REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (24). Páginas 54–88.
Diciembre de 2017. ISSN: 1886-516X. D.L: SE-2927-06.
www.upo.es/revistas/index.php/RevMetCuant/article/view/2880

Selection of Variables in Small Business Failure Analysis: Mean Selection vs. Median Selection

Tascón, María T.
Departamento de Dirección y Economía de la Empresa
Universidad de León (España)
E-mail: m.tascon@unileon.es

Castaño, Francisco J.
Departamento de Dirección y Economía de la Empresa
Universidad de León (España)
E-mail: fjcasg@unileon.es

ABSTRACT

This paper focuses on one of the most determinant processes in business failure assessment: Variable selection. After a preselection of variables based on previous empirical literature, we perform a statistical variable selection on a sample of small firms using both mean and median differences. As the resulting variables differ in each test, we have performed a varied group of business failure assessment methods (linear discriminant analysis, quadratic discriminant analysis, logistic discriminant analysis, k-th nearest-neighbor discriminant analysis, logit, and probit) to identify the implications of using one test or the other. Our results show that the nature of the sample determines not only the statistical variable selection test, but the most appropriate methods to assess business failure, which constitutes our main contribution. Additionally, we contribute new evidence on the addition of qualitative information (payment incidents), with previous evidence for SMEs being scarce.

Keywords: small business failure; variable selection; discriminant analysis; probit; logit; financial ratios; qualitative information.
JEL classification: C14; G17; G33; L25; M41.
MSC2010: 62P20; 91G70; 91E45; 91G40.

Artículo recibido el 6 de julio de 2016 y aceptado 8 de febrero de 2017.

Selección de variables en el análisis de fracaso de empresas pequeñas: selección de medias frente a selección de medianas

RESUMEN

Este trabajo se ocupa de uno de los procesos más determinantes en la evaluación del fracaso empresarial: la selección de variables. Tras una preselección de variables basada en los resultados empíricos de la literatura previa, llevamos a cabo una selección estadística de variables sobre una muestra de empresas pequeñas, utilizando tanto diferencias en medias como diferencias en medianas. Como las variables resultantes difieren con el test, hemos utilizado un variado grupo de métodos de evaluación de fracaso empresarial (LDA, QDA, LogDA, KNNDA, *logit* y *probit*) con el fin de identificar las implicaciones de usar uno u otro test. Nuestros resultados muestran que la naturaleza de la muestra determina no solo el test de selección estadística de variables, sino también los métodos más apropiados para evaluar el fracaso empresarial, lo que constituye nuestra principal contribución. Además, el trabajo proporciona nueva evidencia sobre la adición de información cualitativa (incidencias de pago), siendo escasa la evidencia previa para pymes.

Palabras claves: fracaso en pequeñas empresas; selección de variables; análisis discriminante; *probit*; *logit*; ratios financieros; información cualitativa.
Clasificación JEL: C14; G17; G33; L25; M41.
MSC2010: 62P20; 91G70; 91E45; 91G40.

1. INTRODUCTION

The empirical testing of any hypothesis on business failure encounters problems with one or more of the three key elements that comprise the incomplete theory on business failure: The concept of failure, the model proposed and the selection of the variables chosen as discriminant factors.

Concerning the third issue, several authors (Platt and Platt, 1990; Keasey and Watson, 1991; Sueyoshi and Goto 2009b) have noted that the selection of an appropriate set of variables to implement a particular model may seem trivial but is, in fact, an important part of bankruptcy assessment. Notwithstanding, advances in this line are far less developed than those in methodologies to assess business failure.

Our work focuses on the variable selection process, making contributions to the selection based on statistical analysis. Considering the lack of theory to address our selection of potentially discriminant variables to be included in the statistical process, we check a wide sample of previous empirical works to identify those variables most frequently found significant, and the main economic features underlying the variables.

Once we have made the preselection of variables, we perform a statistical variable selection specifically oriented to the most problematic group of firms concerning data treatment: Small firms. We first apply the mean differences test, which is the most commonly used, and then the rank sum test (a median differences test), which has recently been introduced though not frequently used. Our results show that the significant variables differ radically between the tests. Consequently, the main contribution of our work is the identification of the implications of using mean of median differences to select variables when using a variety of models: Linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discriminant analysis (logDA), *k*th nearest-neighbor discriminant analysis (KNNDA), logit, and probit.

Our results show that the nature of the sample determines not only the most appropriate statistical variable selection test but also the most appropriate methods to assess business failure. Small firms are usually characterized by variables far from being normally distributed. Thus, our sample of small firms is a clear example of inadequacy of the mean differences test to select the most discriminant variables for failure, since the mean is not a good representation of the value for the variables. Furthermore, we find that parametric models obtain more biased results than non-parametric models when the variables studied are not normally distributed. We get to this conclusion after investigating how the results change when three types of changes are applied: (a) the treatment of data to eliminate outliers by winsorization; (b) the addition of a qualitative variable, i.e. legal incidents, contributing to previous literature by providing new supporting evidence on the role of this type of information in business failure analysis for small firms; and (c) the solving of data imbalance, avoiding sample bias by weighting failed firms instead of taking samples of not-failed firms.

2. LITERATURE REVIEW

The stream of literature related to our study is the one referred to business failure modeling, in which we consider both the appropriate variables and the methodological approximations for a good specification of the problem, which should permit an accurate classification of firms.

2.1. Variables

The absence of a formal theoretical model of the connections among the business failure, internal and/or external economic-financial factors, the economic concerns of the various related agents, and management performance has caused changes in the variable selection procedures. Over the last several decades, the variables included in business failure models have been selected less frequently based on economic reasoning and more frequently based on empirical experience gathered in previous studies (Scott, 1981). This is the common use in empirical works trying to accumulate evidence on the performance of known methods previously applied to business failure. In addition, the use of statistical methods, which can quantify the most discriminant statistically significant variables, is growing. However, those works proposing new methods or relevant innovations to any existing method tend to select the same variables as the empirical works they are compared to (Tascón and Castaño, 2012).

The statistical selection of variables may result in some drawbacks, as several authors have noted: Some empirical results are sample-specific and cannot be extrapolated to other samples (Keasey and Watson, 1991); interrelated factors may generate correlation problems, which could explain why the signs of some variables in certain models seem to be contrary to those that are economically sound (Balcaen and Ooghe, 2006); and some *a priori* relevant factors, such as liquidity, indebtedness, profitability, or activity, might be occasionally discarded (Dambolena and Khoury, 1980).

To identify the most discriminant variables used in the literature, we follow Tascon and Castaño[1] (2012) as our main reference because it complements the study of frequencies of individual variables with a study of frequencies of the economic features behind the variables. To determine whether the variables identified as more discriminant in general are also found to be discriminant in SMEs, we have analyzed[2] a sample of 23 empirical works, dated from 1987 to 2012, in which a wide variety of methodologies are applied (detailed information is included in Appendix A). Table 1 shows the economic features underlying the most frequently found significant variables, whereas Table 2 shows the most frequent discriminant variables, which are accounting ratios.

Table 1. Economic features underlying the variables

Economic features	Items	%	Num. ratios
Indebtedness	44	16.36	26
Asset structure	43	15.99	21
Profitability	41	15.24	22
Asset turnover	34	12.64	21
Equilibrium assets-liabilities	33	12.27	18
Profit margin	28	10.41	16
Other ratios	38	14.13	23
Variables (other than ratios)	8	2.97	8
Total	269	100.00	155

[1] Unlike similar previous reviews —for example, Daubie and Meskens (2002), and Bellovary *et al.* (2007)—, it includes Spanish empirical works in its international review and thereby constitutes a better base of comparison for the sample of Spanish small firms used in the subsequent empirical analysis. Tascón and Castaño (2012) analyze forty empirical works, from 1966 to 2009.

[2] Most empirical works internationally are performed on listed firms, but in Spain, which has a small stock exchange market, most empirical studies on business failure have been performed on SMEs. Tables containing all of the information are included in Appendix A.

Table 2. Most frequent ratios

EXPLANATORY VARIABLES	Num.works
Current Assets/Current Liabilities	8
Cash Flow/Total Assets	8
EBIT/Total Assets	7
Total Debt/Total Assets	7
Current Assets/ Total Assets	6
Sales/Total Assets	6
EBIT/Financial expenses	5
Net Income/Total Assets	4
Net income/Sales	4
Financial Expenses/Total Debt	4

In contrast to the reviews on firms of any size, ratios with retained profits are found to be considerably less discriminant, consistent with equity being less representative of the evolution of SMEs. In terms of economic features, indebtedness and asset structure appear as more discriminant than profitability. As expected, accounting variables are used even more frequently (97%) than other variables. After verified that they are a representative group of proxies for the strengths/weaknesses of SMEs as well, we take the 10 financial ratios[3] most frequently found significant in past theoretical and empirical literature on business failure, according to Tascón and Castaño (2012). Next, in this section, we provide with a description of these variables and briefly explain how they contribute to the assessment of business failure.

As a complement to accounting ratios, data from sources other than financial statements have been used in empirical works. Some non-financial variables with incremental explanatory ability include the existence of reservations/observations in the audit report (Peel and Peel, 1987), especially in small firms when the observations change from one period to another (Keasey and Watson, 1987); delays in the publication of financial statements (Ohlson, 1980; Peel and Peel, 1987; Somoza, 2002); the time passed from the end of the period to the date in which the financial statements were published (Peel *et al.*, 1986); and some incidents with payments (Altman *et al.*, 2008). Out of these interesting factors, we only could obtain information on one, legal incidents, which is included in our empirical analysis as a dummy variable to check the added ability of qualitative information to delimitate the potential business failure of SMEs.

TD/TA, total debt/total assets, is a leverage measure that indicates a long-term financial obligation. An increase in leverage would increase the probability of financial distress, as a reduction in cash flows could lead to insufficient funds to service debts, resulting in bankruptcy. In general, a positive relationship can be expected between indebtedness and business failure.

CA/CL, current assets/current liabilities, is a measure of short-term economic-financial equilibrium. As trade credits are a large proportion of small firms' liabilities

[3] Considering the fewer number of works on SMEs, we have decided to use the selection of most frequent ratios resulting from the general survey performed by Tascón and Castaño (2012), in which eight ratios are the same and another one proxies for the same feature as in our survey on SMEs. The first eight variables are the most frequent variables. The ninth and tenth ratios have been selected as the most frequent representatives for the following two features (profit margin and asset turnover) not proxied by the previous eight ratios.

(Altman *et al*., 2008), and credit to customers is extended in periods of financial stress, small firm bankruptcies would be primarily influenced by this type of debt.

EBIT/TA, earnings before interests and taxes/total assets, is a measure of economic profitability.

NI/TA, net income/total assets, is a more general measure of economic profitability.

CA/TA, current assets/total assets, is a measure of economic structure.

RP/TA, retained profit/total assets, is a measure of the cumulative profitability and may be a measure of the age of the firm. It indicates the ability of the firm to protect itself against potential future risks; therefore, a negative correlation with business failure can be expected.

FE/TD, financial expenses/total debt, is a measure of financial cost indicative of indebtedness and risk.

CF/TD, cash flow/total debt, is a measure of liquidity.

NI/SL, net income/sales, is a relevant driver of profitability. The greater the proportion of sales that becomes net income, the more efficient the firm is.

SL/TA, sales/total assets, acts as a multiplier of margin to increase profitability when positive.

Incidents indicates the occurrence of ‘event’ data, such as evidence of company default on credit agreements, trade credit payments and/or debts with any Public Administration organism. Two groups of incidents are considered: Judicial incidents and claims from Public Administration organisms. Judicial incidents gather all of those claims against a firm made to the court by a person of another firm. Claims from Public Administration organisms, such as Social Security, the Treasury or a city/town council, include concepts such as social securities, fines, and taxes that are unpaid at maturity. The variable takes the value 0 if a firm has not received any ‘public’ claim (neither judicial incidents nor Public Administration claims), 1 if the firm has been publicly claimed by only one of these groups of incidents, and 2 if the firm has appeared in both groups of public claims.

As a general idea, we can hypothesize that firms are financially more vulnerable to insolvency and hence, more likely to fail if they have poor liquidity measures, such as: Cash flows; poor profitability measures, such as net income or operating earnings; poor performance measures, such as margin or turnover; and higher indebtedness and risk measures, such as financial expenses and payment incidents.

2.2. Models

As several well-known works on literature reviews (Zavgren, 1983; Jones, 1987; Altman, 1993; Balcaen and Ooghe, 2006) and many empirical studies have compared different methodologies to identify their advantages and drawbacks, we merely report some of their conclusions referred to the methodologies used in our subsequent empirical analysis. In previous decades, business failure literature has been dominated by statistical discriminant analysis (DA) and logistic regression as econometric approaches (Aziz and Dar, 2006; Premachandra *et al.*, 2009).

DA uses a function to get a by-firm score that allows us to classify items into *a priori* defined classes, given a cutoff point. Variables can be quantitative and qualitative. Non-relevant variables for a univariate analysis may contribute significant information in DA models, which combine several variables (Altman, 1968). When the independent variables are quantitative, some assumptions are needed: Multivariate normality in the distribution of independent variables, homoscedasticity, representative sample sizes, as well as discrete, identifiable and not overlapping groups of firms, and finally specified failure probabilities and misclassification costs.

Several authors (Eisenbeis, 1977; Collins and Green, 1982; García Ayuso, 1995) show that accounting ratios hardly satisfy the normality hypothesis. When this limitation is overcome, either by transforming or eliminating extreme values, the model is distorted (Balcaen and Ooghe, 2006). Empirical works tend to transform data to reduce dispersion (Taffler, 1982) instead of using a quadratic model, which would solve the problem using the original data (Zavgren 1983), because the quadratic model only performs better than the linear model in very specific conditions (Balcaen and Ooghe, 2006).

A number of recent techniques have been developed to solve various drawbacks of the basic DA. In the current paper, we have used linear DA and quadratic DA as well as logistic DA and kth nearest-neighbor DA, the latter two being partially parametric and nonparametric techniques, respectively, unlike the first two.

The logistic regression (LR) is applied by Martin (1977) to avoid the drawbacks of DA. The binary logistic regression (binary logit) is a regression analysis in which the dependent variable takes values in the interval [0, 1] to indicate the probability of group membership; e.g., that of healthy firms or that of failed firms. Coefficients of independent variables can be interpreted as measuring the relative weight of the selected factors to explain the likelihood of failure generated by the model (Ohlson, 1980; Zavgren, 1985; Laitinen and Kankaanpää, 1999).

The logistic regression technique is more flexible than DA, as it can work with non-proportional samples (Hair *et al.*, 1999). It does not require normally distributed variables or equal dispersion matrices (Ohlson, 1980; Zavgren, 1983). However, it is assumed that the dependent variable is dichotomic and that both groups are identifiable, discrete and without overlap, which is difficult to fulfill for the population under study. In addition, the cost of misclassification errors has to be considered to establish a model cutoff point between healthy and failed firms, while these costs are highly subjective. Balcaen and Ooghe (2006) note that logistic analysis models are very sensible to multicollinearity, extreme values, and missing values.

A common source of problems in DA and LR models is the incorporation of *a priori* probabilities for the respective groups (Premachandra *et al.*, 2009). Eisenbeis (1977) suggests that an equally likely occurrence of failure and not-failure is assumed in a particular sample. Even though for matched pair sampling this is adequate, when a population or a random sample is analysed, it is unlikely that the population of potential failed firms will be 50%. In our study we have used equal priors in the displayed tables, but we have repeated all the empirical analyses using proportional priors, as a robustness analysis.

Probit models use an accumulative normal distribution instead of a logistic distribution, but previous comments on logit are applicable to the rest. Concerning the comparison between logit and probit, Baum (2006) find that a sample in which the proportion of probabilities is quite different will be sensitive to the choice of the cumulative distribution function. Therefore, as the logistic distribution (logit) has fatter tails, it should result in a better fit of the model.

Empirical literature shows other methods to assess business failure, such as different types of neural networks (Ravi Kumar and Ravi, 2007), data envelopment analysis (Premachandra *et al.*, 2009) and boosting (Sun *et al.*, 2014), for which the variable selection analysis made in this work could be applicable.

3. METHODOLOGY

This section contains a brief description of the methods used. The modeling is performed using Stata software. Besides, we statistically describe the 10 variables selected in Section 2.1. We then present the results obtained in the application of business failure models.

3.1. Models used for business failure assessment

We perform four varieties of discriminant analyses, as well as logit and probit. We use linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discriminant analysis (logDA) and Kth nearest neighbor discriminant analysis (KNNDA). With all four models we use equal prior probabilities[4], and hence geometric calculations (the other alternative, we have performed too as a robustness analysis, is considering proportional probabilities, in which case the probability of landing in a given group is proportional to the geometric area of that group, and they are certainly not equal). With KNNDA we want the straight line between the points, so we use the Euclidean distance.

Linear Discriminant Analysis

Starting from Fisher's approach (Fisher, 1936), LDA seeks the linear combination of the discriminating variables that provides maximal separation between the groups (originally two groups, but later extended to multiple groups). In the software used, maximal separation of groups is determined from an eigen analysis of $W^{-1}B$, where B is the between-group sum-of-squares and cross-products (SSCP) matrix, and W is the within-group SSCP matrix. The eigenvalues (characteristic values) and eigenvectors (characteristic vectors) provide what are called Fisher's linear discriminant functions.

[4] As classifying failed firms as not-failed is considered a bigger problem (due to higher costs) setting prior probability of failed firms higher than the proportion of these firms in the population, the classification of this group is more careful than the classification of the other one. Modifying the computational priors option we can factor in the cost of misclassifying firms.

Quadratic Discriminant Analysis

As introduced by Smith (1947), it is a generalization of LDA. Both LDA and QDA assume that the observations come from a multivariate normal distribution. Unlike the LDA's assumption that the groups have equal covariance matrices, QDA allows the groups to have different covariance matrices. Given the nature of our population, we have to take into account a restriction concerning the sample sizes: if a group is relatively small, the estimation of the covariance matrix for that group may not do a good job of representing the group's population covariance, leading to inaccuracies in classification.

Logistic Discriminant Analysis

LogDA is a partially parametric method falling between parametric discrimination methods such as LDA and QDA and nonparametric discrimination methods such as *k*th-nearest-neighbor (KNN) discrimination (Albert and Lesaffre, 1986). Instead of making assumptions about the distribution of the data within each group, logistic discriminant analysis is based on the assumption that the likelihood ratios of the groups have an exponential form. As multinomial logistic regression provides the basis for logDA, this type of DA is appropriate for both binary and continuous discriminating variables.

*k*th Nearest Neighbor Discriminant Analysis

Even though KNNDA can be found at least as far back as the fifties, advanced treatments are quite more recent (McLachlan, 2004). KNNDA is a nonparametric discrimination method based on the k nearest neighbors of each observation. Compared to other methods of discriminant analysis, KNNDA is able to distinguish irregular-shaped groups.

The *k*th nearest-neighbor algorithm (KNN) classifies objects (firms in our case) based on closest training examples in the feature space. The KNN is one of the simplest machine-learning algorithms: An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). For example, if $k = 1$, the object is simply assigned to the class of its nearest neighbor. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones (a common weighting scheme is to give a weight of $1/d$ to each neighbor, where d is the distance to the neighbor; this scheme is a generalization of linear interpolation). Stata offers us different similarity or dissimilarity measures, being the Euclidean distance the default option.

The neighbors are taken from the set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The KNN is sensitive to the local structure of the data.

Logit

The word "logit" was coined by Berkson (1944) and is analogous to the word "probit". Logit fits maximum likelihood models with dichotomous dependent (left-hand-side) variables coded as 0/1 (or, more precisely, coded as 0 and not-0). Stata makes a maximization supporting the Huber/White/sandwich estimator of the variance and its clustered version.

Stata checks data for collinear variables and omits them if collinearity exists. Therefore, no observations need to be eliminated and model fitting proceeds without the offending variable.

This technique is based in mean analysis. In fact, when the results are prob>chi2 = 0.01; this test would have declared the means different at the 1% level. The same as the parametric discriminant analyses, where a test with prob>F = 0.02 is declaring the means different at the 2% level.

Probit

Probit analysis originated in connection with bioassay, and the word probit (a contraction of "probability unit") was suggested by Bliss (1934). Probit fits maximum likelihood models with dichotomous dependent (left-hand-side) variables coded as 0/1 (more precisely, coded as 0 and not 0). Stata also checks data for collinear variables, following the same procedure as in logit.

3.2. Data and statistical variable selection

All of the data were obtained from *Iberinform Database*. In this database, the accounting variables are compiled from the Official Commercial Registry, information about the firm's status as failed (bankrupt, dissolved, extinct) is compiled from the Spanish Official Gazette of the Commercial Registry, and data on legal incidents are compiled from the Official State Gazette, Official Gazettes of Provinces, Official Gazettes of Autonomous Communities, Website of Social Security and the press (newspapers).

Our sample is made up of Spanish construction firms[5], classified as small according to official accounting standards[6]. The original file contained 3426 firms, of which 3230 had real values for ratio variables. Therefore, our final database consists of 3193 sets of accounts for not-failed firms and 37 for failed firms in 2008, in addition to qualitative data on the status of the firm and its public incidents on payment. The insolvency rate is approximately 1.15%. We first present the statistics and models using full accounting data, and then we present the statistics and modeling results for winsorized accounting ratios.

The variables pre-selected as explained in Section 2.1 are: Total debt/ total assets, current assets/ current liabilities, earnings before interests and taxes/ total assets, net income/ total assets, current assets/ total assets, retained profit/ total assets, financial expenses/ total debt, cash flow/ total debt, net income/ sales, and sales/ total assets. Table 3 displays the correlation matrix among the financial ratios and the additional firm-specific (but not accounting) variable. The correlation coefficients are relatively low, except for those between the profitability ratios and between several ratios divided by total assets.

[5] To develop all the methodologies used in this work, the sample has been selected by taking the whole population of construction firms in a Spanish Autonomous Community (Castilla and Leon) for which the number of firms was near the maximum number that the software could manage.

[6] The European Union definition was updated in 2003 (Commission Recommendation 2003/361/EC May 6). Enterprises qualify as small enterprises if they fulfill the following criteria: Fewer than 50 employees, less than or equal to €10million in sales, and less than or equal to €10 million in total assets.

Table 3. Panel A: Correlation matrix

	Failure	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	incident
Failure	1.0000											
r1= TD/TA	-0.0011	1.0000										
r2= CA/CL	-0.0029	-0.0023	1.0000									
r3= EBIT/TA	-0.0108	-0.9057*	0.0031	1.0000								
r4= NI/TA	-0.0052	-0.9773*	0.0023	0.9692*	1.0000							
r5= CA/TA	0.0147	-0.0024	0.0260	-0.0049	0.0354	1.0000						
r6= FE/TD	0.1593*	-0.0006	0.0375	0.0024	0.0018	-0.0126	1.0000					
r7= RP/TA	-0.0021	0.9995*	-0.0015	-0.8940*	-0.9714*	-0.0011	-0.0006	1.0000				
r8= CF/TD	0.0471*	-0.0036	0.4113*	0.0343	0.0189	0.0096	0.5289*	-0.0018	1.0000			
r9= NI/SL	0.0012	0.0002	-0.0000	0.0084	0.0044	-0.0062	0.0027	0.0001	0.0240	1.0000		
r10= SL/TA	0.0038	0.8138*	-0.0032	-0.6477*	-0.7883*	-0.3350*	-0.0144	0.8173*	-0.0037	0.0015	1.0000	
incidents	0.1794*	0.1057*	-0.0081	-0.0914*	-0.0978*	-0.0199	0.0041	0.1051*	-0.0227	0.0014	0.0808*	1.0000

Panel B: Correlation matrix, winsorized ratios

	Failure	wr1	wr2	wr3	wr4	wr5	wr6	wr7	wr8	wr9	wr10	incident
Failure	1.0000											
wr1= TD/TA	0.0546*	1.0000										
wr2= CA/CL	-0.0047	-0.2632*	1.0000									
wr3= EBIT/TA	-0.0906*	-0.5407*	0.0550*	1.0000								
wr4= NI/TA	-0.0999*	-0.5879*	0.0656*	0.9719*	1.0000							
wr5= CA/TA	0.0215	-0.1396*	0.1469*	0.1595*	0.1724*	1.0000						
wr6= FE/TD	0.0193	-0.0274	0.1200*	0.0830*	0.0011	-0.0811*	1.0000					
wr7= RP/TA	-0.0127	0.3276*	-0.0047	-0.1929*	-0.2100*	-0.0788*	-0.0615*	1.0000				
wr8= CF/TD	-0.0325	-0.4451*	0.2458*	0.5925*	0.6009*	0.0401	-0.0240	0.0182	1.0000			
wr9= NI/SL	-0.0384	-0.2497*	-0.1024*	0.3873*	0.3974*	0.0778*	-0.0056	-0.0880*	0.3025*	1.0000		
wr10= SL/TA	0.0189	0.1653*	-0.1781*	-0.0521*	-0.0957*	0.0208	0.1062*	0.1581*	0.0433	0.1265*	1.0000	
incidents	0.1794*	0.1497*	-0.0346	-0.1304*	-0.1421*	-0.0025	0.0542*	0.0081	-0.1134*	-0.0679*	-0.0026	1.0000

NOTES: TD/TA is total debt/ total assets; CA/CL is current assets/ current liabilities; EBIT/TA is earnings before interests and taxes/ total assets; NI/TA is net income/ total assets; CA/TA is current assets/ total assets; RP/TA is retained profit/ total assets; FE/TD is financial expenses/ total debt; CF/TD is cash flow/ total debt; NI/SL is net income/ sales; and SL/TA is sales/ total assets.
* significance at the 1% level.

To avoid the influence of extreme values on some statistics but retaining the same number of observations, we have winsorized the ratios. For this procedure, the bottom 1% of the values are set equal to the 1st percentile, while the upper 1% are set equal to the 99th percentile. This technique reduces the effect of outliers and brings the corrected data mean closer to the median with a reduction in the standard deviation. After winsorizing the variables, all correlation coefficients are relatively low, except for that between the profitability ratios, r3 and r4 (Table 3, Panel B).

In Table 4, we report the mean, median, standard deviation and skewness of the variables for failed and not-failed firms separately. It can be appreciated that using the full accounting data mean values is quite different than using the median values and both the standard deviation and skewness are high. Once the ratios have been winsorized at the 1% level (Table 4, Panel B), the mean values are closer to the median values and both the standard deviations and the skewness have decreased considerably. The variables express the expected signs.

A well-known method (Peel *et al.*, 1986; Grice and Ingram, 2001; Sueyoshi and Goto, 2009a, 2009b, 2009c; Jin *et al.*, 2011) to select discriminant variables is the analysis of the mean differences (simple univariate analysis) between failed firms and not-failed firms. Table 5 (4th column) shows the results with the t-statistic results. The ratio of financial expenses to total debt is 42.96% and 3.19% for failed firms and not-failed firms, respectively. On the other hand, the ratio of accounting cash flow to total debt is 141.86% and 18.90% for failed firms and not-failed firms, respectively. This spurious result derives from an extraordinary value of 8.328% for this ratio in a specific one failed firm. Once this firm is dropped, the mean value equals is -0.85, being with the difference in means being equally significant. Incidents seem to be a relevant discriminant factor, showing mean values of 0.62 for failed firms and 0.07 for not-failed firms. The difference in means between the two groups is significant at the 1% level for these three variables. The non-default group outperforms the default group in each factor.

Table 4. Panel A: Descriptive statistics

Failure	variable	n	Mean	median	St.dev.	skewness
not-failed	r1= TD/TA	3193	1.464248	0.7783218	35.13914	56.26852
not-failed	r2= CA/CL	3193	4.992888	1.274884	65.81954	30.69193
not-failed	r3= EBIT/TA	3193	-0.0345993	0.0444596	2.19314	-41.647
not-failed	r4= NI/TA	3193	-0.0876411	0.0159537	3.849023	-51.79087
not-failed	r5= CA/TA	3193	0.7268043	0.7864155	0.3425963	9.195928
not-failed	r6= FE/TD	3193	0.0319486	0.0214282	0.0583653	13.70667
not-failed	r7= RP/TA	3193	0.810277	0.1062388	32.54369	56.31467
not-failed	r8= CF/TD	3193	0.1890471	0.0608538	2.319081	23.02114
not-failed	r9= NI/SL	3193	-0.5259805	0.0124977	26.23841	-48.29314
not-failed	r10= SL/TA	3193	1.815489	1.381083	22.46602	18.59462
not-failed	incidents	3193	0.0720326	0	0.3133578	4.695205
Failed	r1= TD/TA	37	1.09844	0.9751387	1.010565	3.7524
Failed	r2= CA/CL	37	3.23058	1.057278	11.17866	5.727102
failed	r3= EBIT/TA	37	-0.2554667	-0.0164002	0.7124744	-2.656331
failed	r4= NI/TA	37	-0.2745696	-0.0354618	0.6439843	-2.230388
failed	r5= CA/TA	37	0.774168	0.8938961	0.2734852	-1.460193
failed	r6= FE/TD	37	0.4296611	0.032181	2.422027	5.832577
failed	r7= RP/TA	37	0.1619858	0.0299024	0.6085422	3.317699
failed	r8= CF/TD	37	1.418611	-0.0286602	14.66338	4.637255
failed	r9= NI/SL	37	-0.2360054	-0.0441338	0.5619837	-3.127511
failed	r10= SL/TA	37	2.612674	1.475444	5.620027	5.245937
failed	incidents	37	0.6216216	0	0.720777	0.694475

Panel B: Descriptive statistics, winsorized ratios

failure	variable	n	Mean	median	St.dev.	skewness
not-failed	wr1= TD/TA	3193	0.7876563	0.7783218	0.4490908	2.483903
not-failed	wr2= CA/CL	3193	2.783909	1.274884	5.769318	5.38778
not-failed	wr3= EBIT/TA	3193	0.022242	0.0444596	0.220357	-2.688866
not-failed	wr4= NI/TA	3193	-0.0056056	0.0159537	0.1911801	-3.040909
not-failed	wr5= CA/TA	3193	0.7263177	0.7864155	0.2373353	-0.9497068
not-failed	wr6= FE/TD	3193	0.0297533	0.0214282	0.032074	1.978609
not-failed	wr7= RP/TA	3193	0.1956719	0.1062388	0.3814027	2.493171
not-failed	wr8= CF/TD	3193	0.1211922	0.0608538	0.3389109	2.552918
not-failed	wr9= NI/SL	3193	-0.0638995	0.0124977	0.475318	-6.155788
not-failed	wr10= SL/TA	3193	1.661431	1.381083	1.448189	2.268183
not-failed	incidents	3193	0.0720326	0	0.3133578	4.695205
failed	wr1= TD/TA	37	1.019683	0.9751387	0.6442166	1.777834
failed	wr2= CA/CL	37	2.527121	1.057278	6.940696	5.563727
failed	wr3= EBIT/TA	37	-0.1692527	-0.0164002	0.4425078	-1.603176
failed	wr4= NI/TA	37	-0.1896237	-0.0354618	0.4113469	-1.710292
failed	wr5= CA/TA	37	0.7744578	0.8938961	0.2726751	-1.451836
failed	wr6= FE/TD	37	0.0355791	0.032181	0.0330995	2.367284
failed	wr7= RP/TA	37	0.1500051	0.0299024	0.4884795	2.808293
failed	wr8= CF/TD	37	0.0164868	-0.0286602	0.5752716	2.43246
failed	wr9= NI/SL	37	-0.2360054	-0.0441338	0.5619837	-3.127511
failed	wr10= SL/TA	37	1.919005	1.475444	1.894715	1.989572
failed	incidents	37	00.6216216	0	0.720777	0.694475

NOTES: TD/TA is total debt/ total assets; CA/CL is current assets/ current liabilities; EBIT/TA is earnings before interests and taxes/ total assets; NI/TA is net income/ total assets; CA/TA is current assets/ total assets; RP/TA is retained profit/ total assets; FE/TD is financial expenses/ total debt; CF/TD is cash flow/ total debt; NI/SL is net income/ sales; and SL/TA is sales/ total assets.

Considering the results obtained, failed firms generate lower cash flows, have higher financial expenses, and suffer legal incidents in a higher proportion than not-failed firms. The mean differences test shows the expected results considering the correlation analysis of the explanatory variables with the failure variable.

However, this type of analysis (mean differences) requires that the populations being compared be normally distributed. In addition, a substantial proportion of the accounting variables shows high-dispersion distributions, in which the median values are more representative than the mean values. Hence, we have performed a nonparametric test to compare two groups: the Wilcoxon rank sum test (also called Mann-Whitney-Wilcoxon). This test has been used to select variables in a bankruptcy assessment process in recent works, such as Premachandra *et al.* (2009). Its main advantages are that it does not assume that the populations being compared are normally distributed, it uses only ranks, and it is not sensitive to outliers.

Table 5. Descriptive statistics by success/failure status of firms and tests on differences

Variable	**Failed firms** **n = 37** Mean **Median** (Std. dev.) [Skewness]	**Not-failed firms** **n = 3193** Mean **Median** (Std. dev.) [Skewness]	**Difference in means** (t-statistic)	**Two sample Wilcoxon rank-sum test between medians (Z)** [p-value> (Z)]
r1= TD/TA	1.0984 **0.9751** (1.0105) [3.7524]	1.464 **0.7783** (35.1391) [56.2685]	0.3658 (0.0633)	**-3.551** [0.0004]***
r2= CA/CL	3.2305 **1.0572** (11.1786) [5.7271]	4.9928 **1.2748** (65.8195) [30.6919]	1.7623 (0.1628)	**1.715** [0.0864]*
r3= EBIT/TA	-0.2554 **-0.0164** (0.7124) [-2.6563]	-0.0345 **0.0444** (2.1931) [-41.647]	0.2208 (0.6121)	**3.093** [0.0020]***
r4= NI/TA	-0.2745 **-0.0354** (0.6439) [-2.2303]	-0.0876 **0.0159** (3.8490) [-51.7908]	0.1869 (0.2953)	**4.112** [0.0000]***
r5= CA/TA	0.7741 **0.8938** (0.2734) [-1.4601]	0.7268 **0.7864** (0.3425) [9.1959]	-0.0473 (-0.8378)	**-1.977** [0.0480]**
r6= FE/TD	0.4296 **0.3221** (2.4220) [5.8325]	0.0319 **0.0214** (0.0583) [13.7066]	-0.3977 (-9.1707)***	**-1.641** [0.1007]
r7= RP/TA	0.1619 **0.0299** (0.6085) [3.3176]	0.8102 **0.1062** (32.5436) [56.3146]	0.6482 (0.1212)	**2.272** [0.0231]**
r8= CF/TD	1.4186 **-0.0286** (14.6633) [4.6372]	0.1890 **0.0608** (2.3190) [23.0211]	-1.2295 (-2.6770)***	**4.476** [0.0000]***
r9= NI/SL	-0.2360 **-0.0441** (0.5619) [-3.1275]	-0.5259 **0.0124** (26.2384) [-48.2931]	-0.2899 (-0.0672)	**4.302** [0.0000]***
r10= SL/TA	2.6126 **1.4754** (5.6200) [5.2459]	1.8154 **1.3810** (22.4660) [18.5946]	-0.7971 (-0.2157)	**-0.436** [0.6625]
Incidents	0.6216 **0** (0.7207) [0.6944]	0.0720 **0** (0.3133) [4.6952]	-0.5495 (-10.3621)***	**-10.838** [0.0000]***

NOTES: The full sample contains 3,230 firm observations during 2008. Variable definitions: TD/TA is total debt/ total assets; CA/CL is current assets/ current liabilities; EBIT/TA is earnings before interests and taxes/ total assets; NI/TA is net income/ total assets; CA/TA is current assets/ total assets; RP/TA is retained profit/ total assets; FE/TD is financial expenses/ total debt; CF/TD is cash flow/ total debt; NI/SL is net income/ sales; and SL/TA is sales/ total assets.
* significance at the 10% level based on a two-tailed test
** significance at the 5% level based on a two-tailed test
*** significance at the 1% level based on a two-tailed test.

Table 6. Descriptive statistics by success/failure status of firms, tests of differences, winsorized ratios

Variable	**Failed firms** **n = 37** Mean **Median** (Std. dev.) [Skewness]	**Not-failed firms** **n = 3193** Mean **Median** (Std. dev.) [Skewness]	**Difference in means** (t-statistic)	**Two sample Wilcoxon rank-sum test between medians (Z)** [p-value> (Z)]
wr1= TD/TA	1.0196 **0.9751** (0.6442) [1.7778]	0.7876 **0.7783** (0.4490) [2.4839]	-0.2320 (-3.1063)***	**-3.551** [0.0004]***
wr2= CA/CL	2.5271 **1.0572** (6.9406) [5.5637]	2.7839 **1.2748** (5.7693) [5.3877]	0.2567 (0.2685)	**1.714** [0.0866]
wr3= EBIT/TA	-0.1692 **-0.0164** (0.4425) [-1.6031]	0.0222 **0.0444** (0.2203) [-2.6888]	0.1914 (5.1689)***	**3.094** [0.0020]***
wr4= NI/TA	-0.1896 **-0.0354** (0.4113) [-1.7102]	-0.0056 **0.0159** (0.1911) [-3.0409]	0.1840 (5.7069)***	**4.118** [0.0000]***
wr5= CA/TA	0.7744 **0.8938** (0.2726) [-1.4518]	0.7263 **0.7864** (0.2373) [-0.9497]	-0.0481 (-1.2245)	**-1.976** [0.0481]**
wr6= FE/TD	0.0355 **0.0321** (0.3309) [2.3672]	0.0297 **0.0214** (0.0320) [1.9786]	-0.0058 (-1.0981)	**-1.638** [0.1015]
wr7= RP/TA	0.1500 **0.0299** (0.4884) [2.8082]	0.1956 **0.1062** (0.3814) [2.4931]	0.0456 (0.7215)	**2.270** [0.0232]**
wr8= CF/TD	0.0164 **-0.0286** (0.5752) [2.4324]	0.1211 **0.0608** (0.3389) [2.5529]	0.1047 (1.8491)*	**4.474** [0.0000]***
wr9= NI/SL	-0.2360 **-0.0441** (0.5619) [-3.1275]	-0.0638 **0.0124** (0.4753) [-6.1557]	0.1721 (2.1849)**	**4.302** [0.0000]***
wr10= SL/TA	1.9190 **1.4754** (1.8947) [1.9895]	1.6614 **1.3810** (1.4481) [2.2681]	-0.2575 (-1.0714)	**-0.434** [0.6641]
incidents	0.6216 **0** (0.7207) [0.6944]	0.0720 **0** (0.3133) [4.6952]	-0.5495 (-10.3620)***	**-10.838** [0.0000]***

NOTES: The full sample contains 3,230 firm observations during 2008. Variable definitions: TD/TA is total debt/ total assets; CA/CL is current assets/ current liabilities; EBIT/TA is earnings before interests and taxes/ total assets; NI/TA is net income/ total assets; CA/TA is current assets/ total assets; RP/TA is retained profit/ total assets; FE/TD is financial expenses/ total debt; CF/TD is cash flow/ total debt; NI/SL is net income/ sales; and SL/TA is sales/ total assets.
* significance at the 10% level based on a two-tailed test
** significance at the 5% level based on a two-tailed test
*** significance at the 1% level based on a two-tailed test.

On the other hand, it also has some disadvantages: Nonparametric methods are often less sensitive (powerful) for finding true differences because they throw away information (they use only ranks); they need the full data set, not just summary statistics; and the results do not include any confidence intervals quantifying the range of possibility for true difference between populations.

Results from a two-way Wilcoxon rank-sum test indicate that the median values for the failed and not-failed firms are significantly different for the following variables: r1, r3, r4, r8, r9, and legal incidents at the 1% level; r5 and r7 at the 5% level; and r2 at the 10% level. These results suggest that, except for r10 and r6, the accounting ratios used in the analysis are appropriate for classifying the firms as failed or healthy[7]. Using the differences in medians, we appreciate that the values in r8 are not affected by outliers, unlike in the case in which the differences in means are used instead.

Using winsorized ratios (see Table 6), the most discriminant variables identified by the mean test change considerably, becoming almost the same as those identified by the median test, which remain unchanged. This finding supports the use of the median test on dispersed and skewed samples of data, such as SME samples.

3.3. Business failure assessment

In this section we apply four discriminant analysis methods (LDA, QDA, LogDA and KNNDA) and two binary regression models (logit and probit). We perform every model with the entire group of ratio variables and with smaller groups of variables using the significant ratios for median and mean selections at the 5% and 1% levels of significance. All of the models are re-estimated with the inclusion of the qualitative non-accounting variable (legal incidents), which is significant at the 1% level in all cases.

For the LDA model (see Table 7), the best percentage of well-classified firms is obtained with the mean selection. The inclusion of the incidents variable considerably improves the selection of failed firms, while it reduces the well-classified not-failed firms in a higher quantity, resulting in a worse global percentage (except for 7 ratios)

Table 7. Linear discriminant analysis (LDA)

	All ratios		**Median selection (5% and 1%)**				**Mean selection(1%)**	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	2 ratios	2 ratios + inc.
	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)
Not-failed n=3,173	3,158 (98.90)	3,003 (94.05)	2,636 (82.56)	3,004 (94.08)	3,102 (97.15)	3,005 (94.11)	3,171 (99.31)	3,007 (94.17)
Failed n= 37	2 (5.41)	20 (54.05)	19 (51.35)	19 (51.35)	6 (16.22)	19 (51.35)	2 (5.41)	20 (54.05)
Total well classified n= 3,210	3,160 (**98.44**)	3,023 (94.17)	2,655 (82.71)	3,023 (94.17)	3,108 (96.82)	3,024 (**94.21**)	3,173 (**98.84**)	3,027 (**94.29**)

NOTES: This table contains results obtained with the LDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/ total assets; r2=CA/CL, current assets/ current liabilities; r3=EBIT/TA, earnings before interests and taxes/ total assets; r4=NI/TA, net income/ total assets; r5=CA/TA, current assets/ total assets; r6=RP/TA, retained profit/ total assets; r7=FE/TD, financial expenses/ total debt; r8=CF/TD, cash flow/ total debt; r9=NI/SL, net income/ sales; and r10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

QDA (see Table 8) shows a big discriminant power concerning failed firms for the ten variables, highly improved when the incident variable is added; this ability is kept with less variables when selected by medians (7 and 5 variables), but not when the variables were

[7] r10 and r6 are not the less significant ratios in previous literature. This finding supports the idea of taking a varied group of variables representing the main economic features of firms and, if necessary, making a second selection, specific for the sample.

selected by means (2 variables). By contrast, not failed firms are poorly identified with 10, 7 or 5 ratios, without the incidents variable, and just a bit better with the incidents variable; while the discriminant power of the two variables selected by means is excellent for not-failed firms and very poor for failed firms.

Table 8. Quadratic discriminant analysis (QDA)

	All ratios		**Median selection (5% and 1%)**				**Mean selection (1%)**	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	2 ratios	2 ratios + inc.
	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)
Not-failed n=3,173	378 (11.84)	390 (12.21)	134 (4.20)	142 (4.45)	98 (3.07)	103 (3.23)	3,161 (99.00)	3,111 (97.43)
Failed n= 37	36 (97.30)	37 (100.00)	35 (94.59)	36 (97.30)	35 (94.59)	35 (94.59)	2 (5.41)	7 (18.92)
Total well classified n= 3,210	414 (12.90)	427 (13.30)	169 (5.26)	178 (5.55)	133 (4.14)	138 (4.30)	3,163 (**98.54**)	3,118 (**97.13**)

NOTES: This table contains results obtained with the QDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/ total assets; r2=CA/CL, current assets/ current liabilities; r3=EBIT/TA, earnings before interests and taxes/ total assets; r4=NI/TA, net income/ total assets; r5=CA/TA, current assets/ total assets; r6=RP/TA, retained profit/ total assets; r7=FE/TD, financial expenses/ total debt; r8=CF/TD, cash flow/ total debt; r9=NI/SL, net income/ sales; and r10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

The results obtained with LogDA (see Table 9) show that not-failed firms are better identified with a fewer number of variables, especially when the incidents variable is added. On the contrary, failed firms are better identified with a higher number of ratios. With 10 or 7 ratios the incidents variable even takes away discriminant power, but when less ratios are used, the addition of the incidents variable considerably improves the discriminant power of the models. In fact, for 7, 5, and 2 ratios with the incidents variable, the discriminant power on failed firms is stable.

Table 9. Logistic discriminant analysis (LogDA)

	All ratios		**Median selection (5% and 1%)**				**Mean selection (1%)**	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	2 ratios	2 ratios + inc.
	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)
Not-failed n=3,173	1,896 (59.38)	2,921 (91.48)	1,691 (52.96)	2,998 (93.89)	2,952 (92.45)	3,006 (94.14)	2,839 (88.91)	2,995 (93.80)
Failed n= 37	26 (70.27)	23 (62.16)	28 (75.68)	19 (51.35)	6 (16.22)	19 (51.35)	8 (21.62)	20 (54.05)
Total well classified n= 3,210	1,922 (59.88)	2,944 (91.71)	1,719 (53.55)	3,017 (**93.99**)	2,958 (**92.15**)	3,025 (**94.24**)	2,847 (**88.69**)	3,015 (93.93)

NOTES: This table contains results obtained with the LogDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/ total assets; r2=CA/CL, current assets/ current liabilities; r3=EBIT/TA, earnings before interests and taxes/ total assets; r4=NI/TA, net income/ total assets; r5=CA/TA, current assets/ total assets; r6=RP/TA, retained profit/ total assets; r7=FE/TD, financial expenses/ total debt; r8=CF/TD, cash flow/ total debt; r9=NI/SL, net income/ sales; and r10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

Using KNNDA (see Table 10), we can appreciate an excellent discriminant power, both for not-failed and failed firms. All the failed firms are identified with any selection of variables. As for the not-failed firms, when only ratios are taken, any median selection (7 or 5 ratios) performs better than the mean selection (2 ratios). Furthermore, 5 ratios produce a better classification than 7 ratios, and 7 ratios perform better than 10 ratios. This type of discriminant analysis is nonparametric; thus, it does not rely on data belonging to any particular distribution. With nonparametric methodology the median selection performs better than the mean selection.

Table 10. *k*th nearest-neighbor discriminant analysis (KNNDA)

	All ratios		Median selection (5% and 1%)				Mean selection (1%)	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	2 ratios	2 ratios + inc.
	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)
Not-failed n=3,173	3,151 (98.68)	3,163 (99.06)	3,153 (98.75)	3,161 (99.00)	3,163 (99.06)	3,156 (98.84)	3,150 (98.65)	3,160 (98.97)
Failed n= 37	37 (100.00)	37 (100.00)	37 (100.00)	37 (100.00)	37 (100.00)	37 (100.00)	37 (100.00)	37 (100.00)
Total well classified n= 3,210	3,188 (99.31)	3,200 (**99.69**)	3,190 (**99.38**)	3,198 (**99.63**)	3,200 (**99.69**)	3,193 (99.47)	3,187 (99.28)	3,197 (99.60)

NOTES: This table contains results obtained with the KNNDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/ total assets; r2=CA/CL, current assets/ current liabilities; r3=EBIT/TA, earnings before interests and taxes/ total assets; r4=NI/TA, net income/ total assets; r5=CA/TA, current assets/ total assets; r6=RP/TA, retained profit/ total assets; r7=FE/TD, financial expenses/ total debt; r8=CF/TD, cash flow/ total debt; r9=NI/SL, net income/ sales; and r10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

Table 11. Panel A: Logit

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios +inc.	2 ratios	2 ratios +inc.
	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)
Not-failed n=3,193	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3,193 (100.00)	3,193 (100.00)
Failed n= 37	2 (5.41)	2 (5.41)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (5.41)	2 (5.41)
Total well classifie n= 3,230	3,195 (98.92)	3,195 (98.92)	3,193 (98.85)	3,193 (98.85)	3,193 (98.85)	3,193 (98.85)	3,195 (98.92)	3,195 (98.92)

Panel B: Probit

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios +inc.	2 ratios	2 ratios +inc.
	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)	number (percentage)
Not-failed n=3,193	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3,193 (100.00)	3,193 (100.00)
Failed n= 37	1 (2.70)	1 (2.70)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (5.41)	1 (2.70)
Total well classifie n= 3,230	3,194 (98.89)	3,194 (98.89)	3,193 (98.85)	3,193 (98.85)	3,193 (98.85)	3,193 (98.85)	3,195 (98.92)	3,194 (98.89)

Panel C: Logit and probit, tests

	All ratios		Median selection (5% and 1%)				Mean selection (1%)	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios +inc.	2 ratios	2 ratios +inc.
LOGIT								
Prob>chi2	0.0038***	0.0000***	0.4951	0.0000***	0.5832	0.0000***	0.0003***	0.0000***
McFadden's R2	**0.0642**	**0.1674**	0.0158	0.1216	0.0093	0.1157	**0.0406**	**0.145**
McFadden's Adj R2	**0.01**	**0.108**	-0.024	0.077	-0.02	0.081	**0.026**	**0.125**
BIC	-25631.91	-25665.59	-25636.61	-25671.31	**-25650.15**	**-25685.06**	**-25687.02**	**-25721.17**
Area under ROC	**0.7101**	**0.8052**	**0.6737**	0.7894	0.5288	0.7099	0.6921	**0.8116**
PROBIT								
Prob>chi2	0.0032***	0.0000***	0.5916	0.0000***	0.6121	0.0000***	0.0003***	0.0000***
McFadden's R2	**0.0653**	**0.1712**	0.0138	0.1252	0.0088	0.1215	**0.0405**	**0.1512**
McFadden's Adj R2	**0.011**	**0.112**	-0.026	0.081	-0.021	0.087	**0.026**	**0.131**
BIC	-25632.39	-25667.13	-25635.78	-25672.74	**-25649.95**	**-25687.41**	**-25686.99**	**-25723.69**
Area under ROC	**0.7145**	**0.8077**	**0.6835**	0.784	0.5612	0.7118	0.6909	**0.8105**

NOTES: Ten ratios include: r1=TD/TA, total debt/total assets; r2=CA/CL, current assets/current liabilities; r3=EBIT/TA, earnings before interests and taxes/total assets; r4=NI/TA, net income/total assets; r5=CA/TA, current assets/total assets; r6=RP/TA, retained profit/total assets; r7=FE/TD, financial expenses/total debt; r8=CF/TD, cash flow/total debt; r9=NI/SL, net income/sales; and r10=SL/TA, sales/total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

Logit and probit results (see Table 11) show that:

- For the same selection of variables, probit tends to perform better than logit.

- McFadden’s R2 decreases as we drop variables.
- At the same level of significance (1%) the mean selection performs better than the median selection. In fact, using the median selection to reduce the number of ratios (5 or 7 ratios) the models are not significant.
- The addition of the incidents variable considerably improves the quality of the models, making any selection of ratios significant.
- Once we have computed McFadden’s adjusted R2 to avoid the influence of the number of variables, we note that the discriminant ability of the models decreases[8]. The mean selection (two variables) shows a better discriminant power than the model with 10 ratios. And the same pattern remains when the incidents variable is added.
- Another possibility to compare logit and probit models is using BIC (Bayesian Information Criterion), based on the likelihood of the models and their degrees of freedom, for which the higher the negative value, the better. The results of the BIC test are better for those models with fewer variables, that is, the mean selection (two ratios). Curiously, the following better models are those with five variables, selected by medians.
- Finally, we have used the area under ROC curves to compare the goodness of the fit of the different models[9]. In this case, it seems that methods with more variables tend to perform better than methods with a lower number of variables, except for the model with two ratios and incidents, which performs the best.

3.4. Winsorized data analysis

After winsorizing ratios, the LDA results (see Table 12) exhibit an irregular pattern. Now, the best percentage of well-classified firms is obtained for the mean selection at the 1% level after the incidents variable is added, but it is only slightly better than the percentage for the model with 10 ratios and the incidents variable. For models using only ratios as independent variables, the use of fewer variables better classifies not-failed firms but classifies failed firms worse, except for the model with three ratios. The inclusion of the incidents variable improves the selection of failed firms, making the number of well-classified failed firms stable, while the reduction of well-classified not-failed firms is small for the median selection and negligible for the mean selection at the 10% level. In general, for our sample, after winsorizing the ratios, LDA tends to select failed firms better but non-failed firms worse.

[8] The adjusted count R2 eliminates from the account those cases related to the higher marginal. It indicates the proportion of correct classifications further derived from putting all firms in the group with the higher marginal. This test would show if the model has a real ability to generate a good classification.

[9] Receiver operating characteristics (ROC) curves are used for example by Altman *et al.* (2008) to assess the performance of logit models. In a ROC curve, the true positive rate (sensitivity) is plotted as a function of the false positive rate (100-specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold to discriminate failed and not-failed firms. The area under the ROC curve is a measure of accuracy in the discrimination, where the value 1 means a perfect model. Both the Gini coefficient and the Kolmorogov-Smirnov (K-S) statistic, commonly used by scoring analysts, derive from ROC curves.

Table 12. Linear discriminant analysis (LDA), winsorized ratios

	All ratios		Median selection				Mean selection			
			5%		1%		5%		1%	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	5 ratios	5 ratios + inc.	3 ratios	3 ratios + inc.
	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)
Not-failed n=3,173	2,776 (86.94)	2,939 (92.05)	2,798 (87.63)	2,935 (91.92)	2,836 (88.82)	2,934 (91.89)	2,836 (88.82)	2,934 (91.89)	2,822 (88.38)	2,944 (92.20)
Failed n= 37	18 (48.65)	22 (59.46)	17 (45.95)	22 (59.46)	13 (35.14)	22 (59.46)	13 (35.14)	22 (59.46)	13 (35.14)	21 (56.76)
Total well classified n= 3,210	2,794 (87.04)	2, 961 (**92.24**)	2,815 (87.69)	2,957 (92.12)	2,849 (**88.75**)	2,956 (92.09)	2,849 (**88.75**)	2,956 (92.09)	2,835 (88.32)	2,965 (**92.37**)

NOTES: This table contains results obtained with the LDA model taking equal priors. Ten ratios include: wr1=TD/TA, total debt/ total assets; wr2=CA/CL, current assets/ current liabilities; wr3=EBIT/TA, earnings before interests and taxes/ total assets; wr4=NI/TA, net income/ total assets; wr5=CA/TA, current assets/ total assets; wr6=RP/TA, retained profit/ total assets; wr7=FE/TD, financial expenses/ total debt; wr8=CF/TD, cash flow/ total debt; wr9=NI/SL, net income/ sales; and wr10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: wr1, wr3, wr4, wr8, wr9; and at the 5% level includes: wr1, wr3, wr4, wr5, wr7, wr8, wr9. The mean selection at the 1% level (3 ratios) includes: r1, r3, r4; and at the 5% level (5ratios) includes: r1, r3, r4, r8, r9, the same ratios as the 1% level mean selection.

Both mean and median selection of variables perform well for not-failed firms, although the median selection shows that the use of fewer variables corresponds to a reduction in discriminant power, while the mean selection produces a better discriminant power with fewer variables (slightly lower than the case with 10 variables). On the other hand, the use of more ratios leads to a better classification, and the mean selection performs worse than the median selection at the same level of significance. Unlike LDA, after winsorizing ratios, QDA (see Table 13) performs much better in identifying not-failed firms, but the identification of failed firms is poorer than the case in which raw data are used.

Table 13. Quadratic discriminant analysis (QDA), winsorized ratios

	All ratios		Median selection				Mean selection			
			5%		1%		5%		1%	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	5 ratios	5 ratios + inc.	3 ratios	3 ratios + inc.
	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)
Not-failed n=3,173	2,947 (92.30)	2,916 (91.32)	2,905 (90.98)	2,829 (88.60)	2,847 (89.16)	2,778 (87.00)	2,847 (89.16)	2,778 (87.00)	2,944 (92.20)	2,873 (89.98)
Failed n= 37	16 (43.24)	25 (67.57)	11 (29.73)	25 (67.57)	11 (29.73)	24 (64.86)	11 (29.73)	24 (64.86)	7 (18.92)	22 (59.46)
Total well classified n= 3,210	2,963 (**92.30**)	2, 941 (**91.62**)	2,916 (90.84)	2,854 (88.91)	2,858 (89.03)	2,802 (87.29)	2,858 (89.03)	2,802 (87.29)	2,951 (**91.93**)	2,895 (**90.19**)

NOTES: This table contains results obtained with the QDA model taking equal priors. Ten ratios include: wr1=TD/TA, total debt/ total assets; wr2=CA/CL, current assets/ current liabilities; wr3=EBIT/TA, earnings before interests and taxes/ total assets; wr4=NI/TA, net income/ total assets; wr5=CA/TA, current assets/ total assets; wr6=RP/TA, retained profit/ total assets; wr7=FE/TD, financial expenses/ total debt; wr8=CF/TD, cash flow/ total debt; wr9=NI/SL, net income/ sales; and wr10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: wr1, wr3, wr4, wr8, wr9; and at the 5% level includes: wr1, wr3, wr4, wr5, wr7, wr8, wr9. The mean selection at the 1% level (3 ratios) includes: r1, r3, r4; and at the 5% level (5ratios) includes: r1, r3, r4, r8, r9, the same ratios as the 1% level mean selection.

After winsorizing ratios, logDA (see Table 14) better identifies not-failed firms with a fewer number of variables, especially when the incidents variable is added. Again, the opposite is found for failed firms: as the number of variables decreases, the number of well-classified firms decreases, although adding the incidents variable leads to a better discriminant power for any selection of variables. In general, after winsorizing ratios, logDA produces more stable results across different selections of variables than the use of raw data (see Table 9). It improves the identification of not-failed firms when using a higher number of ratios and that of failed firms when using a lower number of ratios, the percentages of total well classified firms are better using 10 or 7 variables without the incidents variable (when previous results were poorer) but worse in any other case.

Table 14. Logistic discriminant analysis (LogDA), winsorized ratios

	All ratios		**Median selection**				**Mean selection**			
			5%		**1%**		**5%**		**1%**	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	5 ratios	5 ratios + inc.	3 ratios	3 ratios + inc.
	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)
Not-failed	2,285	2,716	2,277	2,735	2,715	2,826	2,715	2,826	2,747	2,861
n=3,173	(71.56)	(85.06)	(71.31)	(85.66)	(85.03)	(88.51)	(85.03)	(88.51)	(86.03)	(89.60)
Failed	26	26	24	26	15	23	15	23	15	22
n= 37	(70.27)	(70.27)	(64.86)	(70.27)	(40.54)	(62.16)	(40.54)	(62.16)	(40.54)	(59.46)
Total well classified	2,311	2, 742	2,301	2,761	2,730	2,849	2,730	2,849	2,762	2,883
n= 3,210	(71.99)	(85.42)	(71.68)	(86.01)	**(85.05)**	**(88.75)**	**(85.05)**	**(88.75)**	**(86.04)**	**(89.81)**

NOTES:This table contains results obtained with the Log DA model taking equal priors. Ten ratios include: wr1=TD/TA, total debt/ total assets; wr2=CA/CL, current assets/ current liabilities; wr3=EBIT/TA, earnings before interests and taxes/ total assets; wr4=NI/TA, net income/ total assets; wr5=CA/TA, current assets/ total assets; wr6=RP/TA, retained profit/ total assets; wr7=FE/TD, financial expenses/ total debt; wr8=CF/TD, cash flow/ total debt; wr9=NI/SL, net income/ sales; and wr10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: wr1, wr3, wr4, wr8, wr9; and at the 5% level includes: wr1, wr3, wr4, wr5, wr7, wr8, wr9. The mean selection at the 1% level (3 ratios) includes: r1, r3, r4; and at the 5% level (5ratios) includes: r1, r3, r4, r8, r9, the same ratios as the 1% level mean selection.

Using winsorized ratios, KNNDA (see Table 15) achieves an excellent discriminant power: All of the failed firms are identified, and high percentages of well-classified not-failed firms are obtained for any selection of variables. As for the not-failed firms, using only ratios, the median selection performs better at the 1% level of significance. And again, the fewer accounting variables are used, the better the discriminant power obtained. When the incidents variable is added, a good and quite stable behavior is achieved for any selection of variables. The results are similar to those using raw data, but slightly better for not-failed firms.

Table 15. Kth Nearest-neighbor discriminant analysis (KNNDA), winsorized ratios

	All ratios		**Median selection**				**Mean selection**			
			5%		**1%**		**5%**		**1%**	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	5 ratios	5 ratios + inc.	3 ratios	3 ratios + inc.
	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)	number (%)
Not-failed	3,152	3,162	3,153	3,161	3,159	3,153	3,159	3,153	3,143	3,161
n=3,173	(98.72)	(99.03)	(98.75)	(99.00)	(98.94)	(98.75)	(98.94)	(98.75)	(98.43)	(99.00)
Failed	37	37	37	37	37	37	37	37	37	37
n= 37	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)
Total well classified	3,189	3, 199	3,190	3,198	3,196	3,190	3,196	3,190	3,180	3,198
n= 3,210	(99.35)	**(99.66)**	**(99.38)**	**(99.63)**	**(99.56)**	(99.38)	**(99.56)**	(99.38)	(99.07)	**(99.63)**

NOTES: This table contains results obtained with the KNNDA model taking equal priors. Ten ratios include: wr1=TD/TA, total debt/ total assets; wr2=CA/CL, current assets/ current liabilities; wr3=EBIT/TA, earnings before interests and taxes/ total assets; wr4=NI/TA, net income/ total assets; wr5=CA/TA, current assets/ total assets; wr6=RP/TA, retained profit/ total assets; wr7=FE/TD, financial expenses/ total debt; wr8=CF/TD, cash flow/ total debt; wr9=NI/SL, net income/ sales; and wr10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: wr1, wr3, wr4, wr8, wr9; and at the 5% level includes: wr1, wr3, wr4, wr5, wr7, wr8, wr9. The mean selection at the 1% level (3 ratios) includes: r1, r3, r4; and at the 5% level (5ratios) includes: r1, r3, r4, r8, r9, the same ratios as the 1% level mean selection.

After winsorizing the ratios for the logit and the probit model (see Table 16, Panels A and B), the first relevant effect is that all of the models are significant at the 1% level. That is, we can obtain models with good discriminant ability with any selection of significant variables, applying either mean or median differences. Both models identify all not-failed firms but no failed firms when using the ratios only. When incidents is added, the results are similar, except that all models identify one failed firm.

Table 16. Panel A: Logit, winsorized ratios

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed n=3,193	3193 (100.00)	3192 (99.97)	3193 (100.00)	3192 (99.97)	3193 (100.00)	3193 (100.00)	3,193 (100.00)	3,193 (100.00)
Failed n= 37	0 (0.00)	1 (2.70)	0 (0.00)	1 (2.70)	0 (0.00)	1 (2.70)	0 (0.00)	1 (2.70)
Total well classifie n= 3,230	3,193 (98.85)	3,193 (98.85)	3,193 (98.85)	3,193 (98.85)	3,193 (98.85)	3,194 (98.89)	3,193 (98.85)	3,194 (98.89)

Panel B: Probit, winsorized ratios

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed n=3,193	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3193 (100.00)	3,193 (100.00)	3,193 (100.00)
Failed n= 37	0 (0.00)	1 (2.70)	0 (0.00)	1 (2.70)	0 (0.00)	1 (2.70)	0 (0.00)	1 (2.70)
Total well classifie n= 3,230	3,193 (98.85)	3,194 (98.89)	3,193 (98.85)	3,194 (98.89)	3,193 (98.85)	3,194 (98.89)	3,193 (98.85)	3,194 (98.89)

Panel C: Logit and probit, winsorized ratios, tests

	All ratios		Median selection				Mean selection			
			5%		1%		5%		1%	
	10 ratios	10 ratios + inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios + inc.	5 ratios	5 ratios + inc.	3 ratios	3 ratios + inc.
LOGIT										
Prob>chi2	0.0004*	0.0000*	0.0001*	0.0000*	0.0009*	0.0000*	0.0009*	0.0000*	0.0002*	0.0000*
McFadden's R2	**0.0798**	**0.1545**	0.0765	0.1529	0.0515	0.1329	0.0515	0.1329	**0.0497**	**0.1298**
McFadden's Adj R2	**0.025**	**0.095**	0.037	0.108	0.022	0.098	0.022	0.098	**0.030**	**0.105**
BIC	-25638.2	-25660.3	-25661.1	-25683.9	**-25667.1**	**-25692.0**	**-25667.1**	**-25692.0**	-25682.6	-25706.9
Area under ROC	**0.7576**	**0.8326**	**0.7518**	0.8288	0.6742	0.7733	0.6742	0.7733	**0.6908**	**0.7686**
PROBIT										
Prob>chi2	0.0003*	0.0000*	0.0000*	0.0000*	0.0010*	0.0000*0.1	0.0010*	0.0000*0.1	0.0002*	0.0000*
McFadden's R2	**0.0814**	**0.1619**	0.0785	0.1607	0.0510	420	0.0510	420	**0.0485**	**0.1377**
McFadden's Adj R2	**0.027**	**0.102**	0.039	0.116	0.021	0.107	0.021	0.107	**0.029**	**0.113**
BIC	-25638.8	-25663.3	-25661.9	-25687.1	**-25667.0**	**-25695.7**	**-25667.0**	**-25695.7**	-25682.1	-25710.1
Area under ROC	**0.7679**	**0.8389**	**0.7608**	0.8350	0.7061	0.7875	0.7061	0.7875	**0.7103**	**0.7849**

NOTES: Ten ratios include: wr1=TD/TA, total debt/ total assets; wr2=CA/CL, current assets/ current liabilities; wr3=EBIT/TA, earnings before interests and taxes/ total assets; wr4=NI/TA, net income/ total assets; wr5=CA/TA, current assets/ total assets; wr6=RP/TA, retained profit/ total assets; wr7=FE/TD, financial expenses/ total debt; wr8=CF/TD, cash flow/ total debt; wr9=NI/SL, net income/ sales; and wr10=SL/TA, sales/ total assets. The median selection at the 1% level (5 ratios) include: wr1, wr3, wr4, wr8, wr9; and at the 5% level includes: wr1, wr3, wr4, wr5, wr7, wr8, wr9. The mean selection at the 1% level (3 ratios) includes: r1, r3, r4; and at the 5% level (5ratios) includes: r1, r3, r4, r8, r9, the same ratios as the 1% level mean selection.
*Significant at the 1% level.

Panel C in Table 16 includes the results of several tests of significance:

- McFadden's R2 shows the best discriminant power for the model with 10 ratios, followed by the model with two ratios. Both are considerably improved when the incidents variable is added.
- Again, probit tends to perform better than logit, although in fewer models than when using not-winsorized ratios.
- At the same level of significance, the mean selection is better at 1%, while the median selection is better at 5%. And these results remain unchanged based on McFadden's adjusted R2.
- Using BIC, the fewer variables used, the better. Therefore, the mean selection is better for any level of significance, as, in every case, fewer variables are significant for any level with the mean selection.
- Using ROC curves, again the areas under the curves show that models with a higher number of variables perform better.

As mentioned in Section 2.2, LDA requires representative sample sizes, while logit and probit can work with non-proportional samples. Nevertheless, the common practice has been to use a small sample of not-failed firms to obtain a sample size equal to that of the available group of failed firms. In this work, we have avoided sample bias by including the whole population of both failed and not-failed firms. In Appendix B, we include the results of the DA methods, logit and probit, after weighting the failed firms to achieve sample sizes of failed firms that are approximately 50% and 100% those of not-failed firms. As expected, the non-parametric method (KNNDA) offers the same results, and the results of the semi-parametric method (LogDA) changes very little. Contrary to our expectations, LDA and QDA show smaller changes in the percentages of well-classified firms than logit and probit do.

4. CONCLUSIONS

Our work contributes on the variable selection, as one of the three core elements in the empirical study of business failure (jointly with the concept of failure, and the method used to identify or predict the business failure).

Previous empirical studies have evolved from the selection of economically sound variables to statistical selection starting with a group found significant in previous empirical studies. We agree with Altman's inductive reasoning in that it is precisely the empirical evidence that confirms which ones of those variables proposed by the economical reasoning are applicable in business failure analyses of certain populations for a given period and within a given geographical area. In fact, during the last several decades, advances have been made in empirical techniques.

As a cause or a consequence of those big guidelines in the theory on business failure, some business features identified in previous literature as frequently significant are profitability, indebtedness, and economical-financial equilibrium, followed by economic structure, margins and turnovers, whereas the most frequently significant variables proxy for economic profitability and economic-financial short-term equilibrium.

As for the variable selection using statistical tests, our empirical analysis shows that different statistical procedures generate different selections of variables, except in very specific conditions. Our sample of small firms offers a wide dispersion in most of the accounting ratios computed. Outliers increase the standard deviation and skewness, moving the mean away from the median. As a consequence, the selection using mean differences and that using median differences produce radically different groups of significant discriminant variables, when applied to raw data.

After refining the sample by winsorizing the ratios used at the 1% level to homogenize the statistics while maintaining the population size, we find that the mean selection results are closer to the median selection results, which remain unchanged. Therefore, our work shows that mean analysis is not an appropriate method for selecting discriminant variables when they exhibit a wide and skewed dispersion. On the other hand, the mean selection is a proper way to select discriminant variables when the population or the sample analyzed is normally distributed, which is not the case for most accounting variables. Although this condition is attributed to the LDA, our study shows that the lack of normality also biases other parametric methodologies, such as QDA, logit, and probit.

Next, we applied the most commonly used techniques for analyzing business failure, discriminant analysis (DA) and logit and probit binomial regressions. In addition to the well-known linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) approaches, both of which are parametric techniques, we have used a half-parametric technique, logistic discriminant analysis (log DA), and a nonparametric technique, kth nearest-neighbor discriminant analysis (KNNDA). As for the comparison of the two types of binomial regressions, we have incorporated several recent tests to compare the goodness of fit of the two types of binomial regressions with the different groups of variables.

Our empirical results show that a population with widely dispersed variables requires a different variable selection procedure according to the methodology to be used to discriminate or classify failed and not-failed firms. Parametric methods show a better discriminant power with mean selection, while nonparametric methods perform better with median selection. Furthermore, as nonparametric DA methods are superior to parametric ones in classifying both failed and not-failed firms, our results suggest that the parametric methodology is biased when variable data present wide dispersion. Supporting the well-known idea that median is more representative than mean when the variable is dispersed, our results discourage the use of the statistical mean differences test and parametric business failure assessment methodologies for SMEs.

To avoid the sample bias, this work uses all of the failed and not-failed firms in the population under study, which implies a small proportion of failed firms. Weighting the failed firms to 50% and 100% of the not-failed firms, we find that the results do not vary using the non-parametric DA (KNNDA) and are very similar using the semi-parametric DA (logDA). As the weighting of failed firms increases towards paired samples, the bias in both LDA and QDA is mitigated: LDA identifies more failed and fewer not-failed firms, and QDA produces the opposite pattern. Unexpectedly, logit and probit show a more pronounced pattern, in line with LDA.

The same empirical models were also performed after incorporating a qualitative variable, that is, the payment incidents. Our results show a general significant improvement in the discriminant power of all of the models used and more stable proportions of well-classified firms across different selections of variables. It supports the contribution of firm-specific variables from sources other than financial statements in the accuracy of business failure models, as found by the still-scarce collection of relevant works. More specifically, our evidence supports the findings of Altman *et al.* (2008) in a different geographical area, also using a wide sample of small firms, confirming that qualitative variables are even more important for small business failure models, as financial information is quite limited for a large part of SMEs.

In summary, this work contributes to previous evidence on variable selection twofold:

- First, we test two different statistical variable selection procedures, showing the radically different results they may produce and the relevant implications for the methodology choice considering the descriptive statistics of the data. Based on a wide literature review, to our knowledge, this is the first study to address this issue.
- Second, we provide evidence on the DA and binomial models that offer more stable results across different proportions of failed and not-failed firms, indicating a better potential to avoid the imbalance problem. This is the first study to compare so many methods from this viewpoint

- Third, we provide new evidence regarding the effect of the addition of qualitative information to failure models, with previous evidence for SMEs being scarce. Our results show the relevant contribution of a qualitative non-accounting variable (legal incidents) to improve the discriminatory power of all the models applied in our work, for any selection of variables and any weighting of failed firms.

Finally, we contribute to the previous literature by adding new evidence of the application of DA models, logit and probit with several selections of variables. The use of nonparametric DA models is innovative with respect to the previous literature.

REFERENCES

Albert, A. and Lesaffre, E. (1986): "Multiple group logistic discrimination", *Computers and Mathematics with Applications*, 12A(2), 209–224.

Altman, E.I. (1968): "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *Journal of Finance*, 23(4), 589–609.

Altman, E.I. (1993): *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting and Avoiding Distress and Profiting from Bankruptcy*. New York: John Wiley & Sons.

Altman, E. I.; Sabato, G. and Wilson, N. (2008): "The value of qualitative information in SME risk management", *Journal of Credit Risk, 6*(2), 95–127.

Aziz, M.A. and Dar, H.A., (2006): "Predicting corporate bankruptcy: Where we stand", *Corporate Governance*, 6, 18–33.

Balcaen, S. and Ooghe, H. (2006): "35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems", *British Accounting Review*, 38(1), 63–93.

Baum, C.F. (2006): *An introduction to modern Econometrics using Stata*. Texas: Stata Press.

Bellovary, J.L.; Giacomino, D.E. and Akers, M.D. (2007): "A review of bankruptcy prediction studies: 1930 to present, *Journal of Financial Education*, 33, 1–43.

Berkson, J. (1944): "Application of the logistic function to bio-assay", *Journal of the American Statistical Association*, 39, 357–365.

Bliss, C. I. (1934): "The method of probits", *Science*, 79, 38–39.

Collins, R.A. and Green, R.D. (1982): "Statistical methods of bankruptcy forecasting", *Journal of Economics and Business*; 32(4), 349–352.

Dambolena, I.G. and Khoury, S.J. (1980): "Ratio stability and corporate failure", *Journal of Finance*, 35(4), 1017–1026.

Daubie, M. and Meskens, N. (2002): "Business failure prediction: A review and analysis of the Literature". En C. Zopounidis (ed.): *New Trends in Banking Management*. New York: Physica-Verlag, pp. 71–86.

Eisenbeis, R. (1977): "Pitfalls in the application of discriminant analysis in business, finance and economics", *Journal of Finance*, 32, 875–900.

Fisher, R.A. (1936): "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, 7, 179–188.

García-Ayuso, M. (1995): "La necesidad de llevar a cabo un replanteamiento de la investigación en materia de análisis de la información financiera", *Análisis financiero*, 66, 36–61.

Grice, S. and Ingram, R. (2001): "Tests of the generalizability of Altman's bankruptcy prediction model", *Journal of Business Research*, 54, 53–61.

Hair, J.F.; Anderson, R.E.; Tatham, R.L. and Black, W.C. (1999): *Análisis multivariante*, Madrid: Prentice-Hall.

Jin, J.Y.; Kanagaretnam, K. and Lobo, G.J. (2011): "Ability of accounting and audit quality variables to predict bank failure during the financial crisis", *Journal of Banking and Finance*, 35(11), 2811–2819.

Jones F.L. (1987): "Current techniques in bankruptcy prediction", *Journal Accounting Literature*, 6, 131–164.

Keasey, K. and Watson, R. (1987): "Non-financial symptoms and the prediction of small company failure: a test of Argenti's hypothesis", *Journal of Business, Finance and Accounting*, 14(3), 335–354.

Keasey, K. and Watson, R. (1991): "Financial distress prediction models: A review of their usefulness", *British Journal of Management*, 2(2), 89–102.

Laitinen, T. and Kankaanpää, M. (1999): "Comparative analysis of failure prediction methods: The Finnish case". *The European Accounting Review*, 8(1), 67–92.

Martin, D. (1977): "Early warning of bank failure: A logit regression approach", *Journal of Banking & Finance*, 1(3), 249–276.

McLachlan, G. J. (2004): *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

Ohlson, J.A. (1980): "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research*, 18(1), 109–131.

Peel, M.J. and Peel, D.A. (1987): "Some further empirical evidence on predicting private company failure", *Accounting and Business Research*, 18(69), 57–66.

Peel, M.J.; Peel, D.A. and Pope, P.F. (1986): "Predicting corporate failure- Some results for the UK corporate sector", *Omega. International Journal of Management Science*, 14(1) 5–12.

Platt, H.D. and Platt, M.B. (1990): "Development of a class of stable predictive variables: The case of bankruptcy prediction", *Journal of Business, Finance & Accounting*, 17(1), 31–51.

Premachandra, I.M.; Bhabra, G.S. and Sueyoshi, T. (2009): "DEA as a toll for bankrupcy assessment: A comparative study with logistic regression technique", *European Journal of Operational Research*, 193, 412–424.

Ravi Kumar, P. and Ravi, V. (2007): "Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review", *European Journal of Operational Research*, 180(1), 1–28.

Scott, J. (1981): "The probability of bankruptcy", *Journal of Banking and Finance*, 5, 317–344.

Smith, C.A.B. (1947): “Some examples of discrimination”, *Annals of Eugenics*, 13, 272–282.

Somoza López, A. (2002): “Modelos de predicción de la insolvencia: la incorporación de otro tipo de variables”. En F. Doldán M. Rodríguez, M. (eds.): *La gestión del riesgo de crédito*. Madrid, AECA, pp. 139–173.

Sueyoshi, T. and Goto, M. (2009a): “Can R&D expenditure avoid corporate bankruptcy? Comparison between Japanese machinery and electric equipment industries using DEA-discriminant analysis”, *European Journal of Operational Research*, 199, 576–594.

Sueyoshi, T. and Goto, M. (2009b): “Methodological comparison between DEA (data envelopment analysis) and DEA-DA (discriminant analysis) from the perspective of bankruptcy assessment”, *European Journal of Operational Research*, 199, 561–575.

Sueyoshi, T. and Goto, M. (2009c): “DEA-DA for bankruptcy-based performance assessment: misclassification analysis of Japanese construction industry”, *European Journal of Operational Research*, 199, 576–594.

Sun, J.; Li, H.; Huang, Q.-H. and He, K.-Y. (2014): “Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches”, *Knowledge-Based Systems*, 57, 41–56.

Taffler, R.J. (1982): “Forecasting company failure in the UK using discriminant analysis and finance ratio data”, *Journal of the Royal Statistical Society, Series A*, 145(3), 342–358.

Tascón, M.T. and Castaño, F.J. (2012): “Variables and models for the identification and prediction of business failure: Revision of recent empirical research advances”, *Spanish Accounting Review*, 15(1), 7–58.

Zavgren, C.V. (1983): “The prediction of corporate failure: The state of the art”, *Journal of Accounting Literature*, 2(1), 1–38.

Zavgren, C.V. (1985): “Assessing the vulnerability of failure of American industrial firms: A logistic analysis”, *Journal of Banking and Finance*, 12(1), 19-45.

APPENDIX A: Significant variables SMEs

VARIABLES	Keasey, Watson (1987)	Gandía, López-Gracia, Molina (1995)	Lizarraga (1997)	Gallego, Gomez, Yañez (1997)	López, Gandía, Molina (1998)	Ferrando, Blanco (1998)	Lizarraga (1998)	Altman, Sabato (2005)	Pompe, Bilderbeek (2005)	Altman, Sabato (2007)	Somoza, Vallverdu (2007)	Gómez, Torre, Román (2008)	Arquero, Abad, Jiménez (2008)	Labatut, Pozuelo, Veres (2009)	Fantazzini, Figini (2009)	Lugovskaya (2009)	Mora, González (2009)	Pozuelo, Labatut, Veres (2010)	Manzaneque, Benegas, García (2010)	Pederzoli, Torricelli (2010)	Llano, Piñeiro, Rodríguez (2011)	Blanco, Irimia, Oliver (2012)	Authors using the ratio
Account payable/Total assets								X	X													X	3
Added value/Sales														X									1
Added value/Total assets								X	X											X			3
Amortization/Sales															X								1
Average assets/Sales											X												1
Bank debt/(Total assets-Bank debt)								X															1
Capital/Total assets																						X	1
Cash flow/Current liabilities																X				X			2
Cash flow/Sales														X									1
Cash flow/Shareholders equity									X					X								X	3
Cash flow/Total assets			X					X	X	X						X				X	X	X	8
Cash flow/Total debt									X														1
Cash flow/Total liabilities																		X					1
Cash flow/Working capital									X														1
Common earnings/Sales														X									1
Common earnings/Shareholders equity																		X					1
Common earnings/Total assets																		X					1
Common earnings/Total liabilities																		X					1
Cost of goods sold/Trading account liabilities																X							1
Current assets/(Shareholders equity+Long-term liabilities)		X			X																		2
Current assets/Current liabilities		X	X		X										X	X	X					X	7
Current assets/Long-term assets											X												1
Current assets/Sales																X							1
Current assets/Total assets			X						X	X											X		4
Current assets/Total debt																X							1
(Current assets-Current liabilities)/Shareholders equity																	X						1
(Current assets-Inventory)/Current liabilities						X											X						2
(Current assets-Inventory)/Total assets						X																	1
(Current assets-Inventory-Current liabilities)/(Operating expenses-Amortizations-depreciation (no credit))														X									1
Current liabilities/Shareholders equity																X							1
Current liabilities/Total assets																				X		X	2
Current Liabilities/Total liabilities																		X					1
Debt maturity															X								1
Depreciation/Sales															X								1
EBIT/Financial expenses			X							X			X	X									4
EBIT/Sales														X						X			2
EBIT/Total assets							X												X	X			3
EBIT/Total debt						X																	1
EBITDA/Financial expenses										X										X			2
EBITDA/Total assets										X													1
(EBT + Amortizations)/Total assets	X																						1
(EBT + Amortizations+Depreciations)/Capital											x												1
(EBT + Amortizations+Depreciations)/Financial expenses											X												1
(EBT + Amortizations+Depreciations)/Sales											X												1

VARIABLES	Keasey, Watson (1987)	Gandía, López-Gracia, Molina (1995)	Lizarraga (1997)	Gallego, Gomez, Yañez (1997)	López, Gandía, Molina (1998)	Ferrando, Blanco (1998)	Lizarraga (1998)	Altman, Sabato (2005)	Pompe, Bilderbeek (2005)	Altman, Sabato (2007)	Somoza, Vallverdu (2007)	Gómez, Torre, Román (2008)	Arquero, Abad, Jiménez (2008)	Labatut, Pozuelo, Veres (2009)	Fantazzini, Figini (2009)	Lugovskaya (2009)	Mora, González (2009)	Pozuelo, Labatut, Veres (2010)	Manzaneque, Benegas, García (2010)	Pederzoli, Torricelli (2010)	Llano, Piñeiro, Rodríguez (2011)	Blanco, Irimia, Oliver (2012)	Authors using the ratio
EBT/Sales														X		X							2
EBT/Shareholders equity				X																			1
Financial expenses/Total liabilities				X		X						X						X					4
Increase of generated funds													X										1
Increase of total assets													X										1
Innovation/Total assets															X								1
Interests/Sales				X		X																	2
Interests/Total assets															X								1
Interval without credit																					X		1
Inventory/Current assets														X									1
Inventory/Sales			X																	X			2
(Investments+Cash-Financial expenses)/Current assets									X														1
(Long term liabilities+Shareholders equity)/Current liabilities				X																			1
Long-term assets/Shareholders equity											X												1
Long-term assets/Total assets	X							X									X						3
Long-term bank debt/Bank debt								X															1
Long-term debt/Shareholders equity											X												1
Long-term debt/Total assets											X				X					X			3
Long-term liabilities/Shareholders equity																X							1
Long-term liabilities/Total assets								X										X					2
Neperian logarithm of total assets																						X	1
(Net income + Interests) / Shareholders equity	X																						1
(Net income + Interests) / Total debt	X																						1
Net income/Added value														X									1
Net Income/Current Liabilities				X																			1
Net income/Sales						X								X		X							3
Net income/Shareholders equity																X	X						2
Net income/Total assets		X	X		X											X					X		5
Net Income/Total debt																					X		1
Net revenues/Total assets															X					X			2
Operating earnings/Operating revenues																	X						1
Operating earnings/Total assets				X													X						2
Operating margin /Average operating assets												X											1
Operating revenues/Sales																X							1
Operating revenues/Total assets																X	X						2
Personnel expenses/Added value														X				X					2
Personnel expenses/Sales						X								X	X								3
Publication lag									X														1
Quick ratio (Current assets/Current liabilities)	X																						1
Real assets/Total liabilities												X											1
Receivables/Current assets																						X	1
Receivables/Inventory											X												1
Receivables/Operating revenues																		X					1
Receivables/Total assets														X									1
Reserves/Total assets											X												1

VARIABLES	Keasey, Watson (1987)	Gandía, López-Gracia, Molina (1995)	Lizarraga (1997)	Gallego, Gomez, Yañez (1997)	López, Gandía, Molina (1998)	Ferrando, Blanco (1998)	Lizarraga (1998)	Altman, Sabato (2005)	Pompe, Bilderbeek (2005)	Altman, Sabato (2007)	Somoza, Vallverdu (2007)	Gómez, Torre, Román (2008)	Arquero, Abad, Jiménez (2008)	Labatut, Pozuelo, Veres (2009)	Fantazzini, Figini (2009)	Lugovskaya (2009)	Mora, González (2009)	Pozuelo, Labatut, Veres (2010)	Manzaneque, Benegas, García (2010)	Pederzoli, Torricelli (2010)	Llano, Piñeiro, Rodríguez (2011)	Blanco, Irimia, Oliver (2012)	Authors using the ratio
Retained earnings/Total assets							X			X									X				3
ROA - Average cost of liabilities													X										1
Sales/Average number of employees											X												1
Sales/Current liabilities				X																			1
Sales/Inventory																		X					1
Sales/Shareholders equity			X																				1
Sales/Total assets							X			X				X		X				X	X		6
Sales/Total debt						X																	1
Sales/Trading account assets																X							1
Sales/Treasury											X												1
Sales/Working capital-Inventory											X												1
Shareholders equity/Long-term assets				X												X	X						3
Shareholders equity/Total assets									X				X		X					X	X	X	6
Shareholders equity/Total debt			X				X											X			X		4
Shareholders equity/Total liabilities															X								1
(Shareholders equity+long-term liabilities)/Sales														X									1
Short-term debt/Long-term debt															X								1
Short-term debt/Shareholders equity								X		X										X			3
Short-term debt/Total debt															X								1
(Short-term negociated debt+short-term debt without explicit cost-Ext. Assets)/Generated funds before taxes													X										1
Taxes/Added value									X														1
Taxes/Sales															X								1
Total assets																						X	1
Total assets/Operating revenues		X			X									X									3
Total assets/Total liabilities																	X						1
Total debt/Shareholders equity											X						X						2
Total debt/Total assets	X					X				X													3
Total debt/Working capital											X												1
Total liabilities/(Shareholders equity+total liabilities)																			X				1
Total liabilities/(Total liabilities+Shareholders equity)															X								1
Total liabilities/Current assets																						X	1
Total liabilities/Shareholders equity																X							1
Total liabilities/Total assets														X		X							2
Trade debtors/Total assets									X														1
Trade debtors/Total debt										X													1
Trading account assets/Sales			X																				1
Trading account liabilities/Total debt																						X	1
Trading account liabilities/Trading account assets																						X	1
(Treasury+Long-term financial investments)/Total assets				X																			1
(Treasury+Long-term financial investments)/Working capital				X																			1
(Treasury+Receivables)/Current liabilities											X					X							2
(Treasury+Receivables)/Total assets		X			X											X							3
Working capital/Current liabilities																				X			1
Working capital/Operating revenues														X									1
Working capital/Sales																				X			1
Working capital/Total assets				X			X									X				X			4
Number of ratios used by author	6	5	9	11	5	9	5	8	12	10	16	3	6	19	15	22	11	11	3	16	8	13	223

References Appendix A

Altman, E.I. & Sabato, G. (2005): "Effects of the new Basel capital accord on bank capital requirements for SMEs", *Journal of Financial Services Research*, 28(1-3), 15-42.

Altman, E.I. & Sabato, G. (2007): "Modeling credit risk for SMEs: evidence from the U.S. market", *Abacus*, 43(3), 332-357.

Arquero, J.L.; Abad, M.C. and Jiménez, S.M. (2008): "Procesos de fracaso empresarial en pymes, identificación y contrastación empírica", *Revista Internacional de la Pequeña y Mediana Empresa,* 1(2), 64-77.

Blanco, A.; Irimia, A. and Oliver, M.D. (2012): "The prediction of bankruptcy of small firms in the UK using logistic regression", *Análisis Financiero*, 118, 32-40.

Fantazzini, D. and Figini, S. (2009): "Random survival forests models for SME credit risk measurement", *Methodology and Computing in Applied Probability*, 11(1), 29-45.

Ferrando, M. and Blanco, F. (1998): "La previsión del fracaso empresarial en la comunidad valenciana: aplicación de los modelos discriminante y logit", *Revista Española de Financiación y Contabilidad*, 27(95), 499-540.

Gallego, A.M.; Gómez, J.C. and Yáñez, L. (1997): "Modelos de predicción de quiebras en empresas no financieras", *Actualidad Financiera*, 2(5), 3-14.

Gandía, J.L.; López Gracia, J. and Molina, R. (1995): *Estudio empírico de la solvencia empresarial en la Comunidad Valenciana*. Valencia: Instituto Valenciano de Investigaciones Económicas.

Gómez, M.A.; Torre, J.M. and Román, I. (2008): "Análisis de sensibilidad temporal en los modelos de predicción de insolvencia: una aplicación a las pymes industriales", *Revista Española de Financiación y Contabilidad*, 37(137), 85-111.

Keasey, K., and Watson, R. (1987): "Non-financial symptoms and the prediction of small company failure: a test of Argenti's hypothesis", *Journal of Business, Finance and Accounting*, 14(3), 335-354.

Labatut, G.; Pozuelo, J. and Veres, E.J. (2009): "Modelización temporal de los ratios contables en la detección del fracaso empresarial de la pyme española", *Revista Española de Financiación y Contabilidad*, 38(143), 423-448.

Lizarraga Dallo, F. (1997): "Utilidad de la información contable en el proceso de fracaso: análisis del sector industrial de la mediana empresa", *Revista Española de Financiación y Contabilidad,* 26(93), 871-915.

Lizarraga Dallo, F. (1998): "Modelos de predicción del fracaso empresarial: ¿funciona entre nuestras empresas el modelo de Altman de 1968?", *Revista de Contabilidad,* 1(1), 137-164.

Llano, P.D.; Piñeiro, C. and Rodríguez, M. (2011): "A model to forecast financial failure in non financial galician SMES". Paper presented at *XII Ibero-Italian Congress of Financial and Actuarial Mathematics*, Lisboa, 7-9 July. Available at https://core.ac.uk/download/pdf/47130940.pdf.

López, J.; Gandía, J.L. and Molina, R. (1998): "La suspensión de pagos en las pymes: una aproximación empírica", *Revista Española de Financiación y Contabilidad,* 27(94), 71-97.

Lugovskaya, L. (2009): "Predicting default of Russian SMEs on the basis of financial and non-financial variables", *Journal of Financial Services Marketing*, 14(4), 301–313.

Manzaneque, M.; Benegas, R. and Gandía, D. (2010): “Diferentes procesos de fracaso empresarial: un análisis dinámico a través de la aplicación de técnicas estadísticas clúster”, *Revista Europea de Dirección y Economía de la Empresa*, 19(3), 67-88.

Mora, M. C. and González, M. I. (2009): “Caracterización del fracaso empresarial en Colombia: un enfoque con escalado multidimesional”. Working Paper 11/09 in “*Nuevas Tendencias en Dirección de Empresas*” Series.

Pederzoli, C. and Torricelli, C. (2010): “A parsimonious default prediction model for Italian SMEs”, *Banks and Bank Systems*, 5(4), 28-32.

Pompe, P.P.M. and Bilderbeek, J. (2005): “The prediction of bankruptcy of small and medium sized industrial firms”, *Journal of Business Venturing*, 20(6), 847–868.

Pozuelo, J.; Labatut, G. and Veres, E. (2010): “Análisis descriptivo de los procesos de fracaso empresarial en microempresas mediante técnicas multivariantes”, *Revista Europea de Dirección y Economía de la Empresa,* 19 (3), 47-66.

Somoza, A. and Vallverdú, J. (2007): “Predicción de la insolvencia empresarial: comparación de la selección de los ratios contables en los modelos contable-financieros”, *Revista de Contabilidad y Dirección*, *5*, 169-203.

APPENDIX B: Business failure assessment with winsorized ratios and different weightings of failed firms

Across any selection of variables and any weighting of failed firms, the qualitative variable contributes to obtain better identification power and more stable proportions of well classified firms.

Table B.1. Linear discriminant analysis (LDA), winsorized data, weighted failed firms

Panel A. Failed firms 1591, total firms 4784

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed	2.479	2.935	2.487	2.947	2.884	2.956	2.902	2.975
n=3,193	(77.64)	(91.92)	(77.89)	(92.30)	(90.32)	(92.58)	(90.89)	(93.17)
Failed	903	946	903	989	473	946	473	903
n= 1591	(56.76)	(59.46)	(56.76)	(62.16)	(29.73)	(59.46)	(29.73)	(56.76)
Total well classified	3.382	3.881	3.390	3.936	3.357	3.902	3.375	3.878
n= 4,784	(70.69)	(81.12)	(70.86)	(82.27)	(70.17)	(81.56)	(70.55)	(81.06)

Panel B. Failed firms 3182, total firms 6375

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed	2.248	2.933	2.300	2.926	2.858	2.943	2.903	2.976
n=3,193	(70.40)	(91.86)	(72.03)	(91.64)	(89.51)	(92.17)	(90.92)	(93.20)
Failed	2.150	1.978	1.892	1.978	1.032	1.892	860	1.806
n= 3182	(67.57)	(62.16)	(59.46)	(62.16)	(32.43)	(59.46)	(27.03)	(56.76)
Total well classified	4.398	4.911	4.192	4.904	3.890	4.835	3.763	4.782
n= 6,375	(68.99)	(77.04)	(65.76)	(76.93)	(61.02)	(75.84)	(59.03)	(75.01)

NOTES: This table contains results obtained with the LDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/total assets; r2=CA/CL, current assets/current liabilities; r3=EBIT/TA, earnings before interests and taxes/total assets; r4=NI/TA, net income/total assets; r5=CA/TA, current assets/total assets; r6=RP/TA, retained profit/total assets; r7=FE/TD, financial expenses/total debt; r8=CF/TD, cash flow/total debt; r9=NI/SL, net income/sales; and r10=SL/TA, sales/total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

Using LDA, results show similar percentages of well-classified firms for different weightings of failed firms, especially when the qualitative variable is added. Using only ratios, as the weighting of failed firms increases, more failed firms and fewer not-failed firms are well-classified, except for the mean selection (fewer variables).

Table B2. Quadratic discriminant analysis (QDA), winsorized data, weighted failed firms

Panel A: Failed firms 1591, total firms 4784

	All ratios			Median selection (5% and 1%)			Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed	2.991	2.974	3.042	2.952	3.007	2.965	3.032	2.903
n=3,193	(93.67)	(93.14)	(95.27)	(92.45)	(94.17)	(92.86)	(94.96)	(90.92)
Failed	602	1.032	344	903	387	688	258	946
n= 1591	(37.84)	(64.86)	(21.62)	(56.76)	(24.32)	(43.24)	(16.22)	(59.46)
Total well classified	3.593	4.006	3.386	3.855	3.394	3.653	3.290	3.849
n= 4,784	(75.10)	(83.74)	(70.78)	(80.58)	(70.94)	(76.36)	(68.77)	(80.46)

Panel B: Failed firms 3182, total firms 6375

	All ratios			Median selection (5% and 1%)			Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed	3.111	3.030	3.063	3.006	3.035	3.005	3.034	2.904
n=3,193	(97.43)	(94.90)	(95.93)	(94.14)	(95.05)	(94.11)	(95.02)	(90.95)
Failed	688	1.634	688	1.376	774	1.032	516	1.892
n= 3182	(21.62)	(51.35)	(21.62)	(43.24)	(24.32)	(32.43)	(16.22)	(59.46)
Total well classified	3.799	4.664	3.751	4.382	3.809	4.037	3.550	4.796
n= 6,375	(59.59)	(73.16)	(58.84)	(68.74)	(59.75)	(63.33)	(55.69)	(75.23)

NOTES: This table contains results obtained with the QDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/total assets; r2=CA/CL, current assets/current liabilities; r3=EBIT/TA, earnings before interests and taxes/total assets; r4=NI/TA, net income/total assets; r5=CA/TA, current assets/total assets; r6=RP/TA, retained profit/total assets; r7=FE/TD, financial expenses/total debt; r8=CF/TD, cash flow/total debt; r9=NI/SL, net income/sales; and r10=SL/TA, sales/total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

Using QDA, as the weighting of failed firms increases, more not-failed firms are identified, in contrast to well-classified failed firms whose percentage decreases for any selection of variables.

Table B3. Logistic discriminant analysis (LogDA), winsorized data, weighted failed firms

Panel A: Failed firms 1591, total firms 4784

	All ratios			Median selection (5% and 1%)			Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed	2.226	2.869	2.260	2.876	2.792	2.916	2.794	2.942
n=3,193	(69.72)	(89.85)	(70.78)	(90.07)	(87.44)	(91.32)	(87.50)	(92.14)
Failed	1.161	1.032	946	989	602	946	559	903
n= 1591	(72.97)	(64.86)	(59.46)	(62.16)	(37.84)	(59.46)	(35.14)	(56.76)
Total well classified	3.387	3.901	3.206	3.865	3.394	3.862	3.353	3.845
n= 4,784	(70.80)	(81.54)	(67.02)	(80.79)	(70.94)	(80.73)	(70.09)	(80.37)

Panel B: Failed firms 3182, total firms 6375

	All ratios			Median selection (5% and 1%)			Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed	2.241	2.883	2.248	2.892	2.768	2.915	2.755	2.938
n=3,193	(70.18)	(90.29)	(70.40)	(90.57)	(86.69)	(91.29)	(86.28)	(92.01)
Failed	2.236	2.064	1.978	1.978	1.204	1.892	1.204	1.806
n= 3182	(70.27)	(64.86)	(62.16)	(62.16)	(37.84)	(59.46)	(37.84)	(56.76)
Total well classified	4.477	4.947	4.226	4.870	3.972	4.807	3.959	4.744
n= 6,375	(70.23)	(77.60)	(66.29)	(76.39)	(62.31)	(75.40)	(62.10)	(74.42)

NOTES: This table contains results obtained with the LogDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/total assets; r2=CA/CL, current assets/current liabilities; r3=EBIT/TA, earnings before interests and taxes/total assets; r4=NI/TA, net income/total assets; r5=CA/TA, current assets/total assets; r6=RP/TA, retained profit/total assets; r7=FE/TD, financial expenses/total debt; r8=CF/TD, cash flow/total debt; r9=NI/SL, net income/sales; and r10=SL/TA, sales/total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

Using LogDA, as the weighting of failed firms increases, the proportions of well-classified firms do not follow a homogeneous changing pattern. In general, well-classified failed firms tend to reduce and well-classified not-failed firms tend to increase.

Table B4. *k*th-nearest-neighbor discriminant analysis (KNNDA), winsorized data, weighted failed firms
Panel A: Failed firms 1591, total firms 4784

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed n=3,193	3.149 (98.62)	3.163 (99.06)	3.152 (98.72)	3.161 (99.00)	3.163 (99.06)	3.156 (98.84)	3.163 (99.06)	3.164 (99.09)
Failed n= 1591	1.591 (100.00)	1.591 (100.00)	1.591 (100.00)	1.591 (100.00)	1.591 (100.00)	1.591 (100.00)	1.591 (100.00)	1.591 (100.00)
Total well classified n= 4,784	4.740 (99.08)	4.754 (99.37)	4.743 (99.14)	4.752 (99.33)	4.754 (99.37)	4.747 (99.23)	4.754 (99.37)	4.755 (99.39)

Panel B: Failed firms 3182, total firms 6375

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed n=3,193	3.149 (98.62)	3.163 (99.06)	3.151 (98.68)	3.160 (98.97)	3.164 (99.09)	3.157 (98.87)	3.164 (99.09)	3.165 (99.12)
Failed n= 3182	3.182 (100.00)	3.182 (100.00)	3.182 (100.00)	3.182 (100.00)	3.182 (100.00)	3.182 (100.00)	3.182 (100.00)	3.182 (100.00)
Total well classified n= 6,375	6.331 (99.31)	6.345 (99.53)	6.333 (99.34)	6.342 (99.48)	6.346 (99.55)	6.339 (99.44)	6.346 (99.55)	6.347 (99.56)

NOTES: This Table contains results obtained with the KNNDA model taking equal priors. Ten ratios include: r1=TD/TA, total debt/total assets; r2=CA/CL, current assets/current liabilities; r3=EBIT/TA, earnings before interests and taxes/total assets; r4=NI/TA, net income/total assets; r5=CA/TA, current assets/total assets; r6=RP/TA, retained profit/total assets; r7=FE/TD, financial expenses/total debt; r8=CF/TD, cash flow/total debt; r9=NI/SL, net income/sales; and r10=SL/TA, sales/total assets. The median selection at the 1% level (5 ratios) include: r1, r3, r4, r8, r9; at the 5% level includes: r1, r3, r4, r5, r7, r8, r9. The mean selection at the 1% level (the same as at the 5% level) includes: r6, r8.

Using KNNDA, the three weightings of failed firms obtain the same excellent results: 100% of the failed firms and around 99% of the not-failed firms are identified.

Table B5. Logit, winsorized data, weighted failed firms 1591, total firms 4784

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed n=3,193	3035 (95.05)	2989 (93.61)	3045 (95.36)	2983 (93.42)	3053 (95.62)	2983 (93.42)	3.098 (97.02)	2.994 (93.77)
Failed n= 1,591	559 (35.14)	946 (59.46)	516 (32.43)	946 (59.46)	344 (21.62)	946 (59.46)	215 (13.51)	903 (56.76)
Total well classified n= 4,784	3.594 (75.13)	3.935 (82.25)	3.561 (74.44)	3.929 (82.13)	3.397 (71.01)	3.929 (82.13)	3.313 (69.25)	3.897 (81.46)

Using logit, when failed firms are weighted to reach approximately 50% of not-failed firms, well-classified failed firms increase, obtaining a better percentage (from 14% to 35%) as the number of variables is higher. A stable 59% is obtained when the qualitative variable is added. In exchange, the percentage of well-classified not-failed firms goes down to 95% when only ratios are used, and decreases more (93%) when the qualitative variable is added.

Table B6. Probit, winsorized data, weighted failed firms 1591, total firms 4784

	All ratios		Median selection (5% and 1%)				Mean selection(1%)	
	10 ratios number (percentage)	10 ratios +inc. number (percentage)	7 ratios number (percentage)	7 ratios + inc. number (percentage)	5 ratios number (percentage)	5 ratios +inc. number (percentage)	3 ratios number (percentage)	3 ratios +inc. number (percentage)
Not-failed	3057	2996	3064	2992	3065	2986	3.110	2.994
n=3,193	(95.74)	(93.83)	(95.96)	(93.70)	(95.99)	(93.52)	(97.40)	(93.77)
Failed	473	946	473	946	344	946	215	903
n= 1,591	(29.73)	(59.46)	(29.73)	(59.46)	(21.62)	(59.46)	(13.51)	(56.76)
Total well classified	3.530	3.942	3.537	3.938	3.409	3.932	3.325	3.897
n= 4,784	(73.79)	(82.40)	(73.93)	(82.32)	(71.26)	(82.19)	(69.50)	(81.46)

Using probit, the results are similar albeit a bit worse for the highest number of variables when using only ratios.

Table B7. Logit and probit, tests, winsorized data, weighted failed firms 1591, total firms 4784

	All ratios		Median selection (5% and 1%)				Mean selection (1%)	
	10 ratios	10 ratios +inc.	7 ratios	7 ratios + inc.	5 ratios	5 ratios +inc.	3 ratios	3 ratios +inc.
LOGIT								
Prob>chi2	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*
McFadden's R2	**0,161**	**0,284**	0,144	0,269	0,089	0,246	**0,071**	**0,252**
McFadden's Adj R2	**0,157**	**0,28**	0,141	0,266	0,087	0,244	**0,07**	**0,252**
BIC	-35335,323	-36073,537	-35255,376	-36012,621	**-34941,617**	**-35886,589**	-34849,892	-35797,44
Area under ROC	**0,7761**	**0,8438**	**0,7531**	0,83	0,7385	0,8051	**0,7255**	**0,7859**
PROBIT								
Prob>chi2	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*
McFadden's R2	**0,155**	**0,276**	0,139	0,262	0,089	0,241	**0,071**	**0,224**
McFadden's Adj R2	**0,151**	**0,272**	0,136	0,259	0,087	**0,239**	**0,07**	**0,222**
BIC	-35298,948	-36025,258	-35225,7	-35969,951	**-34937,639**	-35857,091	-34847,784	-35770,865
Area under ROC	**0,7761**	**0,8434**	**0,7532**	0,8316	0,7302	0,8059	**0,7237**	**0,7869**

After weighting the failed firms, the tests applied show much better fit than using the real proportion, except the area under ROC that shows only slightly better numbers.