_t the number of defaults in the training (reference) set). Hence, dividing each by the corresponding degrees of freedom and taking the ratio leads to the following F-distributed test statistic, under H_0:

$$F = \frac{s_e^2}{s_{e_t}^2} \sim F_{n-1,n_t-1}$$

Note that deviations from the normality assumption could undermine the validity of this test. Therefore, we propose a second, non-parametric test.

Alternatively, the Ansari-Bradley test [5] can be used to assess whether the cumulative error distribution for the test set, $F_E(u)$, and the cumulative distribution function of the training errors, $F_{E_t}(u)$, are equal, assuming they can only differ in the value of a scale parameter θ:

$$H_0 : F_E(u) = F_{E_t}(u),\ H_a : F_E(\theta u) = F_{E_t}(u)\ with\ \theta > 1$$

In this setting, a test statistic can be derived by calculating the sum of rank scores or weights of the ordered errors in the combined sample containing

both test and training errors, e and e_t; let the size of this sample be $m = n + n_t$. The weights assigned are one to both the smallest and largest error value in the combined sample (i.e. the 'outer edges' of the empirical error distribution), 2 to the next smallest and next largest, etc., until a weight of $\frac{m}{2}$ is assigned to the two middle observations if m is even, or $\frac{m+1}{2}$ to the one middle observation if m is odd (using mid ranks for ties). The test statistic is given by the sum of these weigths (denoted w_e) for the ordered errors e associated with the test set only. For large sample sizes, w_e is asymptotically normally distributed, specifically[3]:

$$Z = \frac{w_e - \frac{n(m+2)}{4}}{\sqrt{\frac{nn_t(m+2)(m-2)}{48(m-1)}}} \sim N(0,1)$$

when m is even, or:

$$Z = \frac{w_e - \frac{n(m+1)^2}{4m}}{\sqrt{\frac{nn_t(m+1)(3+m^2)}{48m^2}}} \sim N(0,1)$$

when m is odd. A lower-tail test is then used to detect larger dispersion in the test sample. As the test requires that E and E_t have identical population medians, Ansari and Bradley [5] recommend subtracting the sample medians and shifting both e and e_t to zero median if this assumption should not be met.

Compared to the F test which draws conclusions based on the actual values of training and test sample variances, the Ansari-Bradley test statistic uses a ranking procedure to determine whether the test sample error distribution is wider than that previously observed in the training (reference) data. Nonetheless, to be able to easily quantify and compare the test performance over the out-of-time validation period at hand, relative to the training performance, we propose an additional metric that will be referred to in this paper as the Ansari-Bradley metric ab_w. This metric is defined as the ratio of the sum of weights of the ordered errors in the combined sample associated

[3]A further modification may be applied to the test statistic variance in this large-sample approximation if ties are present.

with e (w_e) to the total sum of weights in the combined sample associated with both e and e_t ($w_e + w_{e_t}$). Values closer to zero (one) imply greater (smaller) error dispersion in the test data, respectively; 0.5 indicates similar error dispersion in training and test set.

3.3 Error-, correlation- and classification-based tests

In addition to our proposed tests for monitoring model calibration and precision, a number of other metrics are frequently used in the empirical LGD literature to assess model performance. These are error-based (i.e. $RMSE$, MAE, $AOREC$), correlation-based (i.e. R^2, r, ρ, τ) or classification-based (i.e. $AUROC$) metrics. However, the backtesting literature has not always identified readily available statistical hypothesis tests for them or described how to apply these to the problem of detecting model deterioration. The main problem is that it is often not straightforward to determine the theoretical distribution of a test statistic under a null hypothesis based on these metrics. Instead, such a distribution may be estimated via a bootstrapping approach. The basic idea of bootstrapping is that inference about a population can be made by resampling from the available sample data. By doing so, one can produce an empirical distribution for a test statistic under a given null hypothesis when its true distribution is unknown.

A bootstrap test can thus be used to determine to what degree, for a metric of interest, the test performance P is equal to the training performance P_t:

$$H_0 : P = P_t,\ H_a : P < P_t$$

In this case, the test statistic is given by $P_t - P$ if P is one of the commonly used correlation- or classification-based metrics, or $P - P_t$ if the metric of interest is an error-based one (since their values are inversely related to model performance). The distribution of this test statistic under the null hypothesis can be simulated through bootstrapping according to Beran's algorithm [13, 33, 41, 44]. First, the training and test observations, along with their predicted LGD values, are pooled into one larger sample. Next, a training/test bootstrap sample with the same length as the original training/test set is extracted from this pool of observation/predictions through random sampling with replacement. Then, the difference in value for the metric

under consideration between the bootstrap training sample and bootstrap test sample is calculated. This procedure is repeated 1000 times in order to empirically build up the distribution of the test statistic under the null hypothesis. Note that again only one-tailed tests are proposed so as to enhance the statistical power to detect performance deterioration; however, the method can easily be adapted to implement two-tailed tests where needed.

4 Methods

This section evaluates the proposed backtesting framework by applying it to an example LGD model fitted to and tested on real-life data. The experimental set-up is as follows. First, real-life loss data was collected consisting of a variety of characteristics of each respective loan on the one hand and its corresponding observed LGD on the other. Second, a regression analysis is performed over the loss data in order to build a predictive LGD model. Third, the performance of the predictive LGD model is backtested on multiple years of out-of-time data. To this end, the proposed statistical hypothesis tests are run in order to discover any significant model deteriorations. Fourth, the proposed statistical hypothesis tests are empirically evaluated through a statistical power analysis.

4.1 Data collection

The real-life LGD dataset collected in this study consists of corporate loan losses over a time span from 1984 to 2004 and contains 891 observations. Data from 2001 to 2004 is used to annually backtest the constructed LGD model. The model is built with data from 1984 to 2000. This split between training and test data (i.e. letting the training window run to the year 2000) is chosen so as to have sufficient data (about 500 defaults) to train an LGD model while still having sufficient time periods (i.e. four years) to backtest the LGD model. The number of observations used for training and backtesting purposes is given in Table 1.

The empirical distribution of the LGD data used for training and testing is shown in Figure 1. It appears to be predominantly J-shaped with the highest observed frequencies at the right end of the LGD value range. This means

Year	Observations	Purpose
2004	30	Backtesting
2003	47	
2002	140	
2001	155	
1984-2000	519	Training

Table 1: Number of observations

that the dataset is characterized by high LGDs for a majority of defaults. Notice that especially 2001 and 2002 are characterized by high LGDs while this shifts to generally lower LGDs for 2003 and 2004. From the literature, we know that the LGD distribution is indeed typically non-normal and often bimodal; real-life LGD tends to be characterized by high concentrations of either (near-)total recovery (LGD = 0 or close to 0) or total loss (LGD = 1) or both. The majority of the empirical LGD literature reports a large peak at zero and a smaller peak on one [9, 17, 22, 24, 31]. Nonetheless, a few studies also report what we observe in our dataset: a large peak on one and a smaller or non-existing peak on zero [20, 31].

The LGD dataset covers both loans and bonds from large corporates. Apart from the LGD target variable, the dataset includes 42 variables which represent potential LGD drivers, among others rating, level of seniority, country of domicile, type of industry, US default rate. The data covers different sectors such as transportation, finance, public, industrial and real estate. Firms are domiciled in America, Europe and Oceania. The average size of the debts is about $100 million and about 15% of the debts are secured by collateral. For the purpose of predictive modeling, a few pre-processing actions are performed. Continuous variables are transformed to the standard z-score using the sample mean and standard deviation of the training set. Furthermore, categorical variables are quantified by dummy encoding. More information about this dataset is confidential.

4.2 Predictive modeling

First of all, a predictive LGD model is required to estimate future outcomes. This allows the bank to protect itself against default losses whilst remaining

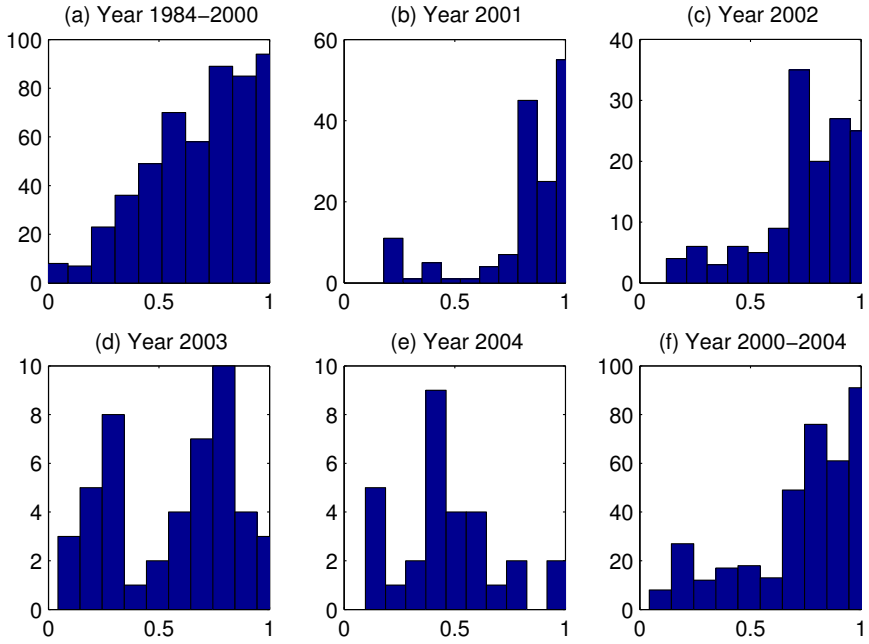

Figure 1: LGD observations histogram

competitive. A second consideration is that the bank may need to provide a comprehensible LGD model typically required by the national regulators in order to ensure that banks fully understand their risks and underlying model relations. Although non-linear models such as Support Vector Machines and Artificial Neural Networks seem to show significantly higher performance on average than linear models in a recent benchmarking study, they are often labelled as black-box models [35]. Therefore, a simple linear model is deliberately chosen in this paper so that the model form remains understandable and its backtesting results can be more easily interpreted. Note that, in the context of this study, the focus is not on building the best possible model, but to illustrate how a given model can be properly backtested.

The LGD model is estimated by applying Ordinary Least Squares (OLS) regression to the training data. In order to improve the generalization ability, i.e. the ability to accurately estimate the LGD on out-of-sample data, a

variable selection method is used to exclude irrelevant or redundant variables from the model. Using a ten-fold cross-validation scheme, a model wrapper searches for a subset of variables that best predicts the LGD by sequentially selecting variables until there is no improvement in minimizing the sum of squared differences between predictions and observations. The selected subset includes two binary variables referring to the level of seniority, i.e. senior unsecured (SU) (true/false) and junior subordinated (JS) (true/false), and one continuous variable, i.e. (standardized) US default rate from the previous year (USDR(t-1)). The output of the variable selection strengthens previous literature studies which stress the importance of seniority and default rate as major predictive drivers [38]:

$$\mathbf{LGD} = 0.74 - 0.15 \cdot \mathbf{SU} + 0.18 \cdot \mathbf{JS} + 0.02 \cdot \mathbf{USDR(t\text{-}1)}$$

The resulting linear model can be interpreted as follows. The baseline LGD estimate is 74%; this estimate decreases with 15% when the loan is senior unsecured or increases with 18% when the loan is junior subordinated (keeping the other variables constant). Similarly, the LGD increases with the US default rate from the previous year. These relations are roughly in line with previous empirical studies. Secured debt and high priority are known to decrease the LGD [1, 2, 3, 4, 6, 7, 18, 19, 25, 29, 32]. Also, LGD was reported to be higher in periods of high defaults [3, 4, 6, 28, 29, 32, 34].

4.3 Significance analysis

Table 2 gives an overview of the performance metrics on which the statistical hypothesis tests of the proposed backtesting framework are based. The name of each performance metric is given in column one while the values in columns two and three show its lower and upper bound. The first two metrics specifically measure the central tendency of the error while the subsequent two metrics measure the dispersion of the error. As explained in sections 3.1 and 3.2, standard (non-)parametric tests are available to test performance deterioration in terms of these metrics. The following eight metrics are a further selection of error-, correlation- and classification-based metrics. To detect performance deterioration based on these metrics, we will use the bootstrapping procedure outlined in section 3.3.

Metric	Worst	Best
$\overline{e}$	-∞	0
w_r	0	0.5
s_e^2	$+\infty$	0
ab_w	0	0.5
RMSE	$+\infty$	0
MAE	$+\infty$	0
AUROC	0.5	1
AOREC	$+\infty$	0
R^2	0[4]	1
r	0	1
ρ	0	1
τ	0	1

Table 2: Performance metrics

To monitor whether the performance according to each metric falls within an acceptable range, the out-of-time performance is compared with the training performance. Each statistical hypothesis test assumes a null hypothesis and if sufficient evidence exists against the null hypothesis, one accepts the alternative hypothesis, i.e. that performance has been affected. This evidence is gathered in the form of a p-value. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. When the resulting p-value is compared to a pre-defined significance threshold, a decision can be made on statistical significance. This pre-defined level is the maximum allowed probability of making a type I error (i.e. the incorrect rejection of the null hypothesis). This is generally denoted as α and can e.g. be set to 5%. Low p-values (i.e. <5%) indicate that H_0 can be more confidently rejected, whereas larger p-values (i.e. >5%) indicate that there is insufficient evidence to do so. Other values of α may be chosen depending on cost considerations, the level of conservatism required by the bank or regulator, etc.

Note that a significance analysis may be extended in various ways. First,

[4]Although R^2 can yield excessive negative values when the model predictions are worse than using the mean from the training set as prediction, these have however the same meaning as zero values, i.e. that the model does not explain any variation at all [37]. Hence, any negative values are replaced by zero to enhance their interpretation.

if required, statistical comparisons may also be performed between the performance of the test year under consideration and the performance of any previous year(s), instead of the performance on the training set. Second, the statistical tests may also be performed on specific segments of the data. This segmentation could be done either on the input data (e.g. different levels of seniority) or on the output data (i.e. different levels or 'grades' from low to high LGD risk). Third, a traffic lights approach may be used to support the visualization of the resulting p-values. Different colors can be assigned depending on the range that the corresponding p-values are in [21]. The choice and number of colors as well as the definition of their underlying p-value bounds are at the discretion of the financial institution, although a minimum number of three is suggested [40]. These extensions are however not put into practice in this paper for reasons of brevity.

4.4 Power analysis

In order to evaluate whether the results of the statistical hypothesis tests are sufficiently reliable, the statistical power π is empirically determined. The power of a test is defined as the probability that the test rejects the null hypothesis when it is indeed false. Note that this is the probability of not making a type II error (i.e. the failure to reject the null hypothesis while it is actually false). The probability of making a type II error is generally denoted as β. To decide upon acceptable statistical power, a threshold of 85% is often used. A test is then considered to be sufficiently powerful when π is higher than 85% or β is lower than 15%. Note that β (and thus also π) is related to the significance level α desired. When α is higher, β is lower or π is higher, and vice versa.

We analyze the statistical power of a test using again Beran's algorithm [13, 33, 41, 44]. First, the distribution of the test statistic under H_a is empirically derived. To do so, a same-sized training/test bootstrap sample is extracted from the original training/test set, respectively, through random sampling with replacement. Subsequently, the test statistic is calculated for each bootstrap sample. This procedure is repeated 1000 times in order to empirically build up a reliable distribution of the test statistic under H_a. Similarly to section 3.3, the distribution of the test statistic under H_0 can be empirically derived by repeatedly extracting a training and test bootstrap

sample but now from a pooled dataset which combines the training and test set values. Next, the probability of making a type II error β is calculated. This is given by the percentile rank of the test statistic's distribution under H_a for the 95th percentile (corresponds to $\alpha = 5\%$) of the distribution of the test statistic under H_0 for the case of a right-tailed test. Finally, the power can be calculated as $\pi = 1 - \beta$.

5 Results and discussion

Metric	**1984-2000**	**2001**	**2002**	**2003**	**2004**
$\overline{e}$	0.00	-0.17	-0.12	0.08	0.16
w_r	0.43	0.15	0.20	0.53	0.83
s_e^2	0.05	0.05	0.05	0.07	0.05
ab_w	0.50	0.24	0.21	0.06	0.05
RMSE	0.23	0.29	0.25	0.28	0.27
MAE	0.18	0.26	0.22	0.25	0.23
AUROC	0.70	0.56	0.55	0.63	0.55
AOREC	0.05	0.08	0.06	0.08	0.07
R^2	0.12	0.00	0.00	0.01	0.00
r	0.34	0.14	0.19	0.30	0.17
ρ	0.33	0.03	0.22	0.24	0.07
τ	0.23	0.03	0.18	0.19	0.06

Table 3: Performance metric values

This section reports and discusses the performance values of the LGD model, the statistical significance values of the performance differences between training and test sets and the statistical power values of the applied statistical hypothesis tests. The performance results of the LGD model for each metric are listed in Table 3. Both training (i.e. data from 1984 to 2000) and test set performances (i.e. data from 2001 to 2004) are given in order to show the evolution of the performances of the subsequent years with respect to the training performance. In order to detect significant performance deteriorations based on these performance values, Table 4 presents the resulting p-values of the statistical hypothesis tests corresponding to each performance metric; bold-face notation is used to denote significant differences ($p < 0.05$).

Test	**2001**	**2002**	**2003**	**2004**
T	**0.00**	**0.00**	0.97	1.00
Wilcoxon	**0.00**	**0.00**	0.98	1.00
F	0.43	0.73	**0.04**	0.52
Ansari-Bradley	0.86	0.29	**0.00**	0.10
RMSE	**0.00**	0.09	**0.03**	0.07
MAE	**0.00**	**0.00**	**0.00**	0.06
AUROC	**0.00**	**0.00**	0.21	0.10
AOREC	**0.00**	0.07	**0.02**	0.07
R^2	**0.00**	**0.00**	0.13	0.18
r	**0.01**	0.05	0.33	0.15
ρ	**0.00**	0.13	0.25	0.10
τ	**0.00**	0.26	0.35	0.08

Table 4: Statistical significance values (p-values rounded to two decimal places)

Finally, Table 5 lists the power values of each statistical hypothesis test so that we can evaluate to what degree they can be sufficiently relied upon to discover performance deteriorations. Power values greater than our example threshold of 0.85 are again put in bold.

The evolution of the central tendency of the error in terms of the mean error $\overline{e}$ or (our variant of) the Wilcoxon metric w_r is represented in the first and second row of Table 3. Regardless of whether the central tendency of the error is measured using $\overline{e}$ or w_r, the same trend is observed. The central tendency is below zero in terms of $\overline{e}$ and below 0.5 in terms of w_r for 2001 and 2002, while it is above zero in terms of $\overline{e}$ and above 0.5 in terms of w_r for 2003 and 2004. The corresponding p-values in Table 4 for both the T test and one-sample Wilcoxon test are (close to) zero for 2001 and 2002 and are (close to) one for 2003 and 2004. This means that both tests agree that the model is significantly underestimating LGD for 2001 and 2002 while this is not the case for 2003 and 2004. The consistent underestimations of the model may point to a more severe economic downturn period than expected by the model. The corresponding power values in Table 5 for both the T test and one-sample Wilcoxon test are at their maximum value for 2001 and 2002 and at their minimum value for 2003 and 2004. Both results should be seen

Test	**2001**	**2002**	**2003**	**2004**
T	**1.00**	**1.00**	0.00	0.00
Wilcoxon	**1.00**	**1.00**	0.00	0.00
F	0.09	0.02	0.54	0.01
Ansari-Bradley	0.05	0.15	**0.94**	0.30
RMSE	**1.00**	0.36	0.79	0.40
MAE	**1.00**	**0.95**	**0.92**	0.57
AUROC	**0.87**	**0.95**	0.15	0.19
AOREC	**1.00**	0.39	0.75	0.39
R^2	**0.98**	**0.91**	0.13	0.09
r	**0.89**	0.50	0.06	0.38
ρ	**0.96**	0.27	0.15	0.50
τ	**0.95**	0.18	0.12	0.55

Table 5: Statistical power values

in conjunction with the significance results obtained for those same years. Note that, even prior to the test results for 2001-2004, the Wilcoxon metric w_r when calculated using the training data only was lower than 0.5, thus indicating that the error is non-normally distributed and mean and median error are different.

The evolution of the dispersion of the error in terms of the observed variance of the error s_e^2 or the Ansari-Bradley metric ab_w is shown in the third and fourth row of Table 3. According to s_e^2, the dispersion of the error remains fairly stable for the subsequent years, except for 2003 which shows a 0.02 increase in error variance. According to ab_w on the other hand, the dispersion of the error seems to worsen over time, gradually dropping further below the reference value of 0.5 tabulated in the training results column. The corresponding p-values in Table 4 for both the F test and Ansari-Bradley test are above the significance level of 5%, except for 2003. This means both tests agree that only for 2003 there is a significant deterioration. However, the corresponding power values in Table 5, with the exception of 2003, are low for both. These low values undermine the usefulness of the p-values greater than 0.05, as they imply that we can not conclude with much certainty that there is no deterioration of the error dispersion in those years. For 2003 however, the detection of significant differences is supported by the greater

power of both tests for that year. Interestingly, the power values of the F test are generally much lower compared to the power values of the Ansari-Bradley test. This is hardly a surprise as previous power analysis studies have found that the F test is most powerful under normal assumptions [43] but extremely sensitive to non-normality [15, 36] whereas the Ansari-Bradley test makes no distribution assumptions and can be applied to relatively small samples [39]. Hence, the non-symmetrical shape of the error distribution observed in training and test samples is likely to be a major factor in explaining the lower power values of the F test.

The observed values for our series of error-, classification- and correlation-based metrics are shown in the last eight rows of Table 3. The corresponding p-values for 2001 in Table 4 are also all below the significance level of 5%. This means that all tests unanimously agree that there is a significant deterioration of the performance for 2001. For 2002 and 2003 however, some metrics still agree on significant performance deterioration but this is no longer true for all of them. In 2004, no significant performance deteriorations could be detected although all metrics show consistently lower test performance compared to the training performance. The corresponding power values shown in Table 5 are generally high for 2001 and decrease for the subsequent years. In other words, the significant differences detected for the bootstrap tests are backed up by large power values. However, in the rest of the cases (i.e. where no significant differences are detected) the bootstrap tests show only moderate power. This leaves decisions about lack of performance deterioration in those later years rather unconclusive.

Summarizing the reported performance results, one can conclude that the model shows significantly worse performance in 2001 and (according to several of the tests) 2002, specifically where it comes to being well-calibrated (see the central tendency tests) as well as in terms of a series of other performance metrics that we tested using a bootstrapping procedure. However, with the exception of significantly lower precision in 2003, performance over the subsequent two years, 2003-2004, is much more acceptable; admittedly some of the tests applied over that period appear to only have moderate power though. The observed performance loss in 2001 and 2002 may be linked to the US recession experienced in the early 2000s and the corresponding increase of the number of defaults in the loan portfolio. As suggested by several (e.g. [8]), higher default rates can be associated with higher LGDs, and our simple lin-

ear model may not adequately quantify this adverse relationship, despite the inclusion of US default rate as a macro-economic factor and the fact that the training data did include loss observations from a previous recession in the 1990s. This suspicion is strengthened by the shift in the actual LGD distribution seen in Figure 1, and the fact that the model, according to both the T test and Wilcoxon test, is consistently underestimating the elevated LGDs during those recession years. The subsequent economic recovery period may explain the slow performance correction for the following years.

When taking into account the resulting p-values and power values, one can conclude that the model performance significantly deteriorates in 2001 and according to some tests also in 2002. For the subsequent years however, the statistical hypothesis tests are not sufficiently powerful to detect performance deterioration even if the model would suffer from it. Hence, in these time periods, many of the tests are of limited value. Part of the explanation for this may lie in the smaller sample sizes available for backtesting in those last two years (see Table 1).

Generally, when the model is well trained but degrades over time, it means that the original training data is no longer representative for the current population. This can be caused by external changes (e.g. new developments in the economic, political or legal environment) or internal changes (e.g. new business strategies, exploration of new market segments or new organizational structure) [21]. A data stability analysis may offer more insight into which variables are causing possible shifts [21]. In this case it is advised to build a new model with more representative (recent) training data.

6 Conclusions

This paper addresses the call for more research on backtesting LGD models, a Basel validation requirement for any bank implementing the advanced IRB approach, by proposing a framework to backtest LGD models using a series of statistical hypothesis tests. The key idea is to evaluate two performance aspects, i.e. model calibration and precision, on an out-of-time test data set, against the original performance on the training data (or some other earlier collected reference sample). For both aspects, potentially suitable

parametric and non-parametric tests are identified. In addition, a bootstrap method is suggested to test for differences in other commonly used performance metrics. One of the main attractions of this framework is that an appropriate reference value is introduced which takes into account the number of observations available for backtesting. The practical implementation of the framework would require the following three steps. First, model performance must be quantified using a selection of metrics so that validators can monitor their evolution over the available time horizon. Second, the corresponding statistical tests should be run to help flag any significant performance degradations. Third, the power of each test could be calculated in order to verify whether weakening performance would likely be picked up by the test. The proposed backtesting framework is illustrated by applying it to an LGD model fitted to real-life corporate loss rate data.

References

[1] V.V. Acharya, S.T. Bharath, and A. Srinivasa. Understanding the recovery rates on defaulted securities. Technical report, CEPR Discussion Papers (nr. 4098), 2003.

[2] E. Altman and V.M. Kishore. Almost everything you wanted to know about recoveries on defaulted bonds. *Financial Analysts Journal*, 52 (6):57–64, 1996.

[3] E. Altman, A. Resti, and A. Sironi. Analyzing and explaining default recovery rates. Technical report, The International Swaps and Derivatives Association, 2001.

[4] E. Altman, A. Resti, and A. Sironi. *Recovery Risks: The Next Challenge in Credit Risk Management*. London: Risk Books, 2005.

[5] A. R. Ansari and R. A. Bradley. Rank-sum tests for dispersions. *The Annals of Mathematical Statistics*, 31(4):1174–1189, 1960.

[6] M. Araten, M. Jacobs, and P. Varshney. Measuring LGD on commercial loans: An 18-year internal study. *RMA Journal*, 86(8):28–35, 2004.

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

[8] B. Bade, D. Rosch, and H. Scheule. Empirical performance of loss given default prediction models. *Journal of Risk Model Validation*, 5(2):25–44, 2011.

[9] J. A. Bastos. Forecasting bank loans loss-given-default. *Journal of Banking and Finance*, 34(10):2510–2517, 2010.

[10] J.A. Bastos. Ensemble predictions of recovery rates. *Forthcoming: Journal of Financial Services Research, DOI:10.1007/s10693-013-0165-3*.

[11] J.A. Bastos. Predicting bank loan recovery rates with neural networks. Technical report, CEMAPRE Working Papers (nr. 1003), Technical University of Lisbon, 2010.

[12] T. Bellotti and J. Crook. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1):171–182, 2012.

[13] R. Beran. Simulated power functions. *The Annals of Statistics*, 14(1):151–173, 1986.

[14] J. Bi and K. P. Bennet. Regression error characteristic curves. In *Twentieth International Conference on Machine Learning, Washington DC*, 2003.

[15] G. Box. Non-normality and tests on variances. *Biometrica*, 40(3/4):317–335, 1953.

[16] M. Bruche and C. Gonzlez-Aguado. Recovery rates, default probabilities and the credit cycle. *Journal of Banking and Finance*, 34(4):754–764, 2010.

[17] R. Calabrese. Estimating bank loans loss given default by generalized additive models. Technical report, Geary Institute Working Papers (nr. 201224), University College Dublin, 2012.

[18] L. Carty and D. Lieberman. Defaulted bank loan recoveries. Technical report, Moody's Investors Service, 1996.

[19] L.V. Carty, D.T. Hamilton, S.C. Keanan, A. Moss, T. Mulvaney, T. Marshella, and M.G. Subhas. Bankrupt bank loan recoveries. Technical report, Moody's Investors Service, 1998.

[20] S. G. Caselli, S. Gatti, 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):134, 2009.

[21] G. Castermans, D. Martens, T. Van Gestel, B. Hamers, and B. Baesens. An overview and framework for PD backtesting and benchmarking. *Journal of the Operational Research Society*, 61(3):359–373, 2009.

[22] R. Chalupka and J. Kopecsni. Modeling bank loan LGD of corporate and SME segments: A case study. *Czech Journal of Economics and Finance*, 59(4):360–382, 2009.

[23] G. Christodoulakis and S. Satchell. *The analytics of risk model validation*. Amsterdam: Elsevier, 2008.

[24] J. Dermine and C. Neto de Carvalho. Bank loan losses-given-default. A case study. *Journal of Banking and Finance*, 30(4):1219–1243, 2005.

[25] R. Eales and E. Bosworth. Severity of loss in the event of default in small business and large consumer loans. *Journal of Lending and Credit Risk Management*, 80(9):58–65, 1998.

[26] B. Engelmann and R. Rauhmeier. *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing (2nd Edition)*. Berlin: Springer, 2011.

[27] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[28] J. Frye. Depressing recoveries. *Risk*, 13(11):108–111, 2000.

[29] J. Frye. LGD in high default years. Technical report, Federal Reserve Bank of Chicago, 2003.

[30] J. Grunert and M. Weber. Recovery rates of commercial lending: Empirical evidence for German companies. *Journal of Banking and Finance*, 33(3):505–513, 2009.

[31] G. Gupton. Advancing loss given default prediction models: How the quiet have quickened. *Economic Notes*, 34(2):185–230, 2005.

[32] G.M. Gupton, D. Gates, and L.V. Carty. Bank loan loss given default. Technical report, Moody's Investors Service, 2000.

[33] P. Hall and S.R. Wilson. Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47(2):757–762, 1991.

[34] D.T. Hamilton, P. Varma, S. Ou, and R. Cantor. Default and recovery rates of corporate bond issuers: A statistical review of Moody's ratings performance 1920-2002. Technical report, Moody's Investors Service, 2003.

[35] G. Loterman, I. Brown, D. Martens, C. Mues, and B. Baesens. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28(1):161–170, 2012.

[36] C. Markowski and P. Markowski. Conditions for the effectiveness of a preliminary test of variance. *The American Statistican*, 44(4):322–326, 1990.

[37] N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrica*, 78(3):691–692, 1991.

[38] Basel Committee on Banking Supervision. Studies on the validation of internal rating systems, working paper no. 14. Technical report, Bank for International Settlements, 2005.

[39] S. Shaffer. Testing for shifts in variability. *Atlantic Economic Journal*, 18(1):86, 1990.

[40] M. Svec. PD backtest empirical study on credit retail portfolio. Technical report, CSOB, 2009.

[41] P.H. Westfall. *Re-sampling based multiple testing: examples & methods for p-Value adjustment*. Hoboken: Wiley, 1993.

[42] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[43] R. S. Witte and J. S. Witte. *Statistics (10th Edition)*. Hoboken: Wiley, 2013.

[44] K. Yuan. Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, 56(1):93–110, 2003.