Validation of default probability models: A stress testing approach

Fábio Yasuhiro Tsukahara a, Herbert Kimura b, Vinicius Amorim Sobreiro b,*, Juan Carlos Arismendi Zambrano c,d

a Midway Finance, 500 Leila XLI, São Paulo, São Paulo, 02526-000, Brazil
b University of Brasilia, Department of Management, Campus Darcy Ribeiro, Brasilia, Federal District, 70910-900, Brazil
Department of Economics, Federal University of Bahia, Rua Barão de Jeremoabo, 668-1154, Salvador, Brazil
ICMA Centre, Henley Business School, University of Reading, Whiteknights, Reading RG6 6BA, United Kingdom

1. Introduction

In summary, the Basel II Accord allows banks to develop internal models for measuring risk (BCBS, 2006; Kiefer, 2009) and the Basel III Accord aims to enhance the stability of the financial system by strengthening risk coverage and highlighting the importance of on- and off-balance sheet risks, including derivatives exposure (BCBS, 2011). In addition, the Accords also require validation of risk models to determine, quantitatively and qualitatively, the models' performance and adherence to the institution's goals. In this context, Stein (2007) states that the validation process is of great importance, since it allows the benefits generated by the use of risk models to be fully obtained. However, effectively validating risk models is still a great challenge, because this is a recent aspect of banking regulation and the primary methods are still under development. In particular, credit model validation has major impediments, i.e., the small number of observations to accurately evaluate model performance (Lopez & Sadedenberg, 2000). Many validation techniques of models for bank risk management have been proposed or submitted in recent years, for market risk (Alexander & Sheedy, 2008; Boucher, Danielsson, Kountchou, & Maillet, 2014), credit risk (Lopez & Sadedenberg, 2000; Agarwal & Taffler, 2008), and model risk (Kerkhof & Meilenberg, 2004; Alexander & Leontsinis, 2011; Alexander & Sarabia, 2012; Colletatz, Hurlin, & Pérignon, 2013). Blöchlinger (2012) presents a methodology where the validation of default probability (DP) is produced over credit rating methodologies. Medema, Koning, and Lensink (2009) proposes a practical methodology for validation of statistical models of DP for portfolio of individual loans where no credit rating can be associated. However, there are no studies that attempt to identify or guide managers regarding which model is most appropriate for a given situation. With regard to the methods for estimating credit risk parameters, DP models are, according to BCBS (2005a), those that have the most developed validation

2 The growth of credit activity is an important aspect of economic development, because credit is a major source of funds for private and public organizations (Hagedoorn, 1996). However, increases in credit supply bring more exposure to credit risk and, in extreme cases, overreliance on credit can compromise the stability of the financial system (Abou-El-Sood, 2015; Arnold, Borio, Ellis, & Moshirian, 2012). Economic crises, such as the one in 2008, indicate a need for greater control and regulation of financial institutions by supervisors and for the development of risk management models. In this context, the Basel I, II, and III Accords are examples of how regulatory agencies are concerned with securing a solid international financial system; they are dynamically adjusting their requirements due to an ever-changing economic environment.
methodology. Tasche (2006) separates the performance validation process for these models into two parts, discriminative ability and calibration.

Our contribution to the literature is twofold. First we evaluate the stress test the adequacy of the primary models for risk management and thereby support the decision-making of managers regarding the model selection process. More specifically, we present the characteristics and main properties of different techniques that allow a manager to choose among classic validation models, such as the Kolmogorov–Smirnov (KS) statistic, Accuracy Ratio (AR), and Brier Score, and newer validation models, such as the Conditional Information Entropy Ratio (CIER) and Measure M. The stress test simulation is carried out in two phases: (i) an assessment of the performance of models to separate good and bad borrowers among the risk groups is performed, (ii) the accuracy of the probabilities estimated by each model is evaluated. The models were applied to credit portfolios, which were compiled using Monte Carlo simulations, to identify good and bad borrowers and how the characteristics (e.g., dependencies or moments) of these portfolios impacted the results of the models. According to Zott (2003), when there are significant limitations on gathering empirical data and variables have complex interrelationships, simulation may be useful and can actually lead to superior insights into the phenomenon. The objective of this study is not to exhaustively explore the subject but rather to enable managers to quickly identify a small number of optimal models.

Second, we analyze the default probability validation metrics using controlled sub-samples of market data. Our empirical stress analysis includes financial data of from 30,686 public US firms from 1950 and 2014, using delisting information as a proxy for default. We develop a methodology that aggregates different groups of years by high–low mean, variance, and correlation related to the financial explanatory variables. Although using empirical data does not allow as total control as using simulated data, the method gives some control over the distribution of credit scores and dependence among variables. Therefore, we can also analyze the behavior of DP evaluation metrics on empirical sub-sample data.

In the case of controlled stress simulations, for independent explanatory variables, we found that (i) the measure $M$ was the only metric able to detect changes in the mean of the explanatory variables, while there was no metric sensitive to changes in the variances; (ii) all metrics were very sensitive to the number of observations; therefore, the study can help in the validation of models for the retail and large corporations segment. In the case of controlled stress simulations, for dependent explanatory variables, we found that (i) the only metric that captured a performance decrease for both increases and decreases in the correlation parameter was measure $M$, all other measures exhibited an increase in performance as the strength of the correlation was decreased; (ii) modeling using the $T$ copula and Gaussian copula provided no difference in the sensitivity results of the metrics.

The remainder of this study is structured as follows: in Section 2, a literature review of credit is presented; Section 3 and 4 address the aspects used to compare the models and their results; Section 5 presents a empirical application; in Section 6, the primary conclusions are presented and discussed.

---

5 Davis, Eisenhardt, and Bingham (2007) presents a reference to the theory developed using simulation methods.

6 Explanatory variables are any variables that can lead to a causal explanation of the relationships in default, such as the ones included in the $Z$-score of Altman (1968).

2. Literature review

The Basel II Accord aims to improve the awareness of the financial institutions regarding their credit risk (Hakenes & Schnabel, 2011). The Basel II Accord first pillar aims to guide the calculation of minimum capital requirements, i.e., it reviews the main ideas presented in the Basel I Accord. The minimum capital requirement is calculated based on the Internal Rating Based (IRB) method, which is generally estimated internally by a bank based on the following parameters: (i) $DP$; (ii) Exposure at Default ($EAD$); (iii) Loss Given Default ($LGD$); and (iv) Maturity ($M$). It is worth noting that in the simplified version of the IRB, it is only necessary to calculate the $DP$ value because the other parameters are defined by regulatory bodies. From this point of view, the calculation of $DP$ becomes crucial.

2.1. Validation tests for default probability models

Two of the most used validation tests are the Cumulative Accuracy Profile (CAP) curve and AR developed by Sobehart, Keenan, and Stein (2000a). Their calculation is performed by ranking all parties based on the scores estimated by the model. Once ranked, for a certain cutoff score, it is possible to identify the fraction of defaults and non-defaults with scores that are less than the cutoff score. The CAP curve is obtained by calculating these fractions for all possible cutoff points, as shown in Fig. 1.

According to Engelmann, Hayden, and Tasche (2003), the AR can be defined by:

$$AR = \frac{a_{0}}{a_{p}}$$

(1)

where $a_{0}, a_{p}$ are the areas defined in Fig. 1. The closer the AR is to one, the greater the discriminative ability of the model.

The Receiver Operating Characteristic (ROC) curve and the area under the ROC curve are other widely used validation measures developed by Tasche (2006). The ROC curve is obtained by plotting $HR(C)$ versus $FAR(C)$, where $HR(C)$ is the hit rate and $FAR(C)$ the false alarm rate at score $C$. According to Engelmann et al. (2003), the higher the area under the ROC curve of the model, the better the performance. Considering the ideal situation, i.e., an ROC area equal to 1, the area may be calculated using Eq. (2):

$$AUROC = \int_{0}^{1} HR(FAR) d(FAR).$$

(2)

The Pietra Index developed by Pietra (1915) is a widely used index, whose geometric interpretation corresponds to half of the shortest distance between the ROC curve and the diagonal. This index can be calculated as:

$$PI = \frac{\sqrt{2}}{4} \max_{x} |HR(C) - FAR(C)|$$

(3)

Sobehart et al. (2000a) defined the CIER measure according to:

$$CIER = \frac{H_{0}(P) - H_{1}}{H_{0}(P)}.$$  

(4)

where $H_{0}(P), H_{1}$ are entropy functions developed by Jaynes (1957) and related to Kullback-Leibler (KL) distance, with the purpose of finding a function with conditions of continuity, monotonicity, and composition law, that represents the uncertainty of a probability distribution. Keenan and Sobehart (1999) defined the measure $H_{0}(p)$ as the entropy of a binary event for which $p$ is the default rate of the sample.

---

7 See Eliazar and Sokolov (2010) for a recent economic application.
The BRIER score was originally proposed by Brier (1950); this metric has the objective of measuring the accuracy of forecasts provided by a given model; and it was initially proposed to measure the accuracy of weather forecasts. The Brier Score can be calculated according to:

$$BRIER = \frac{1}{N} \sum_{i=1}^{N} (P_i - O_i)^2,$$

where \(P_i\) corresponds to the probability of occurrence of the event given by the model for the \(i\)-th component of the sample; and \(O_i\) corresponds to a binary variable \((1/0)\), where one means that the event was observed and zero means that the event was not observed. A perfect DP model would estimate a probability equal to one for observed default events and zero probability for default events that are not observed. Consequently, the Brier Score would be equal to zero, i.e., a Brier Score closer to zero indicates higher accuracy of the model.

Measure \(M\) is proposed by Ostrowski and Reichling (2011), and it aims to evaluate the discriminative ability of the default models. Let \((a_{D,i}, a_{ND,i})\) be areas of default and non-default, and \((r_{D,i} \text{ and } r_{ND,i})\) hit rates of default and non-default for the \(i\)-th rating. (Ostrowski & Reichling, 2011) defined a measure of the performance of the model:

$$m = \sum_{i=1}^{k} \left[ HR_i \left( r_{D,i} - a_{D,i} \right) + FAR_i \left( r_{ND,i} - a_{ND,i} \right) \right],$$

where \(k\) corresponds to the total number of ratings. Note that this measure is not yet standardized, which precludes direct comparison of two different models. To standardise this measure, the values of \(m_{\text{max}}\) and \(m_{\text{min}}\) are calculated and the standardized \(M\) measure is:

$$m_{\text{min}} = \sum_{i=1}^{k} \min \{ HR_i \left( r_{D,i} - a_{D,i} \right), FAR_i \left( r_{ND,i} - a_{ND,i} \right) \}.$$  

The value of \(m\) should be in the range of 0 and 1, where 1 indicates perfect predictive ability of the model.

According to the studies of Kolmogorov-Smirnov, Lilliefors (1967) presents a procedure to test if a set of \(n\) observations is derived from a normal distribution. In a simplified manner, Lilliefors (1967) proposes a hypothesis test for measure \(D\) which is the absolute difference between the accumulated distribution function of the sample; and the normal accumulated distribution function with mean and variance equal to those of the sample. To validate credit risk models, it is worth noting that the aim is not to analyze the normality of a distribution but rather to check whether the model can distinguish defaults and non-defaults. For such a purpose, the KS statistic can be used, as described in Joseph (2005), to quantify the greatest distance between the accumulated distribution of defaults and non-defaults. \(KS\) can be calculated using:

$$KS = \max(|F_D(S) - F_{ND}(S)|).$$

where \(F_D\) corresponds to the accumulated distribution function of default cases,

\(F_{ND}\) corresponds to the accumulated distribution function of non-default cases, and \(S\) corresponds to the score.

The parameter of information value (\(IV\)) proposed by Tasche (2006) measures how default and non-default events are distributed differently among ratings. Let \(R_i\) be the \(i\)-th rating, \(p_D(R_i)\) the ratio of defaults of the \(i\)-th rating, and \(p_{ND}(R_i)\) the ratio of non-defaults of the \(i\)-th rating. Then, the value of \(IV\) can be calculated following (Joseph, 2005) using:

$$IV = \sum_i \left[ p_D(R_i) - p_{ND}(R_i) \right] \times \ln \left[ \frac{p_D(R_i)}{p_{ND}(R_i)} \right].$$

It is important to highlight that high \(IV\) values indicate high discriminative ability (Tasche, 2006).

2.2. Studies of the validation of DP models

Although the process of validation of credit risk models required by Basel II is still relatively new to the global financial market, some studies about model performance measurement techniques had been previously published. Among these studies, the following are noteworthy.

Keenan and Sobehart (1999) presented the following techniques to measure the performance of predictive default models: CAP, AR, CIER, and Mutual Information Entropy (MIE). Using a dataset that included data from 9,000 public companies, covering the years of 1989 through 1999, and containing 530 default events, the authors applied a return based model and four additional prediction models (Altman, 1968; Shumway, 2001; Merton, 1974; Sobehart et al., 2000a), the authors were able to conclude that the tests were effective and measured...
distinct aspects of the model. Keenan and Sobehart (1999) emphasized that the CAP curve and the AR measure the discriminative ability of the default model prediction and the CIER and MIE assess whether different models interact by adding information or are simply redundant; Hanley and McNeil (1982) and Engemann et al. (2003) presented the Receiver Operator Characteristic (ROC) technique and explained its use in the context of validation of rating models. There exists a relationship between the AR and the area under the ROC curve (AUROC) that can be calculated by:

$$AR = 2 \times AUROC - 1.$$  

Karoukas (2004) presented a validation methodology for credit scoring and DP models for the retail segment and small companies. The author also argued that the KS statistic has the limitation of not referring to where the point of maximum distance occurs and that the AUROC is more generic regarding this point and therefore better; Joseph (2005) presented a validation methodology based on several tests, and the final evaluation of the model was based on the average performance of the models in these tests. In addition to the AR, ROC, KS, and Kullback Leibler measures, Joseph (2005) used other measures, such as the mean difference and IV.

Ostrowski and Reichling (2011) found that the AR and AUROC measures can, in certain circumstances, lead to erroneous conclusions and cause low-performance models to be rated well according to these indicators. This observation is in accordance with Engemann et al. (2003), as the author states that if the distribution of default events is bimodal, a perfect model can have an AUROC equal to that of a random model. Furthermore, Ostrowski and Reichling (2011) proposed another measure called M, to measure the performance of the model and applied this new measure in the credit rating models used by the agencies Standard & Poor’s and Moody’s. Considering the period from 1982 to 2001, the authors observed that the AUROC measure behaved in a stable manner compared with measure M, which exhibited high variability in the measurement of model performance.

3. Numerical stress test simulation

Taking into account that different simulated situations lead to distinct behaviors of the metrics it is possible to analyze how the characteristics of the default phenomenon could influence a broad set of evaluation metrics. Consequently, we studied the traditional performance measures like KS, AUROC and AR (BCBS, 2005b; Hand, 2009; Keenan & Sobehart, 1999; Marshall, Tang, & Milne, 2010; Ostrowski & Reichling, 2011; Verbraken, Bravo, Weber, & Baesens, 2014) as well as other less common metrics like Pietra, BRIER, CIER, Kullback-Leibler (KL), Information Value (IV) and measure M (Joseph, 2005; BCBS, 2005b; Ostrowski & Reichling, 2011; Iazzi, Oricchio, & Vitale, 2012).

Our analysis of the validation techniques can be divided into two parts:

1. In the first part, good and bad borrower distributions were simulated according to an arbitrary scoring rule and assigned to variable $Y$. The properties of the distributions were then changed. With this setting, it was possible to analyze the impact of these changes on the values calculated using the validation techniques. This part of the methodology can be summarized by:
   (a). Generation of the variable $Y$ of good and bad borrowers using a Monte Carlo simulation approach with a normal distribution to tag subjects into good and bad borrowers;
   (b). Generation of explanatory variables $X_1, X_2$ by a normal distribution with different mean and volatility, depending on whether the subject was tagged as a good or bad borrower in step 1;
   (c). Generation of default variable $Y$ by a gamma distribution;
   (d). Association of $X_1 = 1.2$ with $Y$ using a bi-stochastic matrix; and,
   (e). Calculation of the performance of the entropy-based validation measures by their sensitivity to changes in $X_1$ and $X_2$.

2. In the second part of the study logistic models were developed from simulated portfolios that contained a default event and other independent variables. As a consequence, it was possible to analyze how changes in the variables, or in the existing relationships between them, affected the values measured by the techniques studied. This part of the methodology includes
   (a). Use of the logistic model to calculate the probability $P$ of the default of a subject depending on $X_1$ and $X_2$;
   (b). Generation of credit scores using $P$, having n-ratings (10 in our numerical simulation) for the classification of the subject;
   (c). Calculation of the number of realized defaults from the variable $Y$ obtained in the first part of the methodology; and,
   (d). Calculation of the performance of rating-based validation measures by their sensitivity to changes in $X_1$ and $X_2$.

The methodology presented associated changes in the distribution properties of $X_1, X_2$, depending on whether is it is a good or bad borrower, with the performance of the techniques for the validation of DP. We then analyzed how different changes in distribution parameters and relationships between variables affect the performance of the validation techniques.

3.1. Normal distribution simulation for good and bad borrowers

Applying the Monte Carlo simulation technique, different portfolios that contained normal distributions of good and bad borrowers were generated. All portfolios consisted of 30,000 simulations and a bad borrower rate of approximately 10%. The parameters changed in the different portfolios were the mean scores of good and bad borrowers and the deviations of both distributions. Although the deviations were changed, this change was performed on both distributions such that in all portfolios, the deviation of the bad borrower distribution was equal to the deviation of the good borrower distribution. The procedure used to construct the portfolio was as follows:

1. Random classification of the portfolio subjects into good and bad borrowers;
2. Determination of the mean score of the distribution of good borrowers, mean score of the distribution of bad borrowers, and standard deviations of both distributions; and,
3. Assignment of a score to each subject using the Monte Carlo simulation technique

Table 1 presents the parameters used in the different simulated portfolios.

We generated simulated data for the relevant variables to analyze the effects of parameters of the credit score distributions of good and bad borrowers on the validation metrics. Once the simulations were completed, the KS, AUROC, AR, and Pietra Index techniques were applied. Subsequently, each portfolio had its score ranked, and the 30,000 components were distributed into 10 ratings of 3,000 components each. Rating 1 contained the lowest scores, and rating 10 contained the highest scores. Through separation in ratings, it was possible to estimate the values of the CIER, Kullback Leibler and IV validation techniques.

---

8 The use of simulated data makes it possible to have control over the various relationships among variables. Therefore, we can analyze the adequacy of the validation techniques in a controlled environment. When using empirical data from real default situations, the very complexity of the phenomenon can preclude an analysis focused solely on the variables of interest, jeopardizing the study of the validation of models. Nevertheless, we also conducted a separate analysis aiming to identify the behavior of validation models under real-world credit events. Therefore, our study also allows the comparison of results using both simulated and real-world data.
3. Assignment of a score to each subject using the Monte Carlo bimodal distributions for bad borrowers. The parameters of the normal distribution for good and bad borrowers were the constant and equal to 5.0. The parameters changed for analysis of the bimodal distribution which mean score of good borrowers was kept equal to 1.0, and the distribution mean of bad borrowers was kept equal to 2.0, and the distribution mean of good borrowers (DM1) and another with a higher mean score (DM2). For these portfolios, the distribution deviations of good borrowers, DM1 and DM2, were kept constant and equal to 1.0, and the distribution mean of good borrowers was kept constant and equal to 5.0. The parameters changed for analysis of the bimodal distribution were the distribution means DM1 and DM2 and the bimodality intensity, i.e., the number of bad borrowers in DM1 and DM2. The total number of bad borrowers, i.e., the number of bad borrowers in the two distributions, is approximately equal to 10% of the portfolio. The procedure used to construct the portfolio contained the following steps:

1. Random classification of the portfolio subjects into good and bad borrowers. When the subject was classified as a bad borrower, a second random classification was performed to determine whether it belonged to distribution DM1 or distribution DM2;
2. Definition of the mean scores of distributions DM1 and DM2; and
3. Assignment of a score to each subject using the Monte Carlo simulation technique.

Table 2 presents the parameters used in the different simulated portfolios. In this analysis we focus on assessing the adequacy of evaluation metrics when the credit score distribution of bad borrowers is bimodal. In particular, we analyze the effects of changing the characteristics of the bimodality of the distribution. The procedure of dividing the portfolio into ratings and the application of validation techniques was performed identically to that for the portfolios with normal distributions for good or bad borrowers.

<table>
<thead>
<tr>
<th>Score Mean</th>
<th>Score Mean</th>
<th>Std. dev.</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5</td>
<td>7.0</td>
<td>6.5</td>
<td>6.0</td>
</tr>
<tr>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
<td>3.5</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Note: The first and second columns depict the mean the third column shows the percentage of bad borrowers in the neighborhood of each of mode of the bimodal distribution.

3.3. Generation of simulated portfolios

For the construction of the portfolios used to obtain the logistic models, the values of the dependent and explanatory variables were simulated. In the case of default probability models, some examples of explanatory variables are the ratio of sales to total assets and the ratio of retained earnings to total assets, which are used in the Altman Z-score model (Altman 1968). Other examples are the rate of indebtedness and economic sector for the corporate credit segment and the income, age, and occupation of an individual for the credit segments that are related to individuals.

For the construction of the dependent variable (default), random values were simulated from a gamma distribution with parameters \( k = 1 \) and \( \theta = 0.1 \). Then, for each gamma distribution value simulated, a random value was generated from a uniform distribution \([0,1]\). If the value of the uniform distribution was less than the value of the gamma distribution, the dependent variable \( Y \) would have value 1 (Default); otherwise, it would have value 0 (non-default). The explanatory variables \( X_{1} \) and \( X_{2} \) were built using randomly simulated values of normal distributions. Variable \( X_{1} \) has mean equal to 7.0 and deviation equal to 2.0. Variable \( X_{2} \) has mean equal to 20.0 and deviation equal to 4.0.

3.3.1. Dependence among variables

For the generation of explanatory variables with dependence, the copula method was used. This method allows, based on Sklar’s theorem, us to formulate joint distributions with several types of dependence. Nelsen (1999) states that copulas are functions that join or couple joint distribution functions to their unidimensional marginal distribution functions. Thus, in this study, portfolios in which \( X_{1} \) and \( X_{2} \) were independent and dependent according to Gaussian copulas were used.

3.3.2. Association between the dependent variable and independent variables

Once the explanatory variables with dependencies are simulated, it is necessary to simulate the default events and use a method that allows us to associate the default events with explanatory variables. This association was made based on a method that uses bi-stochastic matrices. Bi-stochastic matrices can be seen as a discrete version of a copula. Briefly, this method, which is based on that of Hlawatsch and Ostrowski (2011), consists of separating both the dependent and explanatory simulated variables into groups and associating the groups according to a dependency structure. Hence, according to the proposition by Hlawatsch and Ostrowski (2011), the steps used to associate a dependent variable and an independent variable with a negative causal relationship, considering \( Y \) as the dependent variable and \( X \) as the explanatory variable, are as follows:

1. Sort each variable and separate them in blocks such that the first block contains the lowest values and the last block has the highest values;
2. Next, a bi-stochastic matrix \( M \) is constructed with elements \( m_{i,j} \) that represent the probability of observing an element of the \( i-th \) block of \( Y \) associated with an element of the \( j-th \) block of \( X \). The indices \( i \) and \( j \) are natural numbers and range from one to the number of groups (in this study, five groups were used), and the parameters \( m_{i,j} \), that comply with the conditions given by

\[
\sum_{i=1}^{5} m_{i,j} = 1, \quad \sum_{j=1}^{5} m_{i,j} = 1. \tag{11}
\]

Although it may seem counter-intuitive to use a discrete dependence measure to associate two continuous variables, it is possible, considering that there will be some error produced by discretization of the dependence of the continuous variables. This error is reduced by increasing the size of the bi-stochastic matrix.
As an example of the application of this technique, suppose that a matrix $M$ is given by:

$$M = \begin{pmatrix}
0.01 & 0.04 & 0.08 & 0.16 & 0.71 \\
0.04 & 0.07 & 0.18 & 0.55 & 0.16 \\
0.08 & 0.18 & 0.48 & 0.18 & 0.08 \\
0.16 & 0.55 & 0.18 & 0.07 & 0.04 \\
0.71 & 0.16 & 0.08 & 0.04 & 0.01
\end{pmatrix}.$$ 

Once the matrix is built, an observation from the first block of $Y$ and a random number from a uniform distribution $[0-1]$ are selected.

Assuming that the random number is 0.4 and analyzing the first line of the matrix, the observation of the explanatory variable $X$ associated with the observation of the dependent variable $Y$, which was previously selected, belongs to the fifth block because 0.4 is greater than 0.01 + 0.04 + 0.08 + 0.16. If the random number is 0.015, the $X$ observation would be in the second block. This process is replicated until the elements of the first block of $Y$ are exhausted. Then, the process continues to the second block, and the second matrix line is analyzed. The process follows the same form until all associations are performed, i.e., until all $Y$ values have an associated $X$ value. If the $X$ group selected is empty, that is, all elements have been previously selected, a value from the closest non-empty group is selected. Once the process is
completed, a set of observations with a dependency relationship between the explanatory variables and the dependent variable is obtained. If \(X_1\) and \(X_2\) are independent, two bi-stochastic matrices are used to perform the association, where one of the matrices associates \(Y\) to \(X_1\) and the other associates \(Y\) to \(X_2\). For cases of dependency between \(X_1\) and \(X_2\), three variables, \(X_1\), \(X_2\) and \(X_3\), are simulated, where \(X_3\) has a normal distribution with mean zero and deviation 1.0. The correlations between \(X_1\) and \(X_2\) and that between \(X_2\) and \(X_3\) are set to be 0.7. The association between \(Y\), \(X_1\) and \(X_2\) is performed in two steps. In the first, a bi-stochastic matrix is used to associate \(Y\) to \(X_2\). For each \(X_3\) value, there are related \(X_1\) and \(X_2\) values because the variables are not independent; thus, in the second step, \(Y\) is associated with the \(X_1\) and \(X_2\) values that are related to the variable \(X_3\).

### 3.3.3. DP model and separation into ratings

Once the portfolio was built with the binary variable default event observation \(Y\) and independent variables \(X_1\) and \(X_2\), a DP model was developed using the logistic regression technique. The use of logistic models for the development of score models is very common in the financial markets and in academia (Steenackers & Goovaerts, 1989; Bensić, Sarlija, & Zekic-Susac, 2005). Through the logistic regression technique, it is possible to estimate the probability of a default event \((Y = 1)\) given a set of \(n\) explanatory variables. Defining this probability as being \(P(Y = 1|X_1, X_2, \ldots, X_n) = P(X)\), the logistic model can be specified following (O’Connell, 2006) using and (12):

\[
\ln \left( \frac{P(X)}{1 - P(X)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = Z. \tag{12}
\]

Considering (12), the probability of occurrence of a default event for a set of explanatory variables, \(P(X)\), can be obtained using (13):

\[
P(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)} = \frac{\exp(Z)}{1 + \exp(Z)}. \tag{13}
\]

Because in the simulated portfolios, two independent variables \((X_1\) and \(X_2\)) were used, \(Z\) can be calculated using (14), and the model score is obtained using (15):

\[
Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{14}
\]

\[
\text{Score} = 1 - \frac{e^Z}{1 + e^Z} = 1 - P. \tag{15}
\]

In this case, the higher the probability of default of the subject, the lower the score. The number of defaults estimated for the \(k\) — th rating can be obtained as the sum of the probabilities of default of the elements contained in this rating. Hence, for a rating that contains \(m\) elements, the number of estimated defaults is obtained using (16):

\[
QDE = \sum_{i=1}^{m} DP_{ki}. \tag{16}
\]
The objective of the analysis is to identify how changes in the characteristics of independent variables affect the model performance and how each technique responds to these changes. Another important aspect is the sensitivity of the validation techniques to changes in the dependencies between dependent and explanatory variables, i.e., how each technique responds to changes in the bi-stochastic matrices.

3.4.2. Explanatory variables with dependencies simulated via Gaussian copulas

For the case in which the dependencies among explanatory variables were simulated using Gaussian copulas, the following process was performed:

1. Model with zero correlation between $X_1$ and $X_2$:
   - Generation of 20 simulated portfolios with zero correlation between $X_1$ and $X_2$ using the bi-stochastic matrix $M_1$ for association between explanatory variables and $Y$;
   - Development of logistic models from simulated portfolios;
   - Application of the validation techniques to the 20 models developed and determination of a confidence interval for each technique;
   - Choice of one of the 20 models to be used in the samples with changed parameters;
   - Application of the chosen model and validation techniques to the 20 models developed and determination of a confidence interval for each technique.

2. Model with 0.5 correlation between $X_1$ and $X_2$:
   - Generation of 20 simulated portfolios with a correlation of 0.5 between $X_1$ and $X_2$ using the bi-stochastic matrix $M_2$ for association between explanatory variables and $Y$;
   - Development of logistic models from simulated portfolios;
   - Application of the validation techniques to the 20 models developed and determination of a confidence interval for each technique;
   - Choice of one of the 20 models to be used in the samples with changed parameters;
   - Application of the chosen model and validation techniques to the 20 models developed and determination of a confidence interval for each technique.

The objective of the analysis is to identify how changes in the characteristics of independent variables affect the model performance and how each technique responds to these changes. Another important aspect is the sensitivity of the validation techniques to changes in the dependencies between dependent and explanatory variables, i.e., how each technique responds to changes in the bi-stochastic matrices.
the portfolios were the means of the two distributions of bad borrowers. Bases were generated, and the parameters that were varied among normal for good borrowers and bimodal for bad borrowers, 25 simulations of deviations of the distributions. For cases in which the distribution was normal for good borrowers and bimodal for bad borrowers, the distributions were generated. The parameters that were varied among different distributions were the distributions of good and bad borrowers.

4. Results of simulations

To obtain the results of application of the techniques on the logistic models, modelling routines and procedures for generation of variables with or without dependencies were developed in the R language and used in addition to the validation techniques that were implemented in Visual Basic Application (VBA) software. The results for the score distribution simulation were obtained by implementing the simulations and validation techniques using VBA.

4.1. Simulation of score distributions

A total of 15 bases that contained simulations of cases in which good and bad borrowers had normal distributions were generated. The parameters that varied among the portfolios were the means and deviations of the distributions. For cases in which the distribution was normal for good borrowers and bimodal for bad borrowers, 25 simulated bases were generated, and the parameters that were varied among the portfolios were the means of the two distributions of bad borrowers and the bimodal intensity.

4.1.1. Normal distributions of good and bad borrowers

The parameters used in the different score simulations and the values obtained for each validation technique used are presented in Table 3.

For the portfolios with normal distributions of good and bad borrowers, it is expected that the indicators would exhibit a decreased discriminative ability when the means of the distributions approach one another or as the distribution deviations increase. From an analysis of Table 3, it is possible to observe that all of the indicators support this claim, that is, as the distribution means get closer or as the distribution deviation increases, a decrease in all indicators can be observed. Therefore, the results suggest that all the metrics present a loss of performance due to the approximation of the probability distributions of good and bad borrowers.

4.1.2. Normal distribution of good borrowers and bimodal distribution of bad borrowers

As described earlier, for the simulations with normal distributions of good borrowers and bimodal distributions for bad borrowers, the deviations of all distributions (Good, DM1, and DM2) were set to 1.0, and the distribution mean of good borrowers was set to 5.0. The parameters that were varied among different distributions were the distribution means of bad borrowers and the ratio of bad borrowers in each of them. Although the mean of the bad borrower distributions varied, they were equidistant to the distribution mean of good borrowers, as shown in Table 4.

When the performance indicators were analyzed assuming a fixed ratio of bad borrowers for M1 and M2, that is, by varying only the bad borrower distribution means, we observed a decrease in performance for all indicators because the distributions of bad borrowers were closer to the distribution of good borrowers. This occurrence was expected because the approximation between the means of the distribution will result in a less discriminative ability for the model. However, when ob-
In the performance of the discrimination model whereas metrics based about the quality of the scores. Traditional metrics fail to assess changes of bad borrowers, some metrics can better convey information both as good models. The results show that due to the bimodal feature with a ratio of 90% as having good performance and the model with a much less significant drop in model performance, which was unexpected because the distributions M1 and M2 had the same variance and means that were equidistant from the mean value of the good borrower distribution. This decrease in performance can also be observed for the CIER, KL and IV indicators, although the decreases were much less significant for these indicators. By analyzing the results for simulations with mean 7.5 and ratios 90% and 50% for DM1, it is possible to observe that the KS, AUROC, AR, and Pietra metrics classify the model with a ratio of 90% as having good performance and the model with a ratio of 50% as having bad performance; however, CIER, KL and IV classify both as good models. The results show that due to the bimodal feature of bad borrowers, some metrics can better convey information about the quality of the scores. Traditional metrics fail to assess changes in the performance of the discrimination model whereas metrics based on entropy are more sensitive.9

### 4.2. Portfolios with independent explanatory variables

For the portfolios with independent explanatory variables, 20 portfolios were simulated and the validation techniques were applied. Thus, it was possible to define the confidence intervals for each technique. The mean value of the metrics and the confidence interval defined are presented in Table 5.

It is important to stress that the model chosen among the 20 models used for the construction of the confidence interval, which was used in other simulations with variation of the mean and variance parameters, was the regression model developed from sample Indep10 (KS = AUROC = 0.7495, AR = 0.4989, PietraIndex = 0.1352, Brier = 0.0812, CIER = 0.1141, KL = 0.0361, IV = 0.9392, M = 0.8355).

#### 4.2.1 Impact caused by variation of the X1 mean

Simulations in which the mean of the independent variable X1 was varied were performed, and the chosen model was applied. The results of the metrics obtained using the validation techniques can be found in Table 6.

It is possible to observe from the results presented in Table 6 that with the exception of metric M, none of the other metrics exhibited a drop in model performance with displacement of the mean of variable X1 to lower or higher values. Metric M exhibited a drop in model performance in both cases, that is, the value of M decreased as the X1 mean was increased or decreased. Table 7 presents the estimated and observed values for each rating.

where:

$$\text{Est}_M$$ is the number of bad borrowers estimated by the model;  
$$\text{Est}_B$$ is the number of good borrowers estimated by the model;  
$$\text{Obs}_M$$ is the number of bad borrowers observed in the rating; and,  
$$\text{Obs}_B$$ is the number of good borrowers observed in the rating.

According to Table 7, there was considerable variation in the value of bad borrowers estimated for each rating. This variation was caused by the change in the mean X1 value, i.e., the accuracy of the model was greatly affected by the change in the mean of the independent variables, and this aspect can only be observed in the values of measure M. The entropy measures did not undergo considerable changes because the default rate observed for each rating did not vary greatly. The more traditional metrics such as KS, AUROC, Pietra, and AR, remained stable, possibly because the rankings of good and bad borrowers did not undergo major changes. The loss of performance of measure M in Table 6 is a consequence of the lower accuracy of the model. Table 7 corroborates this argument, since the differences between the estimated and the observed values increase with the shift of the mean of the explanatory variable.

### Table 5

<table>
<thead>
<tr>
<th>Rating</th>
<th>Original (Correl = 0.0)</th>
<th>Correl = 0.1</th>
<th>Correl = 0.55</th>
<th>Correl = 0.90</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EstM</td>
<td>EstB</td>
<td>ObsM</td>
<td>ObsB</td>
</tr>
<tr>
<td>1</td>
<td>276</td>
<td>724</td>
<td>260</td>
<td>740</td>
</tr>
<tr>
<td>2</td>
<td>165</td>
<td>835</td>
<td>191</td>
<td>809</td>
</tr>
<tr>
<td>3</td>
<td>124</td>
<td>967</td>
<td>130</td>
<td>870</td>
</tr>
<tr>
<td>4</td>
<td>98</td>
<td>902</td>
<td>120</td>
<td>880</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>920</td>
<td>68</td>
<td>932</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>935</td>
<td>62</td>
<td>938</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>948</td>
<td>38</td>
<td>962</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>959</td>
<td>39</td>
<td>961</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>970</td>
<td>22</td>
<td>978</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>983</td>
<td>17</td>
<td>983</td>
</tr>
</tbody>
</table>

### Table 6

<table>
<thead>
<tr>
<th>Correlation</th>
<th>KS</th>
<th>AUROC</th>
<th>AR</th>
<th>Pietra Index</th>
<th>Brier</th>
<th>CIER</th>
<th>KL</th>
<th>IV</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.343*</td>
<td>0.722*</td>
<td>0.445*</td>
<td>0.121*</td>
<td>0.084</td>
<td>0.088*</td>
<td>0.023*</td>
<td>0.656*</td>
<td>0.762*</td>
</tr>
<tr>
<td>0.20</td>
<td>0.386*</td>
<td>0.737*</td>
<td>0.473*</td>
<td>0.13*</td>
<td>0.087</td>
<td>0.102*</td>
<td>0.034*</td>
<td>0.303*</td>
<td>0.731*</td>
</tr>
<tr>
<td>0.30</td>
<td>0.48*</td>
<td>0.721*</td>
<td>0.441*</td>
<td>0.123*</td>
<td>0.086</td>
<td>0.086*</td>
<td>0.023*</td>
<td>0.646*</td>
<td>0.798</td>
</tr>
<tr>
<td>0.40</td>
<td>0.322</td>
<td>0.713</td>
<td>0.426</td>
<td>0.117</td>
<td>0.082</td>
<td>0.079</td>
<td>0.025</td>
<td>0.595</td>
<td>0.874</td>
</tr>
<tr>
<td>0.55</td>
<td>0.300</td>
<td>0.699</td>
<td>0.398</td>
<td>0.106</td>
<td>0.085</td>
<td>0.072</td>
<td>0.023</td>
<td>0.538</td>
<td>0.869</td>
</tr>
<tr>
<td>0.60</td>
<td>0.298</td>
<td>0.693</td>
<td>0.387</td>
<td>0.106</td>
<td>0.084</td>
<td>0.064</td>
<td>0.02</td>
<td>0.479</td>
<td>0.9</td>
</tr>
<tr>
<td>0.70</td>
<td>0.269*</td>
<td>0.668*</td>
<td>0.336*</td>
<td>0.095*</td>
<td>0.091*</td>
<td>0.049*</td>
<td>0.016*</td>
<td>0.371*</td>
<td>0.83</td>
</tr>
<tr>
<td>0.80</td>
<td>0.264*</td>
<td>0.677*</td>
<td>0.353*</td>
<td>0.093*</td>
<td>0.086</td>
<td>0.054*</td>
<td>0.017*</td>
<td>0.413*</td>
<td>0.874</td>
</tr>
<tr>
<td>0.90</td>
<td>0.247*</td>
<td>0.66*</td>
<td>0.321*</td>
<td>0.087*</td>
<td>0.085</td>
<td>0.044*</td>
<td>0.014*</td>
<td>0.345*</td>
<td>0.832</td>
</tr>
</tbody>
</table>

9 These results demonstrate the importance of validation of DP metrics, as a technique that is unable to properly discriminate different borrowers can produce Type I and Type II errors in terms of good/bad credit. Sobehart et al. (2000a) and Sobehart, Keenan, and Stein (2000b) present an analysis where Type I and Type II errors are discussed: Type I error is associated with a high-credit rating for a bad borrower, which can result in an excess in the number of defaults. Type II error means that a low-credit rating is assigned to a good borrower which is traduced in a lower return given that less number of credits to good borrowers will be assigned as a result of a better bidding rate by the competition. The AUROC measure resulted the best measure in terms of lower Type I and II errors.
4.2.2. Impact caused by changes in the variance of $X_2$

From an application of the chosen model to samples with variation of the $X_2$ variance (originally, the variance had the value 4.0), the results obtained for the validation metrics are presented in Table 8.

It is possible to observe from the results presented in Table 8 that the measures are, in general, not very sensitive to changes in the variance of the independent variable, and only a major change in the variance (deviation equal to 10.00) caused the indicators to be outside the estimated confidence interval. Regarding the correctness of the default values of the ratings, it is possible to observe a result consistent with other metrics, i.e., a lower correctness for the deviation of 10.00, as indicated by Table 9.

4.2.3. Impact caused by changes to the bi-stochastic matrices

The matrices initially used to associate variables $X_1$ and $X_2$ to variable $Y$ were the matrices $M_1$ and $M_1'$. These matrices had high values in their diagonals (the main diagonal in case of $X_1$ and the secondary diagonal in case of $X_2$) that cause a stronger dependency relationship between independent variables and the dependent variable. Table 10 presents the results of the validation techniques obtained using the chosen model for portfolios simulated using the other bi-stochastic matrices.

Based on the results presented in Table 10, it is possible to observe that all metrics except the Brier score exhibited decreased model performance as the dependency between the variables was changed. Measure $M$ was relatively insensitive to the change in dependency because its value changed less sensitively than those of other measures, such as the entropy measures ($\text{CIER}$, $\text{IV}$, and $\text{KL}$), $\text{KS}$, $\text{AUROC}$, $\text{Pietra}$, and $\text{AR}$. In practical terms, the metrics indicate that if the model was developed using a variable that has a strong relationship with the default event and this relationship is weakened over time, the discriminative ability of the model will decrease. Table 11 presents the estimated and observed values for each model rating.

It is possible to observe from the information contained in Table 11 that there was no considerable change in model correctness between the original portfolio and the portfolio simulated from $M_3$ and $M_3'$. There was a more significant change in model correctness for the portfolio simulated from the matrices $M_5$ and $M_5'$. Another aspect that can be observed is that in the original model, the heterogeneity of the rate of defaults observed among the ratings is greater than in the portfolio that used the matrices $M_5$ and $M_5'$. This greater homogeneity in default rates among ratings can be observed from the sharp drop in the value of the entropy measures among portfolios.

4.3. Portfolios with dependent variables simulated using gaussian copulas

Similar to the tests conducted with independent variables, 20 portfolios were generated to determine a confidence interval for the validation techniques. However, in this case, 20 portfolios with zero correlation between $X_1$ and $X_2$ and 20 portfolios with a correlation of

![Fig. 2. Number (blue bars) and proportion (red line) of public companies default from 1950 to 2014. Source: CRSP database.](image-url)
which the correlation between the variables expected because the model was developed using a portfolio for
decrease in model performance with increased correlation, which was

and COMPUSTAT.

Table 15.

![Image](https://example.com/example.png)

4.3.1. Impact caused by correlation emergence

Among the 20 models with zero correlation that were used to deter-
mine the confidence interval, model DepGzero03, i.e., (KS = 0.3808,
AUROC = 0.7410, AR = 0.4820, PietraIndex = 0.1346, Brier = 0.0805,
CIEr = 0.1021, KL = 0.0320, IV = 0.7957, M = 0.8610) was chosen to be
applied to the samples with correlation variation. The results obtained
for the validation techniques for the samples with correlation are pre-
seated in Table 14.

All techniques, with the exception of the Brier score, exhibited a
decrease in model performance with increased correlation, which was
expected because the model was developed using a portfolio for
which the correlation between the variables X₁ and X₂ was zero. The
observed and estimated values for each of the ratings of some models
are presented in Table 15.

Based on the results presented in Table 15, it is possible to observe a
significant drop in model correctness with increased correlation be-
tween variables X₁ and X₂. The heterogeneity of the default rates
among the ratings also decreased, which made the entropy measures
sensitive. In general, the measures indicate that a model developed
with non-correlated variables that explain the default event with good
performance can have its performance compromised if a correlation be-
tween variables comes into existence. The increase in the correlation
generated a relevant impact on model correctness. For instance, taking
into consideration rating 6, the observed default rate overcame the
expected default rate in >60% of cases when the correlation was 0.55.

4.3.2. Impact caused by a change in the correlation

Among the 20 models with a correlation of 0.5 that were used to
determine the confidence interval, model DepG0.5/13, i.e., (KS = 0.3005,
AUROC = 0.6998, AR = 0.3996, PietraIndex = 0.1062, Brier = 0.0858,
CIEr = 0.0691, KL = 0.0224, IV = 0.5365, M = 0.8897) was

chosen to be applied to the samples with distinct correlations. The re-
sults obtained for the validation techniques for these samples are pre-
seated in Table 16.

Except for the Brier Score, all measures were sensitive to variations
in the correlation. However, because the model was developed with a
correlation of 0.5, a decrease in performance was expected if the
correlation values increased or decreased. All sensitive metrics,
with the exception of measure M, exhibited an increase in model per-
formance as the strength of the correlation was decreased. Measure M
exhibited lower model performance for both increased and decreased
correlation, although the values were outside the confidence interval
only for the decreased correlation. Consequently, Table 17 presents
the estimated and observed values of default within the ratings for
the models.

According to Table 17, the model correctness decreased as the corre-
lation was increased or decreased, which explains the values of measure
M for these cases. The default rates remained heterogenous along the
ratings as the correlations were increased or decreased, which explains
the small fluctuation in the entropy measures. The KS, AR, AUROC,
and Pietra measures were more sensitive to the order of the subjects.
Therefore, it can be concluded that although the correlation change
decalibrated the probability values that the model calculates, the order
of the good and bad subjects was not significantly changed.

5. Empirical sub-samples test

In this section, we developed a methodology for testing the adequa-
cy of techniques for the validation of DP with empirical data. The meth-
odology presented in Section 3, where a numerical simulation with
control over the variables (factor variables X₁, X₂ and the response var-
iable of default Y) is used to analyze the effects of the techniques for
the validation of DP, is useful when we try to understand the effectivens
in ideal situations. In particular, we can study the behavior of metrics
due to changes in the distribution, such as when the mean, the variance,
and the correlation are changed. However, the values of the input
parameters used for the numerical simulation results in Section 4 were not calibrated with market data; they were simply used as a numerical application of the methodology.

The purpose of this section is to provide a real-life application analysing the effects of techniques of validation for different market conditions. Comparing our analysis with that of Sobehart et al. (2000a), the contribution of this section is that we use a larger dataset and our results are provided when controlling factor variables \( X_1, X_2 \) by market situations, giving the risk manager an idea of the capacity of the techniques of validation of \( DP \) for historical market situations.

5.1. Methodology

The first part of the methodology involves defining what is a good borrower and a bad borrower. We also define the variables \( X_1, X_2, \) and \( Y \). We adopted Altman1968’s Altman1968 Z-score to determine which companies are in distress. The Z-score, which was initially developed for public manufacturing firms (Altman, 1968), was extended for private firms and non-manufacturing firms (Altman, 2000). In these three models financial ratios with fundamental balance sheet and income statement data are aggregated in a discriminant analysis study to determine companies in financial distress. There are five ratios in the original study of Altman (1968):

\[
T_1 = \frac{\text{Working Capital}}{\text{Assets}}; \\
T_2 = \frac{\text{Retained Earnings}}{\text{Assets}}; \\
T_3 = \frac{\text{EBIT}}{\text{Assets}}; \\
T_4 = \frac{\text{Market Value of Equity/ Liabilities}}{\text{Assets}}; \text{and,} \\
T_5 = \frac{\text{Sales}}{\text{Assets}}.
\]

where Assets and Liabilities are totalled and the regression is given by (17), as in Altman (1968):

\[
Z = 1.27T_1 + 1.4T_2 + 3.3T_3 + 0.6T_4 + 0.99T_5,
\]

This equation is used only for public manufacturing companies. The modified equation for private manufacturing companies is given by (18) as in Altman (2000):

\[
Z = 1.7T_1 + 1.8T_2 + 3.1T_3 + 0.4T_4 + 0.99T_5,
\]

where:

\[
T_1 = \frac{(\text{Curr Assets - Curr Liabilities})}{\text{Assets}}; \text{ and} \\
T_4 = \frac{\text{Book Value of Equity}}{\text{Liabilities}}.
\]

and for non-manufacturing and emerging market companies it is given by (19) as in Altman (2000):

\[
Z = 6.56T_1 + 3.26T_2 + 6.72T_3 + 1.05T_4.
\]

Altman (2000) found some critical values (\( Z < 1.81 \)) for public manufacturing, \( Z < 1.23 \) for private manufacturing, and \( Z < 1.1 \) for non-manufacturing companies. We define a bad borrower to be one with an Altman Z-score that is below the critical level and otherwise it is a good borrower. The factor variables are defined by the ratios, \( X_1 = T_5, X_2 = T_3, \) and the credit status is \( Y = 1 \) in the event of default and zero otherwise.

Extracting the \( DP \) from the market data is a challenging task. As defaults occur during the year, we model the \( DP \) using a discrete approach, defining an annualized dynamic model. Then, using an in-sample approach, the distribution of defaults will be such that if the company defaults during a year the values of the ratios \( X_1 \) and \( X_2 \) will be associated with \( Y = 1 \) and with \( Y = 0 \) otherwise.

5.2. Data

The data for the defaulting companies were proxied by the information on 30,686 delisted public companies that was extracted from the CRSP database and crossed with the financial fundamentals provided by the COMPSTAT database from January 1950 to December 2014. Although all the companies were public, some of them do not have...
enough liquidity to disclose a market value for their equity; in this case, we used the Altman's Z-score for private firms.

5.3. Results

The empirical test for the sensitivity of the validation measures identified AUROC as the most robust measure to detect changes in the distribution of DP. Fig. 2 shows the number of defaults and the proportion of defaults for the 30,686 companies.

5.3.1. Controlling factor variables

One of the main contributions of this research is the ability of the methodology to control the factor variables $X_1, X_2$ when studying the effects of changing conditions over the techniques of the validation of DP. We define three axes for the sensitivity analysis of the numerical simulation of Section 4: changes in the mean of $X_1$, changes in the volatility of $X_1$, and changes in the correlation between $X_1$ and $X_2$. For a discriminant analysis study we classify the market conditions each year by two groups in high-low levels of volatility and correlation and differ only in the mean value. We do the same with the volatility and the correlation. With these two smaller groups, we calculate the KS, AROC, AR, Pietra, CIER, KL, and IV measures for each sub-group and then for the whole population. The mean value of $X_1$ will then differ between the population, the high mean of the $X_1$ group and the low mean of the $X_1$ group.

5.3.2. Impact caused by variation of the $X_1$ mean

Table 18 shows the effects of changing the mean of $X_1$ over the techniques for the validation of DP. For public firms, an increase and decrease of the mean of $X_1$ was recognized by all measures. Fig. 3 shows the mean of $X_1$ for good and bad borrowers from 1950 to 2014. It is evident that the variable selected for the variable $X_1$ ($T_d$) discriminates properly in terms of the mean of the two groups. In the case of the public firms without market value, the years with lower 50th-percentile mean values were not detected for the AR, CIER, KL, and IV measures. For the non-manufacturing firms, AR and KL did not detect the change in the mean. AUROC and measure M proved to be the most sensitive measures in all cases.

5.3.3. Impact caused by changes in the variance of $X_1$

Table 19 shows the effects of changing volatility. For public firms, all measures captured the changes correctly. Nevertheless, KS, AR, Pietra, and KL show decreases in performance in the case of private manufacturing and non-manufacturing firms. AUROC proved to be the best measure.

Comparing the volatility of $X_1$ for good and bad borrowers through the years, Fig. 4 shows a sudden increase during the latter years for the distressed companies. When comparing the empirical distribution of $X_1$ for good borrowers and bad borrowers with Fig. 5, we notice that although there is a difference in the mean and the volatility, it is not possible to perform a simple discrimination of samples from both populations from the distribution given that both distributions overlap.

5.3.4. Impact caused by changes in the pair correlation of $X_1$ and $X_2$

As shown in Table 20, identifying changes in the correlation between $X_1$ and $X_2$ for validation measures was more difficult than for the mean and volatility. Correlation is a more complex characteristic of the distribution, and in Fig. 6 we can see that from the 1990s until the 2010s the difference in correlation of $X_1$ and $X_2$ between good and bad borrowers has not been a good discriminant. This result has impacts on the techniques of the validation of DP. Even in this context, AUROC and measure M again proved to be the best measures.

6. Conclusions

The techniques examined in this study evaluate different aspects of the models. Metrics traditionally used in the market, e.g., the KS statistic and AUROC, allow for the evaluation of the order of good and bad borrowers. However, these more traditional metrics are not adequate for assessing changes in the heterogeneity of default rates over the ratings, which are better evaluated using entropy measures. Metric M, unlike all of the others investigated, was quite sensitive to the accuracy of the estimated number of defaults in each rating. The empirical analysis

---

Table 20

<table>
<thead>
<tr>
<th>Firm sector</th>
<th>$X_1, X_2$ correlation adjustment</th>
<th>KS</th>
<th>AUROC</th>
<th>AR</th>
<th>Pietra</th>
<th>CIER</th>
<th>KL</th>
<th>IV</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public (manufacturing)</td>
<td>Population mean</td>
<td>0.169</td>
<td>0.292</td>
<td>0.139</td>
<td>0.0598</td>
<td>0.188</td>
<td>0.13</td>
<td>1.28</td>
<td>0.589</td>
</tr>
<tr>
<td></td>
<td>Increased</td>
<td>0.214</td>
<td>0.415</td>
<td>0.169</td>
<td>0.0758</td>
<td>0.255</td>
<td>0.175</td>
<td>1.71</td>
<td>0.792</td>
</tr>
<tr>
<td></td>
<td>Decreased</td>
<td>1.07</td>
<td>0.325</td>
<td>0.0722</td>
<td>0.0377</td>
<td>0.16</td>
<td>0.105</td>
<td>1.13</td>
<td>0.416</td>
</tr>
<tr>
<td>Public firms without equity market value (manufacturing)</td>
<td>Population mean</td>
<td>0.107</td>
<td>0.325</td>
<td>0.114</td>
<td>0.0587</td>
<td>0.189</td>
<td>0.12</td>
<td>1.33</td>
<td>0.545</td>
</tr>
<tr>
<td></td>
<td>Increased</td>
<td>0.166</td>
<td>0.443</td>
<td>0.0838</td>
<td>0.0455</td>
<td>0.255</td>
<td>0.17</td>
<td>1.8</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td>Decreased</td>
<td>0.129</td>
<td>0.458</td>
<td>-</td>
<td></td>
<td>0.182</td>
<td>1.82</td>
<td>0.851</td>
<td></td>
</tr>
<tr>
<td>Non-manufacturing</td>
<td>Population mean</td>
<td>0.0718</td>
<td>0.341</td>
<td>0.139</td>
<td>0.0254</td>
<td>0.126</td>
<td>0.0737</td>
<td>0.87</td>
<td>0.405</td>
</tr>
<tr>
<td></td>
<td>Increased</td>
<td>0.047</td>
<td>0.472</td>
<td>-</td>
<td></td>
<td>0.147</td>
<td>0.0731</td>
<td>0.59</td>
<td>0.527</td>
</tr>
<tr>
<td></td>
<td>Decreased</td>
<td>0.107</td>
<td>0.473</td>
<td>-</td>
<td></td>
<td>0.0532</td>
<td>0.202</td>
<td>0.13</td>
<td>1.33</td>
</tr>
</tbody>
</table>
suggests that changes in market conditions are better absorbed by AUROC. However, no validation technique was able to capture all of the impacts generated by changes throughout the tests performed. The techniques are therefore complementary, and none of the techniques can, a priori, be chosen over the others.

The validation techniques presented have been applied in different situations to check the conditions under which a particular technique was more appropriate than the others. Traditional empirical analysis based on a single sample allows one to study a specific market condition, but it is not possible to study the effects that arise from changes in certain parameters. For this purpose, simulated portfolios and controlled empirical samples were used in the stress testing analysis.

First, the distributions of good and bad borrowers along a score scale were simulated. For normal distributions of good and bad borrowers, all the techniques analyzed were effective, i.e., greater similarity of the distributions resulted in decreased performance of all techniques. However, when the simulations were performed using a normal distribution for good borrowers and a bimodal distribution for bad borrowers, it was observed that most traditional techniques were not adequate when the bimodal intensity was high. In these cases, entropy measures were more appropriate. These results are in accordance with those of Engelman et al. (2003). For portfolios with independent explanatory variables, the decreases in the model accuracy generated by changes in the mean values of the variables were captured only by Measure M, and the other measures were not very sensitive. For portfolios with dependent explanatory variables determined using Gaussian copulas, if the model was developed without dependency and dependency came into existence, all metrics except the Brier score exhibited a decrease in model performance.

Second, an empirical analysis with sub-samples of 30,686 delisted US companies was provided, to stress test the validation of DP methodologies. The results were similar to the results found with the numerical simulation, having the AUROC and measure M the most sensitive metrics to changes in the DP distribution main properties (expected return, volatility, and correlation).

Other possibilities for future studies is to investigate other forms of dependency among explanatory variables. Other suggestions for future studies include using other types of explanatory variables (binary, categorical, and beta, for example), changing the distribution used for the definition of the default event, and developing scores from other models, such as decision trees or discriminant analysis, for instance.

Acknowledgments

This research was supported in part by a grant from the Brazilian National Research Council (CNPq).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.irfa.2016.06.007.

References


Fig. 6. Correlation between X1 and X2 for good and bad borrowers of public manufacturing firms, using Altman’s Z-score for discrimination, from January 1950 to December 2014. Source: CRSP and COMPUSTAT.


