THE PROBABILITY OF DEFAULT IN INTERNAL RATINGS BASED (IRB) MODELS IN BASEL II: AN APPLICATION OF THE ROUGH SETS METHODOLOGY[1]

Autoras
Reyes Samaniego Medina. rsammed@upo.es. Universidad Pablo de Olavide
María José Vázquez Cueto. pepi@us.es. Universidad de Sevilla

RESUMEN

El nuevo Acuerdo de Capital de junio de 2004 (Basilea II) da cabida e incentiva la implantación de modelos propios para la medición de los riesgos financieros en las entidades de crédito. En el trabajo que presentamos nos centramos en los modelos internos para la valoración del riesgo de crédito (IRB) y concretamente en la aproximación a uno de sus componentes: la probabilidad de impago (PD).

Los métodos tradicionales usados para la modelización del riesgo de crédito, como son el análisis discriminante y los modelos logit y probit, parten de una serie de restricciones estadísticas. La metodología *rough sets* se presenta como una alternativa a los métodos estadísticos clásicos, salvando las limitaciones de estos.

En nuestro trabajo aplicamos la metodología *rought sets* a una base de datos, compuesta por 106 empresas, solicitantes de créditos, con el objeto de obtener aquellos ratios que mejor discriminan entre empresas sanas y fallidas, así como una serie de reglas de decisión que ayudarán a detectar las operaciones potencialmente fallidas, como primer paso en la modelización de la probabilidad de impago. Por último, enfrentamos los resultados obtenidos con los alcanzados con el análisis discriminante clásico, para concluir que la metodología de los *rough sets* presenta mejores resultados de clasificación, en nuestro caso.

Palabras claves: Calificación de préstamos, Riesgo de crédito, Basilea II, *Rough Sets*

[1] We owe our thanks to the Regional Government of Andalucia (Project of Excellence P06 – SEJ – 01537) for funding.

THE PROBABILITY OF DEFAULT IN INTERNAL RATINGS BASED (IRB) MODELS IN BASEL II: AN APPLICATION OF THE ROUGH SETS METHODOLOGY

ABSTRACT.

The new Capital Accord of June 2004 (Basel II) opens the way for and encourages credit entities to implement their own models for measuring financial risks. In the paper presented, we focus on the use of internal rating based (IRB) models for the assessment of credit risk and specifically on the approach to one of their components: probability of default (PD).

In our study we apply the *rough sets* methodology to a database composed of 106 companies, applicants for credit, with the object of obtaining those ratios that discriminate best between healthy and bankrupt companies, together with a series of decision rules that will help to detect the operations potentially in default, as a first step in modelling the probability of default. Lastly, we compare the results obtained against those obtained using classic discriminant análisis. We conclude that the *rough sets* methodology presents better risk classification results.

Key words: Rating, Credit risk, Basel II, *Rough sets*, *IRB Model.*

JEL Clasification: G21, G32

I. INTRODUCTION

In 1999, the Basel Committee on Banking Supervision, in response to the changes that had been taking place in the banking world in recent years, proposed a re-structuring of the Accord of 1988 on the measurement and control of risks assumed by financial entities. As a result, in June 2004, the definitive document of the new Capital Accord, known as Basel II, was approved. This Agreement represents a novel point of departure not only in the management of risks but also in the relationships that financial entities and their supervisory bodies will have to maintain with each other.

Basel II opens the way for and encourages the implementation of financial entities' own models, known as the internal ratings-based or IRB approach, for measuring their credit risks. The purpose of these models is to calculate the unexpected loss in respect of credit exposures, with the ultimate aim of determining the regulatory capital required. The amount of this unexpected loss depends on a set of factors: probability of default (PD), loss given default (LGD), the exposure at default (EAD), and effective maturity (M). In the Agreement, the IRB method is put forward in two versions: foundation and advanced. Both have in common the need to estimate the probability of default (PD). These estimates must be based on historical data and must represent a conservative view over the medium and long term. The advanced version of IRB also requires the entity to estimate the rest of the components of the unexpected loss. With all this, Basel II re-opens interest in the various models for the prediction of company bankruptcy and for estimating the probability of default.

These models for the prediction of business failure and for estimating the PD have been the subject of innumerable studies, carried out not only by academics but also by the financial sector itself. All the theoretical effort has been focused on the modelling of the stochastic process associated with insolvency and on determining the variables that must be included in these models. Among these traditional models, we can distinguish between univariate and multivariate models. The first type studies separately the behaviour of each of the variables that should explain the insolvency. One of the classic studies, in this respect, was written by Beaver (1966), who found a number of financial ratios that could discriminate between healthy and bankrupt companies, in the period of 5 years before the default actually took place. Other notable studies were those conducted by Courtis (1978) and Altman (1993). Unlike the univariate models, the multivariate models combine the information provided by a set of variables; the pioneer study is that done by Altman in 1968, in which the author proposed a discriminant analysis combining the information provided by 25 financial ratios. There is a wide variety of studies based on discriminant analysis including, among others, Dambolena (1980) and Laitinen (1991). In

Spain, Cabedo et al. (2004) present an adaptation of the discriminant model for calculating the probability of default in companies, applied to a portfolio of hypothetical borrowers belonging to the same sector, with the object of calculating the regulatory capital according to the foundation IRB method of the Basel Accord. The importance of this technique is demonstrated in the bibliographic review carried out by Dimitras (1996); after analysing 158 articles on the prediction of business insolvency, for the period from 1932 to 1994, this author concludes that discriminant analysis is the model most frequently used in the resolution of this type of problem.

Other authors have opted for logit and probit analysis; Ohlson (1980) was the first to apply this type of technique to the prediction of company insolvency. Wilson (1997) developed the CreditPortfolio View model for McKinsey, establishing a discrete process of multiple periods. With this methodology the probability of default is obtained as logit functions of indices of macroeconomic variables that, in some way, represent the functioning of the economy (see Zmijewski, 1984). Dimitras (1996) found that the logit model has been the second most frequently used for resolving the problem of company bankruptcy.

Fernández (2005), in a attempt to combine univariate and multivariate analysis, carried out an empirical study in which a prior univariate analysis was performed with the aim of selecting those ratios with greater discriminant power, within each of the categories of ratios established, from among the 23 ratios initially considered[i]. Subsequently a logit and probit multivariate analysis was performed, from which scores were obtained for each company; these scores enabled a system of rating to be established and default probabilities to be assigned. Trucharte et al. (2002) obtain a system for *rating* borrowers, by estimating a logistic regression model utilising economic and financial information, such that, from the scores obtained, homogeneous categories are established in which the various borrowers are classified or rated, together with the probability of default that can be assigned to each category.

Other methods have been explored, in parallel with these studies, in order to overcome the restrictive hypotheses that models of statistical inference impose on the variables. These hypotheses usually do not conform to reality and distort the results obtained; for these reasons, Eisenbeis (1977), Ohlson (1980) and Zavgren (1983) question the validity of the traditional models. In particular, techniques originating from the field of artificial intelligence began to be used; programs have been produced that are capable of generating knowledge from empirical data and then using that knowledge to make inferences on new data. Within this approach we can distinguish techniques that seek knowledge by finding patterns in the data; among these are various classes of neuronal networks, and others consisting of inferring decision rules from the base data. The methodology of *rough sets* belongs to this last group of techniques. Authors like Dimitras et al. (1998) and Daubie et al. (2002)

have applied this technique to the classification of commercial loans. Other authors, such as Ahn et al. (2000), combine the *rough sets* methodology with neuronal networks in the prediction of company failure. In Spain, various studies can similarly be found whose objective is to apply *rough sets* methodology for the prognosis of company insolvency. Segovia et al. (2003) apply this technique for the prediction of insolvency in insurance companies, and Rodríguez et al. (2005) utilise it for the same purpose in a sample of small and medium – size enterprise (SMEs).

The objective of our study is to apply the *rough sets* methodology to a database, composed of 106 companies that are debtors of the same financial entity, with the object of obtaining those ratios that best discriminate between healthy and bankrupt companies; a second objective is to find a series of decision rules that will help to detect the potentially failed credit operations, as a first step in the modelling of the probability of default. Lastly, we compare the results obtained against those obtained using classic discriminant analysis, and conclude that the *rough sets* methodology presents the best risk classification results.

This article is organised as follows. The most notable aspects of the treatment of credit risk in the Basel II Accord are presented in section 2. In the section 3 we introduce the theory of *rough sets*. We continue, in section 4, with a description of the sample of companies utilised in the empirical study. The empirical application conducted is described in section 5; here we first use the *rough sets* methodology to determine the variables that may explain the default, and then we compare the results of this methodology with those obtained using classic discriminant analysis. Finally, in the section 6, we draw a series of conclusions, followed by the bibliographical references.

II. CREDIT RISK IN THE BASEL II

The treatment of financial risks has now become a strategic factor, not only for financial entities but for any organisation, regardless of its size and of the sector in which it undertakes its activity: it is a factor that can mark the future of any entity. Focusing on the financial entities, the increase in competition, the advances that have taken place in diversification, and the highly significant changes in the regulation, such as the capital requirements now being based on the different risks assumed, have together led banks and other credit entities to seek innovative ways to help them measure and manage these risks.

The main objective of Basel II is to provide an estimate of the regulatory capital requirement that is more sensitive to the financial risks; for this, it has been proposed that the banks may utilise internal methodologies that they have devised themselves. To achieve this objective, the Committee includes in the new Agreement methods or techniques that have not previously been taken into account. Thus, there are novel elements such as: internal techniques for the assessment of credit risk with different degrees of complexity, methods for covering credit risk, and consideration of new classes of risk such as operational risks and interest rate risks in investment portfolios. In addition two complementary pillars are incorporated in the Agreement: one of supervision and the other of market discipline.

The Basel II Accord establishes two methods for estimating the regulatory capital necessary to cover the credit risk. The first is the *standardized* method, which is an improvement of that utilised in the previous Agreement, and which establishes new categories of risk, grouping each type of company in one or other category. And the second, the *Internal method* or *IRB (Internal Rating Based) approach* based on internal classifications; two levels are considered within this method: the *foundation IRB method* in which the entity only estimates the probability of default (PD) in each case, with the supervisor providing the values of the rest of the variables; and the *Advanced IRB method* in which the bank is responsible for estimating each of the variables that are included in the model of credit risk management used by that bank.

The Basel Committee allows a financial entity to quantify these variables with its own model, but does not allow it to determine all the elements necessary for calculating its own capital requirements, since the risk weighting and, therefore, the requirements for capital, are established by combining the values provided by the entity and the specific formulation given by the Committee.

Therefore, to utilise the IRB method, in either of its two versions, it is necessary to be able to calculate the probability of default (PD), which is the fundamental variable for assessing the credit risk.

III. THEORETICAL FRAMEWORK OF THE ROUGH SETS METHODOLOGY

The theory of *rough sets* was proposed by Z. Pawlak in 1982 and has been confirmed as an appropriate tool for dealing with cases where there is considerable vagueness and imprecision. More specifically, the method is efficient for working with problems of multidimensional classification (Pawlak et al., 1994). The basic idea rests on the indiscernibility relation that describes elements that are indistinguishable from each other. Its principal objective is to find basic decision rules that enable new knowledge to be acquired. Its key concepts are discernibility, approximation, reducts and, lastly, decision rules. The point of departure of the method is the existence of an information/decision table where each element is characterised by a set of variables (attributes) and a decision variable that classifies the element in one of two or more categories. Indiscernibility is said to exist when two elements are characterised by the same values of all the variables, and yet the categories in which they are classified do not coincide. This is the basis of *rough sets*. In such a case, for each class of decision or category X and for each subset B of variables, two sets are constructed; these are termed, respectively, the set of lower approximation and the set of upper approximation of the decision class. The set of lower approximation of the decision class X with respect to the variables B, $\underline{B}X$, is given by the group of all the elements that, being characterised by B, belong to class X with complete certainty. The set of upper approximation of the decision class X, $\overline{B}X$, is given by the group of elements that, based on the information B that we possess, may belong to class X but we cannot be sure. The elements that are different between the two sets form the "doubtful" elements; that is, those elements that, using only the information contained in B, are not known with complete certainty to belong to the class X. When these different elements exist, i.e. when the difference is not zero, it is said that class X is a *rough set* with respect to the subset of variables B. This set can be characterised numerically by the quotient between the cardinal of the set of lower approximation and the cardinal of the set of upper approximation. This quotient is known as the "accuracy of approximation". If various decision classes exist, the sum of the cardinals of all the lower approximations divided by the total of all elements is known as the "quality of approximation of the classification, by means of the set B", and this is the percentage of elements correctly classified.

Another important aspect of this technique is the reduction of the initial table of data, eliminating the redundant information. This process is carried out through the reducts. A reduct is a minimum set of variables that conserve the same capacity for the classification of the elements as the full table of information. A reduct is thus the essential part of the knowledge and constitutes the most concise way of differentiating between the decision classes[ii].

The final stage of the analysis by *rough sets* is the creation of decision rules; that is, rules that allow us to say if a given element belongs to particular decision classes. These rules represent knowledge and are generated by combining the reducts with the values of the data analysed. A decision rule is a logical statement of the type: "IF particular conditions are met THEN the element belongs to a particular decision class". These rules allow us to classify new elements easily[iii].

Next, we put an example that illustrates what we have previously exposed. Consider the dataset in Table 1, with six objects on which we have measured three characteristics.

Table 1. Table of information.

	R1	R2	R3
E1	0	1	Yes
E2	0	0	Yes
E3	1	1	No
E4	0	0	No
E5	0	1	Yes
E6	0	0	No

This table, denominated "Table of Information", may contain redundant information in two ways, either because there are objects represented several times or because we have considered unnecessary features. First, we define the relationship "*indiscernibilty*" between elements with respect to a set of characteristics B=[R1, R2, R3]: object X is indiscernible of object Y, xRy, respect to set B if the characteristics in B take exactly the same values. This is a relationship of equivalence that induces a partition on the set of objects. Thus, the equivalence classes that are obtained in accordance with the characteristics considered are:

$[R_1] = \{\{E_3\},\{E_1,E_2,E_4,E_5,E_6\}\}$

$[R_2] = \{\{E_1,E_3, E_5\},\{E_2,E_4,E_6\}\}$

$[R_3] = \{\{E_1,E_2,E_5\},\{E_3,E_4,E_6\}\}$

$[R_1, R_2] = \{\{E_3\},\{E_1,E_5\},\{E_2,E_4,E_6\}\}$

$[R_1, R_3] = \{\{E_3\},\{E_1,E_2,E_5\},\{E_4,E_6\}\}$

$[R_2,R_3] = \{\{E_3\},\{E_1,E_5\},\{E_2\}\{E_4,E_6\}\}$

$[R_1,R_2,R_3] = \{\{E_3\},\{E_1,E_5\},\{E_2\}\{E_4,E_6\}\}$

For the three considered characteristics (B=[R1, R2, R3], we obtain four classes of equivalence; remembering that to represent every class it is enough to take one element, therefore, we can put aside E5 and E6.

Then, we look for the subset of characteristics that proportionate the same partition as [R1, R2, R3], in our case [R2, R3]. This means that the R1 feature is superfluous; it does not give more information than we already have with R2 and R3.

In general there may be several subsets of features that provide the same partition as the set of all the features in study, in this case, those which contain the fewest features and provide the same partition are called reducts. We can choose any of them because they all provide the same information as if we work with all the features of departure.

If we add a variable decision, d, to the initial table of information we have a table of decision (Table 1.1).

Table 1.1 Table of decision.

	R1	R2	R3	d
E1	0	1	Yes	0
E2	0	0	Yes	0
E3	1	1	No	1
E4	0	0	No	0
E5	0	1	Yes	0
E6	0	0	No	1

For the set W0={objects/d=0}={E1,E2,E4,E5} with regard to the set of characteristics B = {R1, R2, R3}, we define B-lower-approximation ($\underline{B}$W0) and B-upper-approximations ($\overline{B}$ W0) of W0, as $\underline{B}$W0={object x /the class of equivalence that x belongs $\subseteq$ W0} and $\overline{B}$ W0={object x /the class of equivalence that x belongs $\cap$ W0 $\neq \phi$ } respectively. In the example we obtain, $\underline{B}W_0 = \{E_1,E_2,E_5\}$ y $\overline{B}\, W_0 = \{ E_1,E_2,E_4,E_5,E_6\}$.

The boundary, defined as FBW0 = $\overline{B}$ W0 - $\underline{B}$W0=(E4, E6), is non-empty, thus, the set W0 is a rough set.

The objects in $\underline{B}$ W0 can be with certainty classified as members of W0, while the objects in $\overline{B}$ W0 can be only classified as possible members of W0. The set FBW0 consists of those objects that we cannot decisively classify into W0 on the basis of knowledge in B. In our case for E1, E2 and E5, d = 0, and for E3, d = 1 and E4 and E6 cannot be classified.

From here, we get the rules of decision that will enable us to classify (assign d=0 or d=1) new elements, that is, knowing your values for R1, R2 and R3 we can assign the value of d. In most cases metaheuristics procedures will be used to extract the rules. In our case we can do it manually and for our example, the rules are: IF R3 = Yes THEN d = 0, rule that classify E1, E2, E5; IF R2= 1 and R3 = No THEN d = 1, rule that classify E3.

We see that there has been no need to use the feature R1, and that these rules don't classify 100% of the objects: E4 and E6 are not classified. In fact, those news objects on which R2 = 0 and R3 = No, can not be classified with the established rules. This occurs because $\underline{B}$ W0 $\cup$ $\underline{B}$ W1 $\neq$ {E1,E2,E3,E4,E5,E6}.

Indeed, defining W1={objects/d=1}={E3,E6}, can be seen that $\underline{B}$W1 = {E3}, and $\overline{B}$ W1 = {E3,E4,E6}, and thus $\underline{B}$W0 $\cup$ $\underline{B}$W1={ E1,E2,E3,E5} $\neq$ {E1,E2,E3,E4,E5,E6}, in fact {E1,E2,E3,E4,E5,E6}-($\underline{B}$W0 $\cup$ $\underline{B}$ W1) = {E4, E6} which can not be classified.

In our example, the quality of the classification, defined as (cardinal $\underline{B}W0$ + cardinal $\underline{B}W1$)/Nº objects, is (3+1)/6 = 66,7 %; and the precision of the classification, given by (cardinal W0 + cardinal W1) / (cardinal W0 + cardinal W1), is (3+1) / (5+3) = 50 %.

IV. DATA AND VARIABLES UTILISED

The following approach has been adopted both in the selection of the sample and in the choice of the independent variables utilised in our empirical study. Following Altman (1968) we have paired together, under criteria of size and sector, a number of healthy and bankrupt companies, thus taking a sample with the bankrupt companies representing 50% of the total. It also appears to be of particular relevance, when selecting the sample that the data considered should be obtained for the same period of time in healthy and bankrupt companies alike. However, the companies in bankruptcy or suspension of payments tend to delay the presentation of their accounting data in the time period prior to the declaration of insolvency. To overcome this inconvenience, we have taken the accounting data of the last full year prior to the bankruptcy from the most recent data available.

i) Selection of the sample.

In the development of our model, we have employed a database provided by a Spanish savings bank that contains information on companies that requested and obtained a loan from this entity. These companies were divided into two groups: healthy and bankrupt. In particular, the sample of bankrupt companies used for the analysis only included those companies with loans from the financial entity whose unpaid debt, whether of interest or principal, amounted to a percentage of more than 10% of the full risk accepted. The date of computing was 31 December 2003.

The group of healthy companies, that is, those that did not generate situations of default in the time horizon considered, was selected by the technique of individual pairing, controlled by those characteristics that could affect the relationships between financial ratios and failure. Each company of the failed group has been matched with a healthy company of the same industry and same approximate size. In relation to the sector, the pairing was done at a level of four digits of the C.N.A.E. of 1993. The criterion adopted for pairing by size is total assets.

As a homogenising factor for all the companies, we have controlled so that the total of the customer's operations with the financial entity, or live risk, should exceed 60,120 euros and that they should all be public limited company (plc), which would facilitate access to their accounting statements.

In total, the sample comprised 106 companies, 53 failed and 53 healthy, with a very diverse spread of economic activities[iv].

ii) Selection of the independent variables of the models

The independent variables chosen for the construction of the models were selected from the financial statements, principally from the Balance Sheet and Profit and Loss Account, of the companies that comprise the sample. These accounting statements were extracted from the SABI database of the company Informa, S.A., which includes more than 95% of the companies that present their accounts in the Mercantile Register in Spain. Given that most of the companies that went bankrupt did not present their financial statements in the preceding year nor even in the two years prior to the date of default, we have taken the latest data available as corresponding to the year prior to the company bankruptcy, as already explained above. Thus, the year t-1 corresponds to that of the latest available accounts.

The accounting information derived from the sample selected was subjected to a meticulous study with the aim of detecting and resolving, where found, possible anomalies or significant incidents that could distort the final analysis. Those atypical companies with clear and insuperable anomalies in their accounts were excluded from the sample. In this respect, for example, those companies that presented profits despite being in a situation of default, were eliminated.

The selection of ratios was made by choosing a broad set of variables, 25 in total, that are potentially explanatory of company bankruptcy on the basis of the frequency and efficacy with which they have been utilised in other predictive models of company insolvency, or in the analysis of banking risks.

The variables utilised include ratios of liquidity, indebtedness, structure, rotation, generation of resources, and profitability. The specific ratios considered in the analysis are given in table 2.

Table 2. *Ratios considered in the analysis*

LIQUIDITY RATIOS	Degree to which the company's assets that can be liquidated, in the short term, are sufficient to meet the payments required for the short-term debts contracted.	R1=Current assets / current liabilities
		R2=(Quick + available assets) / current liabilities
		R3=Available assets / current liabilities
		R4=(Quick + available assets - current liabilities) / (Operating costs + Personnel costs + Variation provisions + Other operating costs)
RATIOS OF INDEBTEDNESS	Relationship between the different components of the liabilities, in the short and long term, and the own funds; and between the cost of the debt and the liabilities or the profits and funds generated.	R5=Long Term Debt / Net Worth
		R6=Net Worth / Total Liabilities
		R7=Long Term Debt / (Long Term Debt + Current Liabilities)
		R8=Financial Costs / (Long Term Debt + Current Liabilities)
		R9= Financial Costs / (Gross Profits + Provision for Amortization)
		R10=Financing costs / Gross profits
		R11=Long Term Debt / Total Liabilities
STRUCTURAL RATIOS	Proportionality between the balance sheet items of assets and liabilities, and in the composition of these items.	R12= (Current Assets - Current Liabilities) / Total Assets.
		R13=Current Assets / Total Assets.
ROTATION RATIO	Measure of the dynamism of the business activity in relation to the structure of the company.	R14=(Current Assets – Current Liabilities) / (Net Turnover + Other Income from Operations)

RATIOS OF RESOURCE GENERATION	Relationship of the self-financing capacity of the company to various accounting magnitudes.	R15=(Net Profit/Loss for period + Amortization Provision) / (Net Turnover + Other Income from Operations)
		R16=(Net Profit/Loss for period + Amortization Provision) / Current Liabilities
		R17=(Net Profit/Loss for period + Amortization Provision) / (Long Term Debt + Current Liabilities)
		R18=(Net Profit/Loss for period + Amortization Provision) / Total Liabilities
		R19=(Gross Profits + Amortization Provision) / Current Liabilities
PROFITABILITY RATIOS	Comparison of the profit obtained at various levels, with the resources invested	R20=(Operating Profit/Loss + Financial Income + Profits from financial investments + Exchange rate gains) / Total Assets.
		R21=Profit/Loss from ordinary activities / Total Liabilities.
		R22=Pre-tax Profits / Net Worth.
		R23=Pre-tax Profits / Total Liabilities.
		R24= Profit/Loss for the period / Net Worth.
		R25=Gross Profits / Total Assets

Source: Trujillo et al. (2004)

V. EMPIRICAL APPLICATION

For the empirical application of this approach, described in the present article, the values of the 25 economic – financial ratios shown in table 1 have been calculated for each of the 53 bankrupt companies, for the financial year before entering into default, and a similar procedure was adopted for each matched healthy company. This produces a table of information containing 106x25 items of data. An additional column indicative of the situation of bankruptcy or health of the company in question is included in the table. Thus we have assigned the value 0 to the bankrupt company and 1 to the matched healthy company. We thus obtain an information-decision table of 106x26 items of data.

From these data, we determine which ratio or ratios, of these 25, serve to explain the company being in default, as the first step to calculating the probability of default.

First, Napierian logarithms of the values of the ratios were taken in order to avoid problems with the normality of the variables when applying the discriminant analysis. Then, given the nature of the variables considered, we proceeded to discretize the values. This is not an essential requirement for the application of the technique, but it facilitates the interpretation of the results, and it is more consistent to identify bankrupt or healthy companies, not when the values of the variables considered coincide exactly but when these fall within the same range. For this we have utilised the codification given in table 3[v].

Table 3. *Codification ranges of the variables.*

Variables	***CODIFIED VALUE***			
	0	***1***	***2***	***3***
R1	(-inf, 0.00434058)	(0.00434058, 0.00437793)	(0.00437793, 0.00467188)	(0.00467188, +inf)
R2	(-inf, 0.00131066)	(0.00131066, 0.00215777)	(0.00215777, 0.00653424)	(0.00653424, +inf)
R3	(-inf, 3.01154e-005)	(3.01154e-005, 4.19928e-005)	(4.19928e-005, 0.000818944)	(0.000818944, +inf)
R4	(-inf, -0.00126886)	(-0.00126886, -0.000469693)	(-0.000469693, -0.000412292)	(-0.000412292, +inf)
R5	(-inf, -0.000651859)	(-0.000651859, 0.000583821)	(0.000583821, 0.00176257)	(0.00176257, +inf)
R6	(-inf, 1.38138e-005)	(1.38138e-005, 0.00072438)	(0.00072438, 0.00077608)	(0.00077608, +inf)
R7	(-inf, 1.46802e-005)	(1.46802e-005, 6.811e-005)	(6.811e-005, 0.000260608)	(0.000260608, +inf)
R8	(-inf, 4.45802e-005)	(4.45802e-005, 6.1332e-005)	(6.1332e-005, 0.000253627)	(0.000253627, +inf)
R9	(-inf, -0.00032564)	(-0.00032564, 0.00129406)	(0.00129406, 0.00412803)	(0.00412803, +inf)
R10	(-inf, -0.00032564)	(-0.00032564, 0.00170486)	(0.00170486, 0.00355543)	(0.00355543, +inf)
R11	(-inf, 0.00308829)	(0.00308829,	(0.00355161,	(0.00430769, +inf)

		0.00355161)	0.00430769)	
R12	(-inf, -0.00169722)	(-0.00169722, 3.00724e-005)	(3.00724e-005, 0.000163319)	(0.000163319, +inf)
R13	(-inf, 0.00131974)	(0.00131974, 0.00255195)	(0.00255195, 0.00265304)	(0.00265304, +inf)
R14	(-inf, -0.000937993)	(-0.000937993, 0.000271131)	(0.000271131, 0.00183654)	(0.00183654, +inf)
R15	(-inf, 9.4422e-006)	(9.44223e-006, 7.94461e-005)	(7.94461e-005, 0.000744896)	(0.000744896, +inf)
R16	(-inf, 3.03552e-005)	(3.03552e-005, 0.000232274)	(0.000232274, 0.000290301)	(0.000290301, +inf)
R17	(-inf, 6.49981e-006)	(6.49981e-006, 0.000152804)	(0.000152804, 0.000620383)	(0.000620383, +inf)
R18	(-inf, 6.4581e-006)	(6.45818e-006, 1.90195e-005)	(1.90195e-005, 0.000153841)	(0.000153841, +inf)
R19	(-inf, 5.4507e-005)	(5.45079e-005, 5.56532e-005)	(5.56532e-005, 0.000473715)	(0.000473715, +inf)
R20	(-inf, -0.000115164)	(-0.000115164, 5.14362e-005)	(5.14362e-005, 0.00121164)	(0.00121164, +inf)
R21	(-inf, -0.000148011)	(-0.000148011, 4.68013e-005)	(4.68013e-005, 0.00120903)	(0.00120903, +inf)
R22	(-inf, 1.16457e-005)	(1.16457e-005, 0.000203684)	(0.000203684, 0.00417805)	(0.00417805, +inf)
R23	(-inf, 7.0378e-007)	(7.0378e-007, 1.22549e-005)	(1.22549e-005, 0.000117764)	(0.000117764, +inf)
R24	(-inf, 1.10402e-005)	(1.10402e-005, 0.000426896)	(0.000426896, 0.00416212)	(0.00416212, +inf)
R25	(-inf, 1.2036e-005)	(1.2036e-005, 2.53061e-005)	(2.53061e-005, 0.000161758)	(0.000161758, +inf)

The next step in our study was to determine the accuracy provided by the explanatory variables, using the ROSE software. Quality of the approximation is one[vi].

Next we constructed the reducts. Since there are correlations between the explanatory variables introduced in the analysis, the number of reducts that the program gives is very high - specifically 18,241 reducts, of between 6 and 12 variables, and there are no core elements; that is, there is no variable that is essential for the classification or that is shown to be more relevant than any other. The frequency of appearance of each variable is shown in table 4.

Table 4. Frequency of appearance of each variable in the reducts.

Variable	R1	R2	R3	R4	R5	R6	R7
Frequency	23.78%	30.55%	42.46%	23.26%	26.31%	29.58%	24.89%
Variable	**R8**	**R9**	**R10**	**R11**	**R12**	**R13**	**R14**
Frequency	44.01%	32.34%	48.32%	33.71%	23.26%	38.31%	28.20%
Variable	**R15**	**R16**	**R17**	**R18**	**R19**	**R20**	**R21**
Frequency	31.96%	31.96%	24.75%	25,60%	24.33%	28.39%	30.72%
Variable	**R22**	**R23**	**R24**	**R25**			
Frequency	39.72%	37.99%	34.67%	29.52%			

From all the possible reducts we have selected three. The selection criteria are: first, that it should contain the smallest possible number of variables; second, that the variables should present a high frequency of appearance in the different reducts; and lastly, that they should be formed by the smallest number of ratios for each category considered. The ratios belonging to each of the reducts are shown in table 5.

Table 5. Selected reducts

Reducts	Variables
1	{R3,R10,R13,R17,R22,R25}
2	{R3,R10,R13,R14,R17,R20,R25}
3	{R3,R11,R13,R14,R15,R23}.

The two first reducts have been chosen because they include the ratios R3, R10 and R13, the quotients with high percentages of appearance, and which contain a ratio in each of the categories studied. The third reduct was chosen because it presents ratios from all the categories. We limited our search to those reducts formed by a maximum of seven ratios, one more than the number of categories considered.

With each of the three reducts, decision rules were generated using the lem2 procedure[vii]; these are shown in table 6. From reading this table the following points can be noted. We can observe that the number of rules varies from one reduct to another; we find that 24, 28 and 22 rules, respectively, are needed to classify correctly 100% of the observations. The first reduct is the one that requires the fewest rules to identify the bankrupt companies. We can see also that it is the reduct that has rules with the greatest power of classification. Thus, the rules 11 [(R22 = 0) => (D1 = 0)] and 12 [(R10 = 1) & (R22 = 2) & (R25 = 3) => (D1 = 1)] of this first reduct classify, respectively, 37.74% of the bankrupt companies and 49.06% of the healthy ones. No other individual rule of the second and third reducts reaches such high percentages.

These two rules tell us firstly that, if the value of the Napierian logarithm of the profitability ratio R22 (Pre-tax profits / Net Worth) is less than 1.16457e-005, the company must be classified as bankrupt, and secondly that, if the value of the Napierian logarithm of the ratio of indebtedness, R10 (Financial Costs /Gross Profits) is between (0.00032564 and 0.00170486), that of the profitability ratio R22 (Pre-tax profits / Net Worth) is between (0.000203684 and 0.00417805) and that of the ratio R25 (Gross Profits / Total Assets) is greater than 0.000161758, the company must be classified as healthy.

Table 6. Selected reducts and corresponding decision rules.

6.1.

Reduct		Rules of Classification	% correct classification
R3,R10,R13,R17,R22,R25	(D1 = 0): HEALTHY	1. (R3 = 2) & (R13 = 3) & (R17 = 1) & (R25 = 3)	9.43%
		2. (R3 = 0) & (R10 = 3)	26.42%
		3. (R10 = 2) & (R17 = 2) & (R22 = 1) & (R25 = 3)	3.77%
		4. (R17 = 0)	33.96%
		5. (R10 = 3) & (R17 = 2) & (R22 = 2)	5.66%
		6. (R3 = 2) & (R13 = 1) & (R25 = 2)	11.32%
		7. (R3 = 0) & (R17 = 1)	13.21%
		8. (R10 = 1) & (R13 = 3) & (R22 = 3)	3.77%
		9. (R3 = 0) & (R17 = 3)	5.66%
		10. (R3 = 2) & (R13 = 0) & (R25 = 2)	5.66%
		11. (R22 = 0)	37.74%,
	(D1 = 1): BANKRUPT	12. (R10 = 1) & (R22 = 2) & (R25 = 3)	49.06%,
		13. (R10 = 2) & (R17 = 2) & (R22 = 2) & (R25 = 3)	16.98%
		14. (R3 = 2) & (R17 = 3) & (R22 = 2)	9.43%
		15. (R13 = 2)	9.43%
		16. (R3 = 3) & (R17 = 3)	24.53%
		17. (R3 = 2) & (R13 = 3) & (R25 = 2)	5.66%
		18. (R25 = 1)	3.77%
		19. (R3 = 2) & (R13 = 1) & (R25 = 3)	15.09%
		20. (R3 = 0) & (R17 = 2) & (R25 = 2)	1.89%

		21. (R3 = 2) & (R17 = 2) & (R22 = 3)	3.77%
		22. (R3 = 1) & (R25 = 3)	7.55%
		23. (R3 = 2) & (R10 = 3) & (R13 = 3) & (R17 = 2) & (R22 = 1)	1.89%
		24. (R3 = 3) & (R25 = 2)	3.77%

6.2.

R3,R10,R13,R14,R17,R20,R25	(D1 = 0): HEALTHY	1. (R13 = 3) & (R14 = 1) & (R17 = 1) & (R25 = 3)	11.32%
		2. (R10 = 3) & (R25 = 2)	22.64%
		3. (R10 = 3) & (R14 = 2)	9.43%
		4. (R10 = 0)	16.98%
		5. (R3 = 2) & (R14 = 0) & (R17 = 2)	5.66%
		6. (R10 = 2) & (R14 = 3)	1.89%
		7. (R10 = 3) & (R13 = 0)	15.09%
		8. (R3 = 0) & (R13 = 3)	26.42%
		9. (R3 = 3) & (R13 = 0) & (R17 = 2)	1.89%
		10. (R20 = 0)	32.08%
		11. (R3 = 0) & (R17 = 3)	5.66%
		12. (R14 = 1) & (R20 = 1) & (R25 = 2)	9.43%
		13. (R13 = 0) & (R14 = 0)	16.98%
		14. (R10 = 2) & (R13 = 3) & (R20 = 1)	3.77%
		15. (R20 = 3)	1.89%
		16. (R10 = 3) & (R20 = 2)	1.89%
	(D1 = 1): BANKRUPT	17. (R10 = 1) & (R17 = 3) & (R20 = 2)	28.30%
		18. (R10 = 2) & (R14 = 1) & (R17 = 2) & (R20 = 2)	15.09%
		19. (R3 = 2) & (R10 = 1) & (R20 = 2)	20.01%
		20. (R3 = 2) & (R13 = 1) & (R25 = 3)	15.09%
		21. (R13 = 2)	9.43%
		22. (R10 = 1) & (R14 = 2)	32.08%
		23. (R3 = 1) & (R25 = 3)	7.55%
		24. (R3 = 0) & (R13 = 1) & (R17 = 2)	3.77%
		25. (R3 = 2) & (R13 = 3) & (R25 = 2)	5.66%
		26. (R3 = 3) & (R13 = 1)	15.09
		27. (R14 = 2) & (R17 = 3)	26.42%

		28. (R3 = 2) & (R10 = 3) & (R13 = 3) & (R14 = 1) & (R17 = 2) & (R20 = 1)	1.89%

6.3.

R3,R11,R13,R14,R15,R23	(D1 = 0): HEALTHY	1. (R3 = 0) & (R11 = 2)	26.42%,
		2. (R11 = 3)	26.42%,
		3. (R15 = 1) & (R23 = 3)	3.77%
		4. (R3 = 2) & (R13 = 1) & (R14 = 0)	3.77%
		5. (R23 = 0)	47.17%
		6. (R13 = 0) & (R23 = 1)	3.77%
		7. (R11 = 2) & (R13 = 3) & (R15 = 2) & (R23 = 2)	5.66%
		8. (R15 = 3) & (R23 = 2)	7.55%
		9. (R14 = 0) & (R15 = 3)	5.66%
		10. (R3 = 2) & (R11 = 0) & (R14 = 3)	7.55%
		11. (R13 = 1) & (R23 = 1)	1.89%
		12. (R11 = 0) & (R13 = 1) & (R14 = 1) & (R23 = 2)	1.89%
	(D1 = 1): BANKRUPT	13. (R13 = 1) & (R15 = 2) & (R23 = 3)	20.75%
		14. (R3 = 2) & (R15 = 2) & (R23 = 3)	30.19%
		15. (R3 = 2) & (R11 = 2) & (R15 = 1) & (R23 = 2)	7.55%
		16. (R11 = 0) & (R14 = 2)	30.19%
		17. (R3 = 3) & (R15 = 2)	30.19%
		18. (R11 = 1) & (R23 = 2)	11.32%
		19. (R13 = 2)	9.43%
		17. (R3 = 3) & (R15 = 2)	30.19%
		18. (R11 = 1) & (R23 = 2)	11.32%
		19. (R13 = 2)	9.43%
		20. (R3 = 1) & (R14 = 3)	1.89%
		21. (R3 = 1) & (R15 = 2)	5.66%
		22. (R3 = 2) & (R11 = 2) & (R13 = 3) & (R23 = 1)	1.89%

To check these classification results, we have performed a cross validation with ten passes; the results for each reduct are presented in table 7.

Table 7. Percentages of correct classification.

Reduct	**Correct classification: bankrupt**	**Correct classification: healthys**	**Correct classification: total**
R3,R10,R13,R17,R22,R25	83.98 %	87.90 %	86.00 %
R3,R10,R13,R14,R17,R20,R25	66.67 %	84.83 %	76.27 %
R3,R11,R13,R14,R15,R23	86.58 %	88.82 %	88.82 %

The first and third reducts are clearly more robust, in respect of percentages of correct classification, than the second. The third reduct, in addition to meeting the previously imposed requirements, is the one that presents fewest type I and II errors, these being 13.42% and 11.18%, respectively, for the validation sample.

By means of this study we have been able to confirm that the *rough sets* methodology employed leads to good results in the classification of healthy and bankrupt companies, and indicates which are the most relevant of all the variables considered. Thus, by applying these rules to new credit operations, a bank would be able to detect the possible defaults.

Comparison with the Discriminant Analysis

As we have reported, discriminant analysis was the first technique used, and is the one most frequently utilised to date, for the measurement of company insolvency. Therefore we shall compare the results obtained using the *rough sets* methodology with those that would be produced using discriminant analysis.

We shall use the same set of data to perform this comparison[viii]. First we have utilised Box's M statistic to test the equality of the matrices of variance-covariance between the two groups. From the results we can accept this hypothesis at level of significance of 5%. The discriminant function, obtained by means of the *ascending stepwise* method, with Snedecor's F criterion of entry between 0.06 and 0.09, and the statistics associated with the model are shown in table 8.

Table 8. Discriminant functions and statistics of the models

Discriminant function (discriminant canonical function with standardised coefficients)	**Wilks' lambda**	**P-Value**
Z = 0.554 R4 + 0.769 R6 – 0.337 R9	0.691	<0.0000

From reading the above table we can deduce that, according to the discriminant analysis, the liquidity and indebtedness ratios are the most relevant ratios for determining the possible insolvency of a company. Thus the liquidity ratio affects positively the probability of being classified as healthy. With respect to the ratios of indebtedness, their influence will depend on their definition. The ratio R6 (Net Worth / Total liabilities) has a positive influence, while ratio R9 (Financial Costs / (Gross Profits + provision for amortization)) has a negative influence on the probability of the company in question being considered healthy.

Table 9 gives the results of the classification with the discriminant function, applied to the original sample, and validated by means of the cross procedure.

Table 9. Percentages of correct classification.

	Discriminant function		
	Correct classification: bankrupt	**Correct classification: healthy**	**Correct classification: total**
Original sample	71.7%	83.0%	77.4%
Validation sample	71.7%	83.0%	77.4%

From reading tables 7 and 9 together, it can be deduced that the *rough sets* methodology is demonstrated to be a more useful tool than discriminant analysis, for the classification of the defaulting companies in the database considered for our empirical application. With the correct choice of reducts, only six of the twenty five variables considered in the analysis are required to classify correctly 100% of the original observations; furthermore it is found that the percentages of correct classifications are better for the samples of validation, giving type I and II errors of 13.42% and 11.18% respectively, against 28.3% and 17% that are given by the application of discriminant analysis, utilising the *forward* method with the twenty five variables.

VI. CONCLUSION

The objective of Basel II is to set a regulatory capital requirement that is more sensitive to risks in general and to credit risk in particular. For this, with reference to credit risk, it has been proposed that banks may utilise internal measurement methodologies that they have devised themselves.

In our article we present an alternative methodology to classic discriminant analysis for determining the variable or variables that serve to explain the failure of a company to meet its debt repayments, as the first step in determining the PD.

Basing our arguments on a sample of healthy and bankrupt companies, and on a set of 25 financial ratios that are potential explanatory factors of the defaults occurring in the sample, we have concluded that the *rough sets* methodology can be a valid alternative to discriminant analysis when there is a need to classify objects in two classes of decision. Not only are acceptable percentages of correct classification obtained using *rough sets* but also there is no need to require any kind of prior statistical behaviour of the variables that are involved in the classification, unlike discriminant analysis, which requires normality of the distributions and equality of the matrices of variance-covariance. Furthermore, the variables are included as they are presented, with no need for any prior transformation. Among the more significant advantages of this methodology are that it eliminates redundant information and that it expresses the dependencies between the variables considered and the result of the classification, by means of decision rules whose language is closer to that normally utilised by the experts.

Notes

[ii]These categories were liquidity, leverage, activity, debt cover, and productivity.

[ii] The reducts are obtained on the basis of the equivalence classes that define the indiscernibility relation, on the set of observations.

[iii] For more detail on the formal mathematical aspects of the methodology, Komorowski et al. (1999) may be consulted.

[iv] We have excluded from the analysis property development and property sales companies, since these have characteristics that are very peculiar and different from the rest of companies, and because, in the assessment of the application for a loan made by this type of company, the decisive factor for granting the loan is the viability of the specific project for which the loan is sought, and this information is not reflected in the corresponding accounting statements.

[v] We have discretized the variables grouping them in four ranges based on the number of observations belonging to each range. For this we have utilised the ROSE software, provided by the Institute of Computing Science of Poznan University of Technology, and we thank the Institute for making this available to us.

[vi] The quality of the approximation is expressed by the ratio between the number of companies classified correctly and the total number of companies that comprise the sample.

[vii] Chan et al., 1994

[viii]The Napierian logarithms of the data have been used, and their values have been typified.

References:

Altman, E.I. (1968). Financial Ratios, Discriminate Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance (September) 589-609.*

Altman, E.I. (1993). Corporate Financial Distress and Bankruptcy. John Wiley. New York.

Ahn, B.S., Cho, S.S., Kim, C.Y. (2000).The Integrated Methodology of Rough Set Theory and Artificial Neural Network for Business Failure Prediction. *Expert Systems with Applications, 18.*

Basle Committee On Banking Supervision (2001).The Internal Ratings-Based Approach. *Consultive Document. Supporting Document to the New Basle Capital Accord, January.*

Basel Committee On Banking Supervision (1999). Credit Risk Modelling: Current Practices and Applications. April.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Empirical Research in Accounting: Selected Studies, (Supplement of The Accounting Review).*

Cabedo,J., Reverte, J.A., Tirado, J.M. (2004). Riesgo de Crédito y Recursos Propios Mínimos en Entidades Financieras. *Revista Europea de Dirección y Economía de la Empresa, volumen 13, nº. 2.*

Chan. C., Grzymala-Busse, J. (1994).On the two local inductive algorithms: PRISM and LEM2. *Foundations of Computing and Decision Sciences*, 19, 185–204.

Comité De Supervisión Bancaria De Basilea (2004). Convergencia Internacional de Medidas y Normas de Capital. June.

Courtis, J.K. (1978). Modelling Financial Ratios: Categories Framework. *Journal of Business, Finance and Accounting*, 5 (4).

Dambolena, I.G., Khoury, S.J. (1980). Ratio Stability and Corporate Failure. *The Journal of Finance*, 35.

Daubie, M.; Leveck, P.; Meskens, N. (2002). A Comparison of the Rough Sets and Recursive Partitioning Induction Approaches: An Application to Commercial Loans. *International Transactions in Operational Research* 9, 681-694.

Dimitras, A.I., Zanakis, S.H., Zopounidis, C. (1996). A Survey of Business Failures with an Emphasis on Prediction Methods and Industrial Applications. *European Journal of Operational Research*, 90.

Dimitras, A.I., Slowinski, R., Susmaga, R, Zopounidis, C. (1998). Business Failure Prediction Using Rough Sets. *European Journal of Operational Research*, 114.

Eisenbeis, R.A. (1977). Pitfalls in the Application of Discriminant Analysis in Business and Economics. *Journal of Finance*, 32, 875-900.

Fernandez, J.E. (2005). Corporate Credit Risk Modeling: Quantitative Rating System and Probability of Default Estimation. Http://www.defaultrisk.com, april.

Komorowski, Z. (1999). Rough Sets: A Tutoriain Rough-Fuzzy Hybridization - A New Trend in Decision Makin. S.K. Pal *and A. Skowron, Eds.*, 3-98, Springer-Verlag Singapore Pte Ltd.

Laitinen, E.K. (1991). Financial Ratios and Different Failure Processes. *Journal of Business Finance and Accounting,* 18/5.

Ohlson, J.A. (1980). *Financial Ratios and the Probabilistic Prediction of Bankruptcy*. Journal of Accounting Research (Spring) 109-131.

Pawlak Z. (1982). Rough Sets. Int. J. *Computer and Information Sci.*, 11,341-356

Pawlak,Z.; Slowinski,R.(1994). Decision Analysis Using Rough Sets. *International Transactions in Operational Research* 1, 107-114.

Rodriguez, M., Díaz, F. (2005). La Teoría de los Rough Sets y la Predicción del Fracaso Empresarial. Diseño de un Modelo para Pymes. *XIII Congreso AECA*.

Segovia, M.J., Gil, J.A., Vilar, L., Heras, A.J. (2003). La Metodología Rough Set frente al Análisis Discriminante en la Predicción de Insolvencia en Empresas Aseguradoras. *Anales del Instituto de Actuarios Españoles*, 9.

Trujillo, A., Martin, J.L. (2004). El Rating y la Fijación de Precios en Préstamos Comerciales: Aplicación mediante un Modelo Logit. *Revista Europea de Dirección y Economía de la Empresa*, 13. Nº. 2.

Trucharte, C., Marcelo, A. (2002). Un Sistema de Clasificación (Rating) de Acreditados (1). *Estabilidad Financiera*, 2.

Wilson, T. (1997). Portfolio Credit Risk. *Risk Magazine*, Sept., 111-117.

Zavgren, C.V.(1983). The Prediction of Corporate Failure: the State of the Art. *Journal of Financial Literature* 2, 1-37.

Zmijewski, M.E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Studies on Current Econometric Issues in Accounting Research.*