Instituto Superior de Ciências do Trabalho e da Empresa

Departamento de Contabilidade e Finanças

CORPORATE CREDIT RISK MODELING

João Eduardo Dias Fernandes

Tese submetida como requisito parcial para obtenção do grau de

Doutor em Gestão

Especialidade Finanças

Orientador:

Prof. Doutor Miguel Almeida Ferreira

[Dezembro, 2006]

Instituto Superior de Ciências do Trabalho e da Empresa

Departamento de Contabilidade e Finanças

CORPORATE CREDIT RISK MODELING

João Eduardo Dias Fernandes

Tese submetida como requisito parcial para obtenção do grau de

Doutor em Gestão

Especialidade Finanças

Orientador:

Prof. Doutor Miguel Almeida Ferreira

[Dezembro, 2006]

CORPORATE CREDIT RISK MODELING

João Eduardo D. Fernandes

Agradecimentos

Gostaria de agradecer em primeiro lugar ao Banco BPI, SA e, em particular ao Dr. Rui Martins do Santos por possibilitarem a utilização de dados e recursos internos para este trabalho de investigação.
Um agradecimento especial ao Prof. Miguel Almeida Ferreira pelo apoio e orientação. Gostaria igualmente de agradecer a todas as pessoas que contactei ao longo destes últimos 4 anos pelo apoio e comentários da maior relevância. Em particular agradeço aos meus colegas Drs. Jorge Barros Luís, Luís Ribeiro Chorão, Carla Cavaco Martins, Dária Adriano Marques, Isabel Moutinho e Yuneza Abdul Latif, e à Prof. Cristina Neto de Carvalho da Universidade Católica Portuguesa.
Pelo encorajamento e paciência agradeço e dedico este trabalho aos meus pais e irmão e à minha Carla.
Por último gostaria de deixar uma palavra de agradecimento à FCT – Fundação para a Ciência e Tecnologia pelo apoio financeiro prestado a este projecto.

Resumo

A modelização do risco de crédito de empréstimos a empresas sem emissões cotadas em mercados financeiros é limitada, apesar do peso elevado deste segmento nas carteiras de crédito dos bancos. O objectivo deste estudo é contribuir para este ramo de literatura ao aplicar técnicas de medição dos dois principais parâmetros de risco de crédito a uma amostra aleatória extraída da base de dados de um banco europeu. A dissertação é composta por dois capítulos, o primeiro trata a modelização da probabilidade de incumprimento (PD), e o segundo a modelização da perda em caso de incumprimento (LGD). O primeiro capítulo começa por apresentar e comparar alternativas para a medição do *credit score* dos clientes, incluindo modelo de equações sectoriais múltiplas e outro com amostra ponderada. Em seguida é abordada a problemática de agrupar scores individuais em classes de risco com PDs associadas. Para tal, duas alternativas são propostas, a primeira usa técnicas de *clustering*, enquanto que a segunda baseia-se no mapeamento entre as classificações internas e uma escala de referência externa. No final do primeiro capítulo, e usando as estimativas de PD anteriormente calculadas, determinam-se os requisitos de capital regulamentar à luz do novo acordo de capital de Basileia, em contraste com os requisitos previstos no acordo actual. No segundo capítulo comparam-se duas alternativas para a modelização do LGD. Os modelos são estimados sob uma amostra aleatória de 7 anos, considerando-se como variáveis explicativas características dos empréstimos, garantias e clientes. Ambas as alternativas têm em consideração o facto da variável dependente ser uma fracção e de ter uma distribuição não normal. A primeira alternativa é baseada na transformação Beta da variável dependente, enquanto que a segunda é baseada em *Generalized Linear Models*.

Palavras-Chave: Risco de Crédito, Probabilidade de Incumprimento, Perda em Caso de Incumprimento, Basileia II
Classificação JEL: C13, G21

Abstract

Corporate credit risk modeling for privately-held firms is limited, although these firms represent a large fraction of the corporate sector worldwide. This study is an empirical application of credit scoring and rating techniques to a unique dataset on private firms bank loans of a European bank. It is divided in two chapters. The first chapter is concerned with modeling the probability of default. Several alternative scoring methodologies are presented, validated and compared. These methodologies include a multiple industry model, and a weighted sample model. Furthermore, two distinct strategies for grouping the individual scores into rating classes with PDs are developed, the first uses cluster algorithms and the second maps internal ratings to an external rating scale. Finally, the regulatory capital requirements under the New Basel Capital Accord are calculated for a simulated portfolio, and compared to the capital requirements under the current regulation. On the second chapter, we model long-term Loss-Given-Default on loan, guarantee and customer characteristics using a random, 7-year sample. Two alternative modeling strategies are tested, taking in consideration the highly non-normal shape of the recovery rate distribution, and a fractional dependent variable. The first strategy is based on Beta transformation of the dependent variable, while the second is based on Generalizes Linear Models. The methodology can be used for long-term LGD prediction of a corporate bank loan portfolio and to comply with the New Basel Capital Accord Advanced Internal Ratings Based approach requirements.

KEYWORDS: Credit Risk, Probability of Default, Loss-Given-Default, Basel II
JEL CLASSIFICATION: C13, G21

Contents

Agradecimentos ..... **i**

Resumo ..... **ii**

Abstract ..... **iii**

Contents ..... **iv**

List of Tables ..... **vii**

List of Figures ..... **ix**

Chapter I - Quantitative Rating System And Probability Of Default Estimation

1 Introduction **1**

2 Data Description **5**

3 Financial Ratios and Bivariate Analysis **9**

4 Scoring Model and Validation **14**

4.1 Model Validation 14

4.1.1 Efficiency 15

4.1.2 Statistical Significance 17

4.1.3 Economic Intuition 19

4.2 Model A – Multiple Industry Equations Model 20

4.3 Model B – Standard Model 22

4.4 Model C – Weighted Sample Model 23

4.5 Analysis of the Results 26

5 Applications **33**

5.1 Quantitative Rating System and Probability of Default Estimation 33

5.1.1 Cluster Methodology 34

5.1.2 Historical / Mapping Methodology 38

5.1.3 Rating Matrixes and Stability 41

5.2 Regulatory Capital Requirements 44

6 Conclusion **46**

Bibliography **48**
Appendix 1 – Description of Financial Ratios and Accuracy Ratios **53**
Appendix 2 – Estimating and Comparing the Area Under the ROC curves **54**
Appendix 3 – Binomial Logistic Regression Estimation and Diagnostics **56**
Appendix 4 – Estimation Results **63**
Appendix 5 – Kolmogorov-Smirnov and Error Type Analysis **67**
Appendix 6 – K-Means Clustering **71**
Appendix 7 – IRB RWA and Capital Requirements for Corporate Exposures **72**
Appendix 8 - IRB Capital Requirements Figures **74**

Chapter II - Loss Given Default Estimation
1 Introduction **78**
2 Sample Description **81**
3 LGD Definition **83**
4 Recovery Drivers **87**
4.1 Guarantee Characteristics 88
4.1.1 Guarantee Type 88
4.2 Loan Characteristics 89
4.2.1 Loan Value 89
4.2.2 Loan Maturity 90
4.2.3 Default-to-Loan Value and Seasoning-to-Loan Maturity Ratios 91
4.2.4 Interest Rate 93
4.3 Customer Characteristics 94
4.3.1 Firm Industry 94
4.3.2 Age of Relationship 95
4.3.3 Firm Age 95
4.3.4 Geographic Location 96
5 LGD Modeling **98**
5.1 Beta Transformation Methodology 99
5.2 GLM Methodology 103
6 Conclusion **107**

Bibliography **109**
Appendix 1 – Kaplan Meier Survival Analysis **111**
Appendix 2 – Industry Groups by Economic Activity **112**
Appendix 3 – Regression Fit Analysis **113**

List of Tables

Chapter I

Table 1 – Estimated Model Variables and Parameters, Model (A) 21
Table 2 – Estimated Model Variables and Parameters, Model (B) 22
Table 3 – Estimated Model Variables and Parameters, Model (C) 25
Table 4 – AUROC, AR and KS Statistics..... 30
Table 5 – Testing the Differences between AUROC's..... 30
Table 6 – Calinski-Harabasz CH(k) index for k = 2 up to k = 20..... 35
Table 7 – Annual Global Issuer-Weighted Default Rate Descriptive Statistics, 1920-2003..... 38
Table 8 – Model A 1 Year Transition Matrix (Cluster Method) 42
Table 9 – Model B 1 Year Transition Matrix (Cluster Method)..... 42
Table 10 – Model C 1 Year Transition Matrix (Cluster Method)..... 42
Table 11 – Model A 1 Year Transition Matrix (Historical Method) 42
Table 12 – Model B 1 Year Transition Matrix (Historical Method) 43
Table 13 – Model C 1 Year Transition Matrix (Historical Method) 43
Table 14 – Average RWA and Total Capital Requirements..... 45

Chapter II

Table 15 – Sample Distribution by Original Loan Maturity..... 81
Table 16 – Cumulative Recovery Rate Summary Statistics 86
Table 17 – List of Explanatory Variables 87
Table 18 – Guarantee Weight and Recovery Rate by Guarantee Type 89
Table 19 – Loan Value Summary Statistics, in EUR..... 90
Table 20 – Loan Frequency and Recovery Rate by Loan Maturity..... 91
Table 21 – Default-to-Loan Value and Seasoning-to-Maturity Ratios Summary Statistics 93

Table 22 – Loan Frequency and Recovery Rate by Interest Rate....93
Table 23 – Loan Frequency and Recovery Rate by Industry Group....94
Table 24 – Age of Relationship Summary Statistics, in Years....95
Table 25 – Loan Frequency and Recovery Rate by Age of Relationship....95
Table 26 – Firm Age Summary Statistics, in Years....96
Table 27 – Loan Frequency and Recovery Rate by Firm Age....96
Table 28 – Loan Frequency and Recovery Rate by Geographic Location....97
Table 29 – Beta Fit Descriptive Statistics....100
Table 30 – Maximum-Likelihood estimates of long-term cumulative recovery rates, Beta Methodology....101
Table 31 – Maximum-Likelihood estimates of long-term cumulative recovery rates, GLM Methodology....105
Table 32 – Industry Groups by Economic Activity Classification....112

List of Figures

Chapter I

Figure 1 – Economy-Wide vs. Main Sample Industry Distribution 6
Figure 2 – Sample Industry Distribution 7
Figure 3 – Accounting Statement Yearly Distribution 7
Figure 4 – Size (Turnover) Distribution, Millions of EUR 8
Figure 5 – Average Default Frequency by *Liquidity / Current Liabilities* Ratio Decile 9
Figure 6 – Average Default Frequency by *Current Ratio* Decile 10
Figure 7 – Average Default Frequency by *Liquidity / Assets* Ratio Decile 10
Figure 8 – Average Default Frequency by *Debt Service Coverage* Ratio Decile 11
Figure 9 – Average Default Frequency by *Interest Costs / Sales* Ratio Decile 11
Figure 10 – Average Default Frequency by *Productivity Ratio* Decile 12
Figure 11 – Average Default Frequency by *Current Earnings and Depreciation / Turnover* Ratio Decile 13
Figure 12 – Model A: Hosmer-Lemeshow Test 21
Figure 13 – Model B: Hosmer-Lemeshow Test 23
Figure 14 – Weighted vs. Unweighted Score 24
Figure 15 – Model C: Hosmer-Lemeshow Test 25
Figure 16 – Smoothed Lowess and Fractional Polynomial Adjustment for the *Interest Costs / Sales* Ratio 27
Figure 17 – Receiver Operating Characteristics Curves 28
Figure 18 – Cumulative Accuracy Profiles Curves 29
Figure 19 – Default Frequency by Rating Class (Cluster Method) 36
Figure 20 – Number of Observations Distribution by Rating Class (Cluster Method) 37
Figure 21 – Default Frequency by Rating Class (Historical Method) 39
Figure 22 – Number of Observations Distribution by Rating Class (Historical Method) 40
Figure 23 – Model A: Kolmogorov-Smirnov Analysis 67

Figure 24 – Model A: Types I & II Errors 68
Figure 25 – Model B: Kolmogorov-Smirnov Analysis 68
Figure 26 – Model B: Types I & II Errors 69
Figure 27 – Model C: Kolmogorov-Smirnov Analysis 69
Figure 28 – Model C: Types I & II Errors 70
Figure 29 – Model A - IRB Capital Requirements (Cluster Method) 74
Figure 30 – Model B - IRB Capital Requirements (Cluster Method) 74
Figure 31 – Model C - IRB Capital Requirements (Cluster Method) 75
Figure 32 – Model A - IRB Capital Requirements (Historical Method) 75
Figure 33 – Model B - IRB Capital Requirements (Historical Method) 76
Figure 34 – Model C - IRB Capital Requirements (Historical Method) 76

Chapter II

Figure 35 – Sample Distribution by Year of Loan and Year of Default 82
Figure 36 – Sample Distribution by Original Loan Size, thousand EUR 82
Figure 37 – Empirical Recovery Rate Distribution 84
Figure 38 – Cumulative Recovery Rate Growth by Year 85
Figure 39 – Average Cumulative Recovery Rate Growth 85
Figure 40 – Average Recovery Rate by Loan Value Percentile 90
Figure 41 – Average Recovery Rate by Default-to-Loan Value Ratio Percentile 92
Figure 42 – Average Recovery Rate by Seasoning-to-Maturity Ratio Percentile 92
Figure 43 – Beta Distribution Fit to Empirical Recovery Rate Distribution 100
Figure 44 – Pearson Residuals Plot by Estimated Recovery Rate, Beta Methodology 102
Figure 45 – Cloglog, Logit and Loglog Link Functions 104
Figure 46 – Pearson Residuals Plot by Estimated Recovery Rate, GLM Methodology 106

Chapter I

Quantitative Rating System And Probability Of Default Estimation

1 Introduction

The credit risk modeling literature has grown extensively since the seminal work by Altman (1968) and Merton (1974). Several factors contribute for an increased interest of market practitioners for a correct assessment of the credit risk of their portfolios: the European monetary union and the liberalization of the European capital markets combined with the adoption of a common currency, increased liquidity, and competition in the corporate bond market. Credit risk has thus become a key determinant of different prices in the European government bond markets. At a worldwide level, historically low nominal interest rates have made the investors seek the high yield bond market, forcing them to accept more credit risk. Furthermore, the announced revision of the Basel capital accord will set a new framework for banks to calculate regulatory capital[1]. As it is already the case for market risks, banks will be allowed to use internal credit risk models to determine their capital requirements. Finally, the surge in the credit derivatives market has also increased the demand for more sophisticated models.

There are three main approaches to credit risk modeling. For firms with traded equity and/or debt, Structural models or Reduced-Form models can be used. Structural Models are based on the work of Black and Scholes (1973) and Merton (1974). Under this approach, a credit facility is regarded as a contingent claim on the value of the firm's assets, and is valued according to option pricing theory. A diffusion process is assumed for the market value of the firm's assets and default is set to occur whenever the estimated value of the firm hits a pre-specified default barrier. Black and Cox (1976) and Longstaff and Schwartz (1993) have extended this framework relaxing assumptions on default barriers and interest rates.

For the second and more recent approach, the Reduced-Form or Intensity models, there is no attempt to model the market value of the firm. Time of default is

[1] For more information see Basel Committee on Banking Supervision (2003).

modeled directly as the time of the first jump of a Poisson process with random intensity. These models were first developed by Jarrow and Turnbull (1995) and Duffie and Singleton (1997).

A third approach, for privately held firms with no market data available, accounting-based credit scoring models are the most common alternative. Since most of the credit portfolios of commercial banks consist of loans to borrowers that have no traded securities, these will be the type of models considered in this research[2]. Although credit scoring has well known disadvantages, it remains as the most effective and widely used methodology for the evaluation of privately-held firms' risk profiles[3].

The corporate credit scoring literature as grown extensively since Beaver (1966) and Altman (1968), who proposed the use of Linear Discriminant Analysis (LDA) to predict firm bankruptcy. On the last decades, discrete dependent variable econometric models, namely logit or probit models, have been the most popular tools for credit scoring. As Barniv and McDonald (1999) report, 178 articles in accounting and finance journals between 1989 and 1996 used the logit model. Ohlson (1980) and Platt and Platt (1990) present some early interesting studies using the logit model. More recently, Laitinen (1999) used automatic selection procedures to select the set of variables to be used in logistic and linear models which then are thoroughly tested out-of-sample.

The most popular commercial application using logistic approach for default estimation is the Moody's KMV RiskCalc Suite of models developed for several countries[4]. Murphy et al. (2002) presents the RiskCalc model for Portuguese private firms. In recent years, alternative approaches using non-parametric methods have been developed. These include classification trees, neural networks, fuzzy algorithms and k-nearest neighbor. Although some studies report better results for the non-parametric methods, such as in Galindo and Tamayo (2000) and Caiazza (2004), we will only consider logit/probit models since the estimated parameters are more

[2] According to the Portuguese securities market commission (CMVM), at 31 December 2004 only 82 firms had listed equity or debt (CMVM 2005).
[3] See, for example, Allen (2002).
[4] See Dwyer et al. (2004).

intuitive, easily interpretable and the risk of over-fitting to the sample is lower. Altman, Marco and Varetto (1994) and Yang et al. (1999) present some evidence, using several types of neural network models, that these do not yield superior results than the classical models. Another potential relevant extension to traditional credit modeling is the inference on the often neglected rejected data. Boyes et al. (1989) and Jacobson and Roszbach (2003) have used bivariate probit models with sequential events to model a lender' decision problem. In the first equation, the decision to grant the loan or not is modeled and, in the second equation, conditional on the loan having been provided, the borrowers' ability to pay it off or not. This is an attempt to overcome a potential bias that affects most credit scoring models: by considering only the behavior of accepted loans, and ignoring the rejected applications, a sample selection bias may occur. Kraft et al. (2004) derive lower and upper bounds for criteria used to evaluate rating systems assuming that the bank storages only data of the accepted credit applicants. Despite the findings in these studies, the empirical evidence on the potential benefits of considering rejected data is not clear, as shown in Crook and Banasik (2004).

The first main objective of this research is to develop an empirical application of credit risk modeling for privately-held corporate firms. This is achieved through a simple but powerful quantitative model built on real data randomly drawn from the database of one of the major Portuguese commercial banks. The output of this model will then be used to classify firms into rating classes, and to assign a probability of default for each one of these classes. Although a purely quantitative rating system is not fully compliant with the New Basel Capital Accord (NBCA), the methodology applied could be regarded as a building block for a fully compliant system[5].

The remainder of this paper is organized as follows. Section 2 describes the data and explains how it is extracted from the bank's database. Section 3 presents the

[5] For example, compliant rating systems must have two distinct dimensions, one that reflects the risk of borrower default and another reflecting the risk specific to each transaction (Basel Committee on Banking Supervision 2003, par. 358). The system developed in this study only addresses the first dimension. Another important drawback of the system presented is the absence of human judgment. Results from the credit scoring models should be complemented with human oversight in order to account for the array of relevant variables that are not quantifiable or not included in the model (Basel Committee on Banking Supervision 2003, par. 379).

variables considered and their bivariate relationship with the default event. These variables consist of financial ratios that measure Profitability, Liquidity, Leverage, Activity, Debt Coverage and Productivity of the firm. Factors that exhibit a weak or unintuitive relationship with the default frequency will be eliminated and factors with higher predictive power for the whole sample will be selected. Section 4 combines the most powerful factors selected on the previous stage in a multivariate model that provides a score for each firm. Two alternatives to a simple regression will be tested. First, a multiple equation model is presented that allows for alternative specifications across industries. Second, a weighted model is developed that balances the proportion of default to non-default observations on the dataset, which could be helpful to improve the discriminatory power of the scoring model, and to better aggregate individual firms into rating classes. Results for both alternatives are compared and thoroughly validated. All considered models are screened for statistical significance, economic intuition, and efficiency (defined as a parsimonious specification with high discriminatory power). In Section 5 several applications of the scoring model are discussed. First, two alternative rating systems are developed, using the credit scores estimates from the previous section. One alternative consists on grouping individual scores into clusters, while the other consists on indirectly deriving rating classes through a mapping procedure between the resulting default frequencies and an external benchmark. Next, the capital requirements for an average portfolio under both the NBCA and the current capital accord are derived and compared. Section 6 concludes.

2 Data Description

A random sample of 11,000 annual, end-of-year corporate financial statements is extracted from the financial institution's database. These yearly statements belong to 4,567 unique firms, from 1996 to 2000, of which 475 have had at least one defaulted loan in a given year. The default definition considered is compliant with the NBCA proposed definition, it classifies a loan as default if the client misses a principal and/or interest payment for more than 90 days.

Furthermore, a random sample of 301 observations for the year 2003 is extracted in order to perform out-of-sample testing. About half of the firms in this testing sample are included in the main sample, while the other half corresponds to new firms. In addition, the out-of-sample data contains 13 defaults, which results in a similar default ratio to that of the main sample (about 5%). Finally, the industry distribution is similar to the one in the main sample.

Firms belonging to the financial or real-estate industries are excluded, due to the specificity of their financial statements. Furthermore, firms owned by public institutions are also excluded, due to their non-profit nature.

The only criteria employed when selecting the main dataset is to obtain the best possible approximation to the industry distribution of the Portuguese economy. The objective is to produce a sample that could be, as best as possible, representative of the whole economy, and not of the bank's portfolio. If this is indeed the case, then the results of this study can be related to a typical, average credit institution operating in Portugal.

Figure 1 shows the industry distribution for both the Portuguese economy and for our dataset[6]. The distributions are similar, although the model and test samples have higher concentration on industry *D – Manufacturing*, and lower on *H – Hotels & Restaurants* and *MNO – Education, Health & Other Social Services Activities*.

[6] Source: INE 2003.

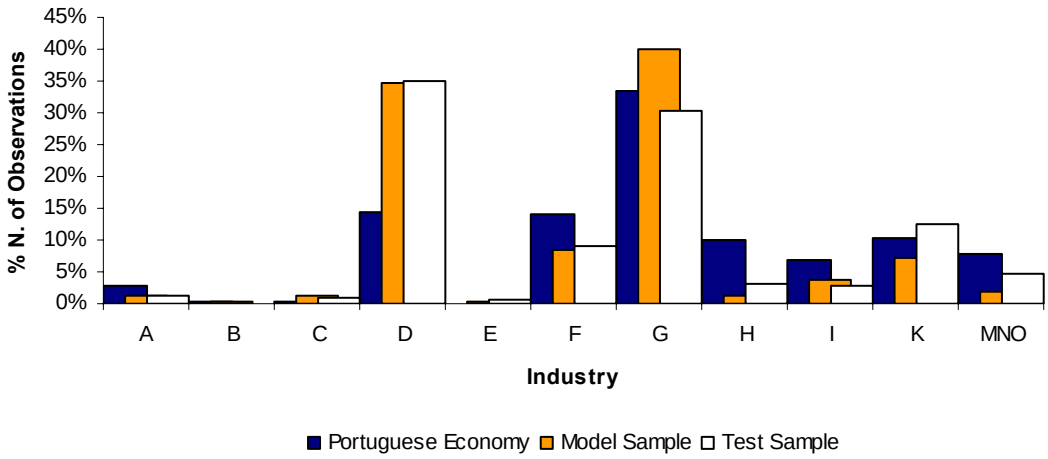

Figure 1 – Economy-Wide vs. Study Samples Industry Distribution

This figure displays the industry distribution for the firms in our model and test sample and for the Portuguese economy in 2003 (INE 2003). The industry types considered are: A – Agriculture, Hunting & Forestry; B – Fishing; C – Mining & Quarrying; D – Manufacturing; E – Electricity, Gas & Water Supply; F – Construction; G – Wholesale & Sale Trade; H – Hotels & Restaurants; I – Transport, Storage & Communications; K – Real Estate, Renting & Business Activities; MNO – Education/ Health & Social Work/ Other Personal Services Activities.

Figures 2, 3 and 4 display the industry, size (measured by annual turnover) and yearly distributions respectively, for both the default and non-default groups of observations of the model dataset.

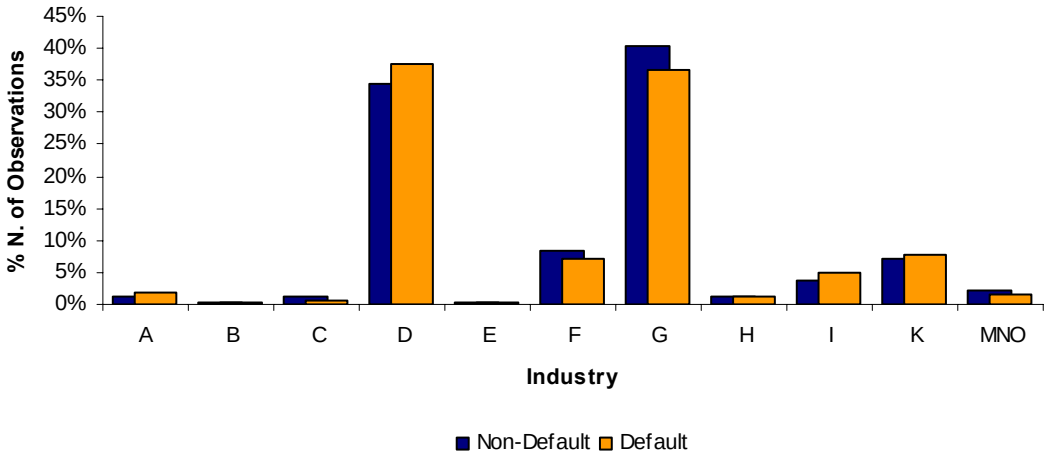

Figure 2 – Sample Industry Distribution

This figure shows the model sample industry distributions for the firms that defaulted and for the firms that have not defaulted. The industry types considered are: A – Agriculture, Hunting & Forestry; B – Fishing; C – Mining & Quarrying; D – Manufacturing; E – Electricity, Gas & Water Supply; F – Construction; G – Wholesale & Sale Trade; H – Hotels & Restaurants; I – Transport, Storage & Communications; K – Real Estate, Renting & Business Activities; MNO – Education/ Health & Social Work/ Other Personal Services Activities.

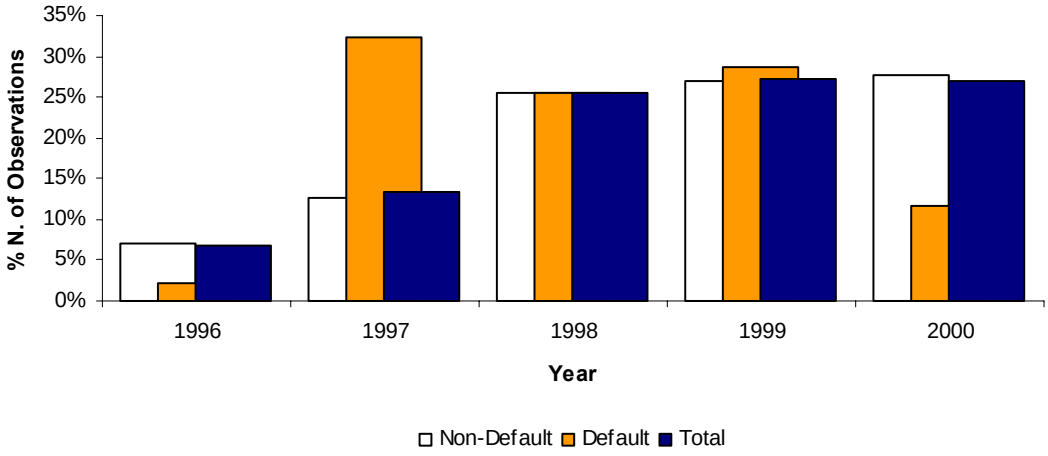

Figure 3 – Accounting Statement Yearly Distribution

The figure above displays the yearly distribution of the financial statements in the dataset for the default, non-default and total observations.

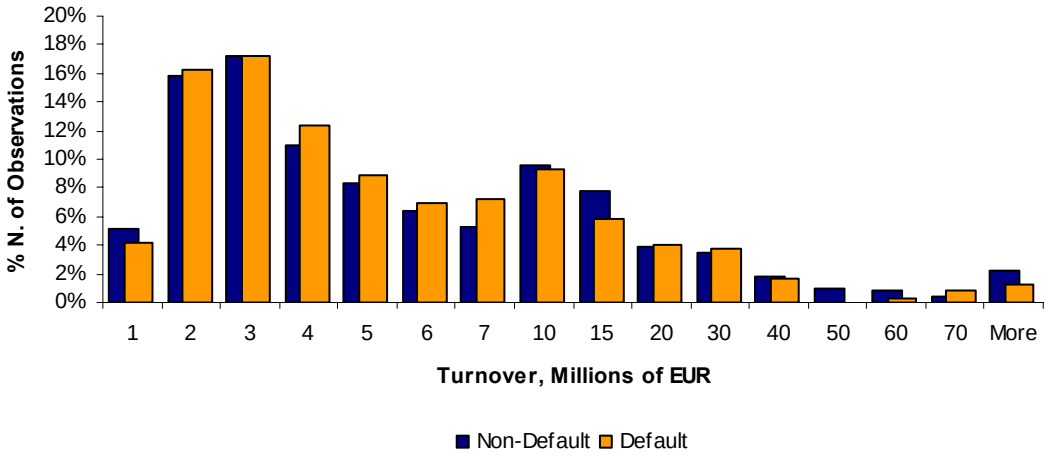

Figure 4 – Size (Turnover) Distribution, Millions of EUR
The figure shows the size distribution of the default and non-default observations. Size is measured by the firms' Turnover, defined as the sum of total sales plus services rendered.

Analysis of industry distribution (Figure 2) suggests high concentration on industries *G – Trade* and *D – Manufacturing*, both accounting for about 75% of the whole sample. The industry distributions for both default and non-default observations are very similar.

Figure 3 shows that observations are uniformly distributed per year, for the last three periods, with about 3,000 observations per year. For the non-default group of observations, the number of yearly observations rises steadily until the third period, and then remains constant until the last period. For the default group, the number of yearly observations has a great increase in the second period and clearly decreases in the last.

Figure 4 shows size distribution and indicates that most of the observations belong to the Small and Medium size Enterprises - SME segment, with annual turnover up to 50 million EUR (according to the NBCA SME classification). The SME segment accounts for about 95% of the whole sample. The distributions of both non-default and default observations are very similar.

3 Financial Ratios and Bivariate Analysis

A preliminary step before estimating the scoring model is to conduct a bivariate analysis for each potential explanatory variable, in order to select the most intuitive and powerful ones. In this study, the scoring model considers exclusively financial ratios as explanatory variables. A list of twenty-three ratios representing six different dimensions – Profitability, Liquidity, Leverage, Debt Coverage, Activity and Productivity – is considered. The bivariate analysis relates each of the twenty-three ratios and a default indicator, in order to assess the discriminatory power of each variable. Appendix 1 provides the list of the variables and how they are constructed. Figures 5 – 10 provide a graphical description, for some selected variables, of the relationship between each variable individually and the default frequency. The data is ordered in ascending order by the value of each ratio and, for each decile, the default frequency is calculated (number of defaults divided by the total number of observations in each decile).

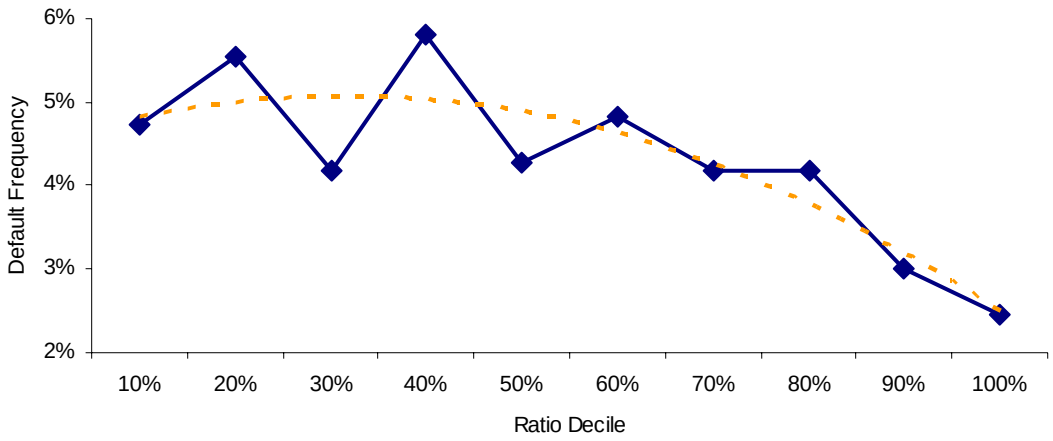

Figure 5 – Average Default Frequency by *Liquidity / Current Liabilities* Ratio Decile
The figure shows the relationship between the Liquidity / Current Liabilities Ratio and the historical default frequency. Loans are ranked in ascending order, in terms of the value of the ratio. For each decile the average default frequency is calculated.

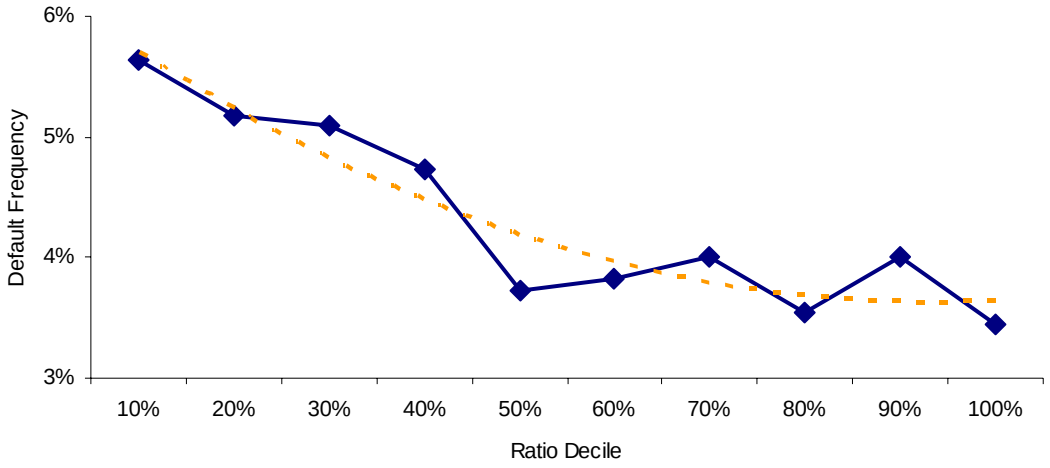

Figure 6 – Average Default Frequency by *Current Ratio* Decile

The figure shows the relationship between the Current Ratio and the historical default frequency. Loans are ranked in ascending order, in terms of the value of the ratio. For each decile the average default frequency is calculated.

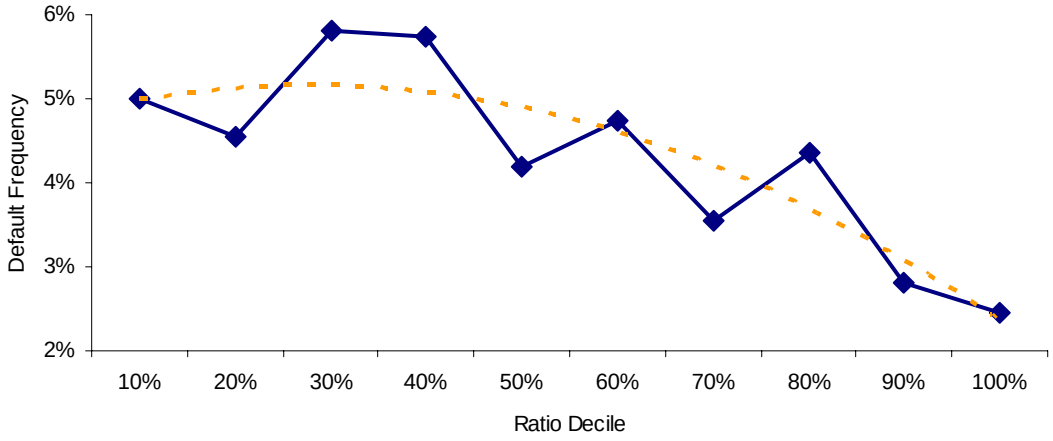

Figure 7 – Average Default Frequency by *Liquidity / Assets* Ratio Decile

The figure shows the relationship between the Liquidity / Assets Ratio and the historical default frequency. Loans are ranked in ascending order, in terms of the value of the ratio. For each decile the average default frequency is calculated.

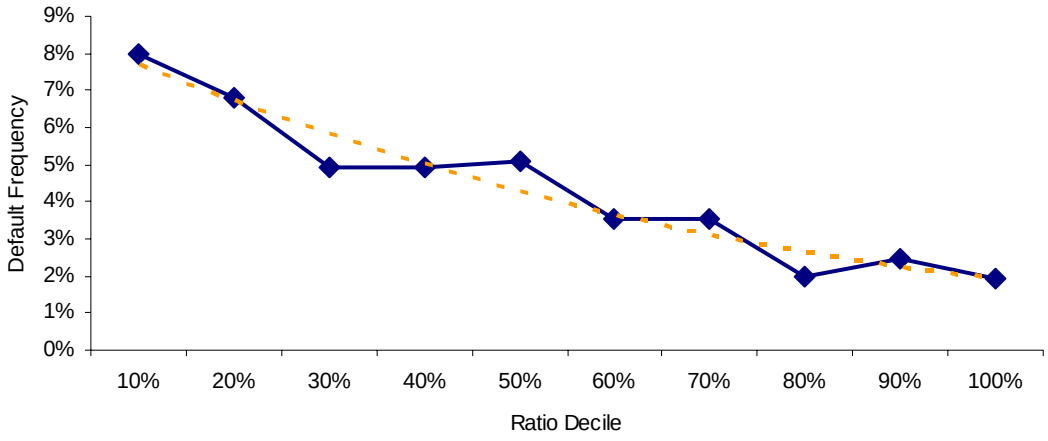

Figure 8 – Average Default Frequency by *Debt Service Coverage* Ratio Decile

The figure shows the relationship between the Debt Service Coverage Ratio and the historical default frequency. Loans are ranked in ascending order, in terms of the value of the ratio. For each decile the average default frequency is calculated.

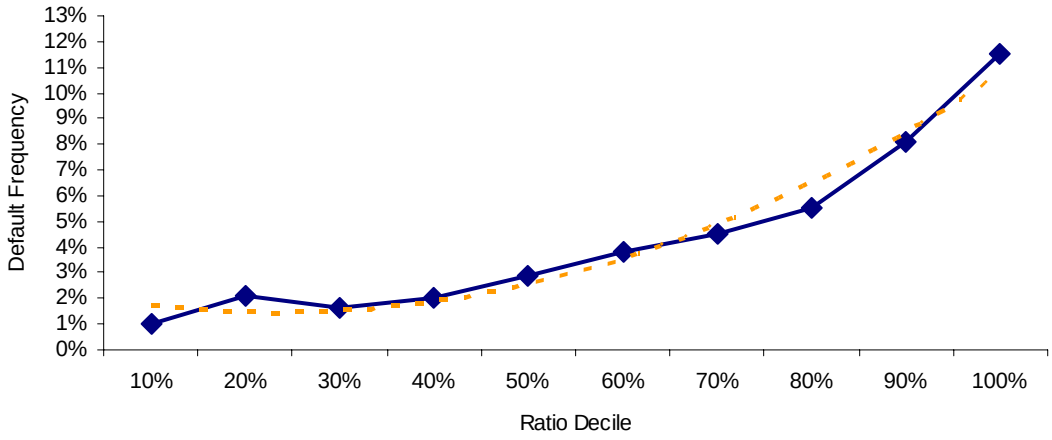

Figure 9 – Average Default Frequency by *Interest Costs / Sales* Ratio Decile

The figure shows the relationship between the Interest Costs / Sales Ratio and the historical default frequency. Loans are ranked in ascending order, in terms of the value of the ratio. For each decile the average default frequency is calculated.

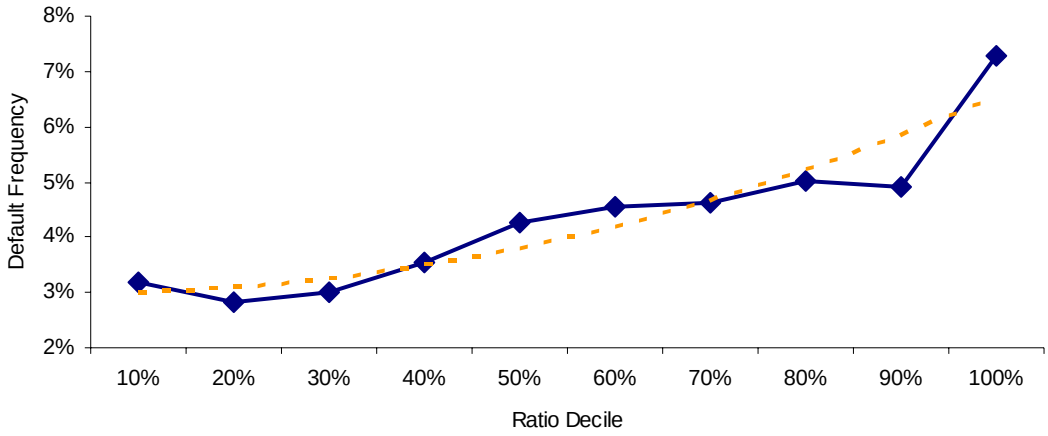

Figure 10 – Average Default Frequency by *Productivity Ratio* Decile
The figure shows the relationship between the Productivity Ratio and the historical default frequency. Loans are ranked in ascending order, in terms of the value of the ratio. For each decile the average default frequency is calculated.

In order to have a quantitative assessment of the discriminating power of each variable, the Accuracy Ratio is used[7]. The computed values of the Accuracy Ratios are reported in Appendix 1.

The selected variables for the multivariate analysis comply with the following criteria:

- They must have discriminating power, with an Accuracy Ratio higher than 5%;
- The relationship with the default frequency should be clear and economically intuitive. For example, ratio *Current Earnings and Depreciation / Turnover* should have a negative relationship with the default frequency, since firms with a high percentage of EBITDA over Turnover should default less frequently; analyzing Figure 11, there seems to be no clear relationship for this dataset;
- The number of observations lost due to lack of information on any of the components of a given ratio must be insignificant. Not all firms have the same

[7] The Accuracy Ratio can be used as a measure of the discriminating power of a variable, comparing the ability of the variable to correctly classify the default and non-default observations against that of a random variable, unrelated to the default process. Section 4.1.1 provides a more detailed description.

degree of accuracy on their accounting reports, for example, ratios *Bank Debt / Accounts Payable* and *P&L / L-T Liabilities* have a significant amount of missing data for the components *Debt to Credit Institutions* and *Long-Term Liabilities* respectively.

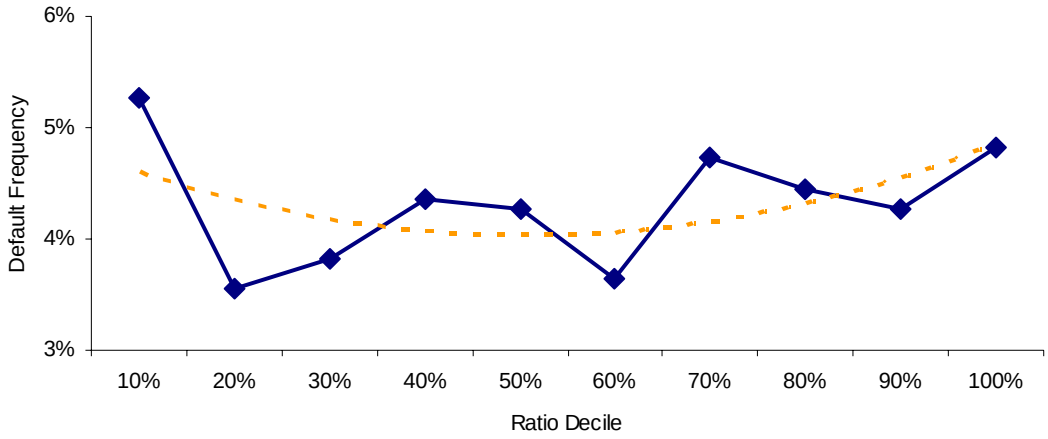

Figure 11 – Average Default Frequency by *Current Earnings and Depreciation / Turnover* Ratio Decile

The figure shows the relationship between the Current Earnings and Depreciation / Turnover Ratio and the historical default frequency. Loans are ranked in ascending order, in terms of the value of the ratio. For each decile the average default frequency is calculated.

At this point, nine variables are eliminated and are not considered on the multivariate analysis. All the remaining variables are standardized in order to avoid scaling issues[8].

[8] Standardization consists on subtracting the value of the variable by its average on the sample and dividing the result by its sample standard deviation.

4 Scoring Model and Validation

The dependent variable Y_{it} is the binary discrete variable that indicates whether firm i has defaulted (one) or not (zero) in year t. The general representation of the model is:

$$Y_{it} = f\left(\beta_k, X_{it-1}^k\right) + e_{it}, \tag{1}$$

where X_{it-1}^k represents the values of the k explanatory variables of firm i, one year before the evaluation of the dependent variable. The functional form selected for this study is the Logit model[9]. Alternative specifications can be considered, such as Probit, Linear Probability Model, or even Genetic Algorithms, although there is no evidence in the literature that any alternative specification can consistently outperform the Logit specification in credit default prediction (Altman, Marco and Varetto, 1994 and Yang et al., 1999).

Using both forward and backward procedures, the selected model is the one that complies with the validation criteria and has the higher discriminating power, measured by the Accuracy Ratio.

4.1 Model Validation

The variables selected on Section 3 are pooled together in order to obtain a model that is at the same time:

- Parsimonious but powerful: high discriminating power with few parameters to estimate;
- Statistically significant: all variables individually and the model as a whole must be significant, with low correlation between the variables;

[9] Refer to Appendix 3 for a description of the Logit model.

- Intuitive: the sign of the estimated parameters should make economic sense and the selected variables should represent the various relevant risk factors.

4.1.1 Efficiency

A model with high discriminatory power is a model that can clearly distinguish the default and non-default populations. In other words, it is a model that makes consistently "good" predictions relative to few "bad" predictions. For a given cut-off value[10], there are two types of "good" and "bad" predictions:

		Estimated	
		Non-Default	**Default**
Observed	**Non-Default**	True	False Alarm (Type II Error)
	Default	Miss (Type I Error)	Hit

- The "good" predictions occur if, for a given cut-off point, the model predicts a default and the firm does actually default (Hit), or, if the model predicts a non-default and the firm does not default in the subsequent period (True).
- The "bad" prediction occurs if, for a given cut-off point, the model predicts a default and the firm does not actually defaults (False-Alarm or Type II Error), or if the model predicts a non-default and the firm actually defaults (Miss or Type I Error).
- The Hit Ratio (HR) corresponds to the percentage of defaults from the total default population that are correctly predicted by the model, for a given cut-off point.

[10] The cut-off point is the value from which the observations are classified as "good" or "bad". For example, given a cut-off point of 50%, all observations with an estimated score between 0% and 50% will be classified as "good", and those between 50% and 100% will be considered "bad".

- The False Alarm Ratio (FAR) is the percentage of False Alarms or incorrect default predictions from the total non-defaulting population, for a given cut-off point.

Several alternatives could have been considered in order to analyze the discriminating power of the estimated models. In this study, both ROC/CAP analysis and Kolmogorov-Smirnov (KS) analysis are performed.

Receiver Operating Characteristics (ROC) and Cumulative Accuracy Profiles (CAP) curves are two closely related graphical representations of the discriminatory power of a scoring system. Using the notation from Sobehart and Keenan (2001), the ROC curve is a plot of the HR against the FAR, while the CAP curve is a plot of the HR against the percentage of the sample.

For the ROC curve, a perfect model would pass through the point (0,1) since it always makes "good" predictions, and never "bad" predictions (it has FAR = 0% and a HR = 100% for all possible cut-off points). A "naïve" model is not able to distinguish defaulting from non-defaulting firms, thus will do as many "good" as "bad" predictions, though for each cut-off point, the HR will be equal to the FAR. A better model would have a steeper curve, closer to the perfect model, thus a global measure of the discriminant power of the model would be the area under the ROC curve. This can be calculated as[11]:

$$AUROC = \int_0^1 HR(FAR)d(FAR), \tag{2}$$

For the CAP or Lorenz curve, a perfect model would attribute the lowest scores to all the defaulting firms, so if x% of the total population are defaults, then the CAP curve of a perfect model would pass through the point (x,1). A random model would make as many "good" as "bad" predictions, so for the y% lowest scored firms it would have a HR of y%. Then, a global measure of the discriminant power of the model, the Accuracy Ratio (AR), compares the area between the CAP curve of the

[11] Refer to Appendix 2 for a technical description of the AUROC calculation.

model being tested and the CAP of the random model, against the area between the CAP curve of the perfect model and the CAP curve of the random model.

It can be shown that there is a linear relationship between the global measures resulting from the ROC and CAP curves[12]:

$$AR = 2\left(AUROC - 0.5\right), \tag{3}$$

The KS methodology considers the distance between the distributions of *1 – HR* (or Type I Errors) and *1 – FAR* (or True predictions) [13]. The higher the distance between the two distributions, the better the discriminating power of the model. The KS statistic corresponds to the maximum difference for any cut-off point between the *1 – FAR* and *1 – HR* distributions.

4.1.2 Statistical Significance

All estimated regressions are subject to a variety of statistical tests, in order to ensure the quality of the results at several levels:

i. Residual Analysis is performed with the purpose of testing the distributional assumption of the errors of the regression. Although the logistic regression assumes that the errors follow a binomial distribution, for large samples (such as the one in this study), it approximates the normal distribution. The standardized residuals from the logistic regressions should then follow a standard normal distribution[14]. At this stage, severe outliers are identified and eliminated. These outliers are observations for which the model fits poorly

[12] See, for example, Engelmann, Hayden and Tasche (2003).
[13] The Kolmogorov-Smirnov statistic is a non-parametric statistic used to test whether the density function of a variable is the same for two different groups (Conover, 1999).
[14] The standardized residuals correspond to the residuals adjusted by their standard errors. This adjustment is made in logistic regression because the error variance is a function of the conditional mean of the dependent variable.

(has an absolute studentized residual[15] greater than 2), and that can have a very large influence on the estimates of the model (a large DBeta[16]).

ii. The significance of each estimated coefficient is tested using the Wald test. This test compares the maximum likelihood value of the estimated coefficient to the estimate of its standard error. This test statistic follows a standard normal distribution under the hypothesis that the estimated coefficient is null. For the three models, all of the estimated coefficients are significant at a 90% significance level.

iii. In order to test the overall significance of each estimated model, the Hosmer-Lemeshow (H-L) test is used. This goodness-of-fit test compares the predicted outcomes of the logistic regression with the observed data by grouping observations into risk deciles.

iv. After selecting the best linear model, the assumption of linearity between each variable and the logit of the dependent variable is checked. This is performed in four stages:

1- The Box-Tidwell test (Box-Tidwell, 1962) is performed on all continuous variables, in order to confirm the linearity assumption;

2- For all variables that failed the linearity test in the previous step, a plot of the relationship between the covariate and the logit is presented, providing evidence on the type of non-linear relationship;

3- For all continuous variables with significant non-linear relationships with the logit, the fractional polynomial methodology is implemented (Royston and Altman, 1994) in order to adequately capture the true relationship between the variables;

4- Check whether the selected transformation makes economic sense.

v. The last assumption to be checked is the independence between the explanatory variables. If multicolinearity is present, the estimated coefficients

[15] The studentized residual corresponds to the square root of the change in the -2 Log Likelihood of the model attributable to deleting the case from the analysis. It follows an asymptotical normal distribution and extreme values indicate a poor fit.

[16] DBeta is an indicator of the standardized change in the regression estimates obtained by deleting an individual observation.

will be unbiased but their estimated standard errors will tend to be large. In order to test for the presence of high multicolinearity, a linear regression model using the same dependent and independent variables is estimated, and the tolerance statistic is calculated for each independent variable[17]. If any of the tolerance statistics are below 0.20 then it is assumed that we are in the presence of high multicolinearity, and the estimated regression is discarded.

4.1.3 Economic Intuition

All estimated coefficients follow economic intuition in the sense that the sign of the coefficients indicates the expected relationship between the selected variable and the default frequency. For example, if for a given model the estimated coefficient for variable *Productivity Ratio* is +0.123, this means that the higher the Personnel Costs relative to the Turnover, the higher the estimated credit score of the firm. In other words, firms with lower labor productivity have higher credit risk. For the non-linear relationships it is best to observe graphically the estimated relationship between the independent variable and the logit of the dependent. As for the linear case, this relationship should be monotonic, either always positive or negative. The difference is that the intensity of this relationship is not constant, it depends on the level of the independent variable.

During the model estimation two hypotheses are tested:

1. Whether a system of unrelated equations, by industry group yields better results than a single-equation model for all industries;
2. Whether a model where the observations are weighted in order to increase the proportion of defaults to non-defaults in the estimation sample, performs better than a model with unweighted observations.

[17] The tolerance statistic corresponds to the variance in each independent variable that is not explained by all of the other independent variables.

4.2 Model A – Multiple Industry Equations Model

In order to test the hypothesis that a system of unrelated equations by industry group yields better results than a single-equation model for all industries, the dataset is broken into two sub-samples: the first one for *Manufacturing & Primary Activity* firms, with 5,046 observations of which 227 are defaults; and the second for *Trade & Services* firms, with 5,954 observations and 248 defaults. If the nature of these economic activities has a significant and consistent impact on the structure of the accounting reports, then it is likely that a model accommodating different variables for the different industry sectors performs better than a model which forces the same variables and parameters to all firms across industries[18]. The model is:

$$\hat{Y}_i = \frac{\exp(\hat{\mu}_i)}{1+\exp(\hat{\mu}_i)}, \tag{4}$$

for the two-equation model,

$$\hat{\mu}_i = \begin{cases} X_i^a{}'\hat{\beta}^a & \text{if } i \text{ belongs to industry a} \\ X_i^b{}'\hat{\beta}^b & \text{if } i \text{ belongs to industry b} \end{cases}, \tag{5}$$

for the single-equation model,

$$\hat{\mu}_i = X_i{}'\hat{\beta} \quad \forall i, \tag{6}$$

For the final model, the selected variables and estimated coefficients are presented in the table below[19]:

[18] Model performance is measured by the ability to discriminate between default and non-default populations, which can be summarized by the Accuracy Ratio.
[19] Refer to Appendix 4 for full estimation results.

Industry a			Industry b		
Variable	$\hat{\beta}$	**Wald Test P-Value**	**Variable**	$\hat{\beta}$	**Wald Test P-Value**
Liquidity / CurLiabilities	-0.381	0.003	Current Ratio	-0.212	0.005
Debt Service Coverage	-0.225	0.021	Liquidity / Assets	-0.160	0.063
Interest Costs / Sales_1	2.011	0.002	Debt Service Coverage	-0.184	0.041
Interest Costs / Sales_2	-0.009	0.000	Interest Costs / Sales_1	1.792	0.000
Productivity Ratio	0.200	0.028	Interest Costs / Sales_2	-0.009	0.038
Constant	-3.259	0.000	Constant	-3.426	0.000
Number of Observations		5,044	**Number of Observations**		5,951
-2 LogLikelihood		1,682	**-2 LogLikelihood**		1,913
H-L Test P-Value		0.415	**H-L Test P-Value**		0.615

Table 1 – Estimated Model Variables and Parameters, Model (A)

The table above presents the estimation results for the two industry equation model. *Industry a* represents Manufacturing & Primary Activity firms, and *Industry b* Trade & Services firms. The sign of the estimated parameters for both regressions is in accordance with economic intuition. The significance of each parameter is demonstrated by the low p-values for the Wald test, while the overall significance of each regression is verified by the high p-values for the Hosmer-Lemeshow test.

The Hosmer-Lemeshow test is a measure of the overall significance of the logistic regression. Through the analysis of Figure 12 we can conclude that the estimated logistic regressions significantly fit the observed data.

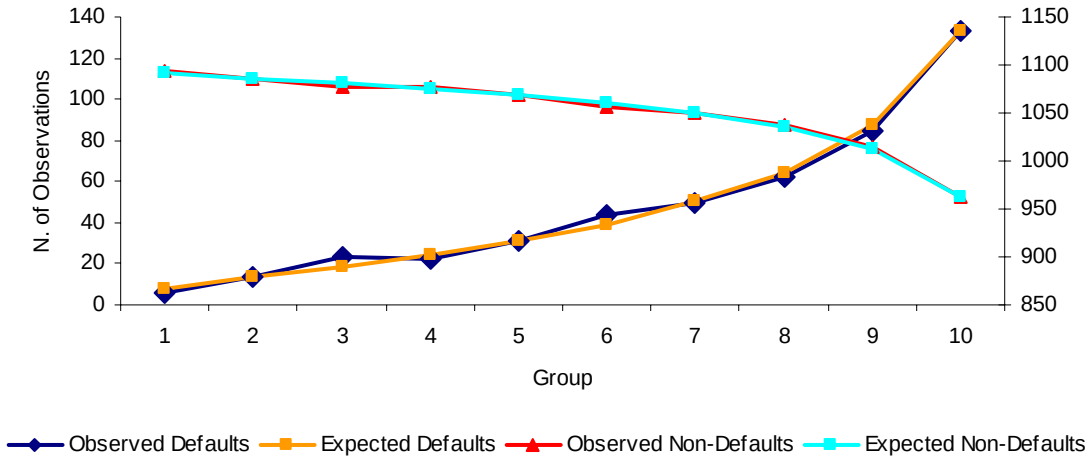

Figure 12 – Model A: Hosmer-Lemeshow Test

This figure presents the comparison between the observed and expected number of default and non-default observations for each of the 10 groups comprised in the Hosmer-Lemeshow test. The number of the default observations is represented on the left y axis, while the number of non-default observations is represented on the right y axis.

4.3 Model B – Standard Model

In order to test our two hypotheses both the Two-Equation Model and the Weighted Sample Model will be evaluated against the standard setting of a single equation across all industries, using an unweighted sample. Table 2 summarizes the final results under this standard setting and Figure 13 provides a graphical description of the overall significance of the estimated model.

Variable	$\hat{\beta}$	Wald Test P-Value
Current Ratio	-0.171	0.001
Liquidity / Assets	-0.211	0.002
Debt Service Coverage	-0.231	0.001
Interest Costs / Sales_1	1.843	0.007
Interest Costs / Sales_2	-0.009	0.000
Productivity Ratio	0.124	0.003
Constant	-3.250	0.000
Number of Observations		10,995
-2 LogLikelihood		3,600
H-L Test P-Value		0.973

Table 2 – Estimated Model Variables and Parameters, Model (B)

This table displays the estimation results for the single-equation model. The sign of the estimated parameters agrees with economic intuition. The significance of each parameter is demonstrated by the low p-values for the Wald test, while the overall significance of the regression is verified by the high p-values for the Hosmer-Lemeshow test.

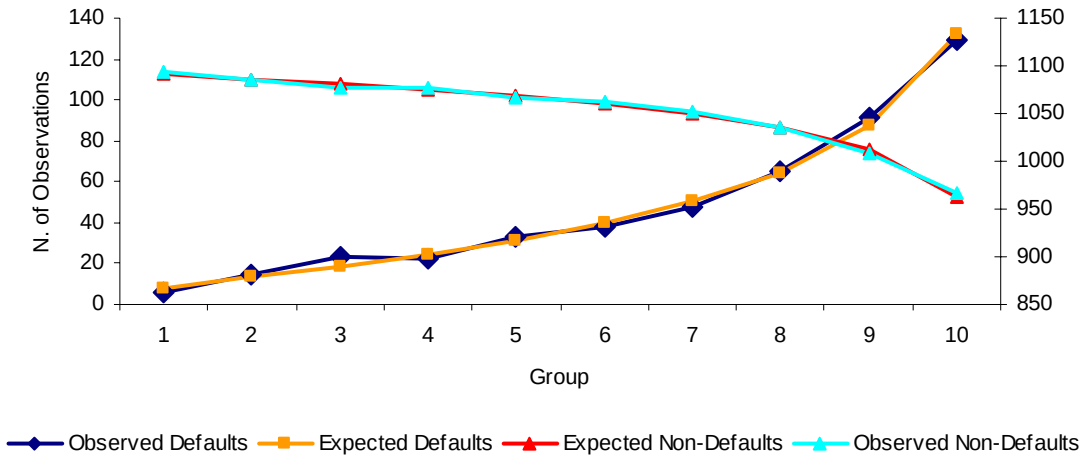

Figure 13 – Model B: Hosmer-Lemeshow Test
This figure presents the comparison between the observed and expected number of default and non-default observations for each of the 10 groups comprised in the Hosmer-Lemeshow test. The number of the default observations is represented on the left y axis, while the number of non-default observations is represented on the right y axis.

4.4 Model C – Weighted Sample Model

The proportion of the number of defaults (450) to the total number of observations in the sample (11,000) is artificially high. The real average annual default frequency of the bank's portfolio and the Portuguese economy is significantly lower than the 4.32% suggested by our sample for the corporate sector. However, in order to be able to correctly identify the risk profiles of "good" and "bad" firms, a significant number of observations for each population is required. For example, keeping the total number of observations constant, if the correct default rate is about 1%, extracting a random sample in accordance to this ratio would result in a proportion of 110 default observations to 11,000 observations.

A consequence of having an artificially high proportion of default observations is that the estimated scores cannot be directly interpreted as real probabilities of default. Therefore, these results have to be calibrated in order to obtain default probabilities estimates.

A further way to increase the proportion of the number of default observations is to attribute different weights to the default and non-default observations. The

weightening of observations could potentially have two types of positive impact in the analysis:

1. As mentioned above, a more balanced sample, with closer proportion of default to non-default observations, could help the Logit regression to better discriminate between both populations;
2. The higher proportion of default observations results in higher estimated scores. As a consequence, the scores in the weighed model are more evenly spread throughout the]0,1[interval (see Figure 14). If, in turn, these scores are used to group the observations into classes, then it could be easier to identify coherent classes with the weighed model scores. Thus, even if weightening the observations does not yield a superior model in terms of discriminating power, it might still be helpful later in the analysis, when building the rating classes.

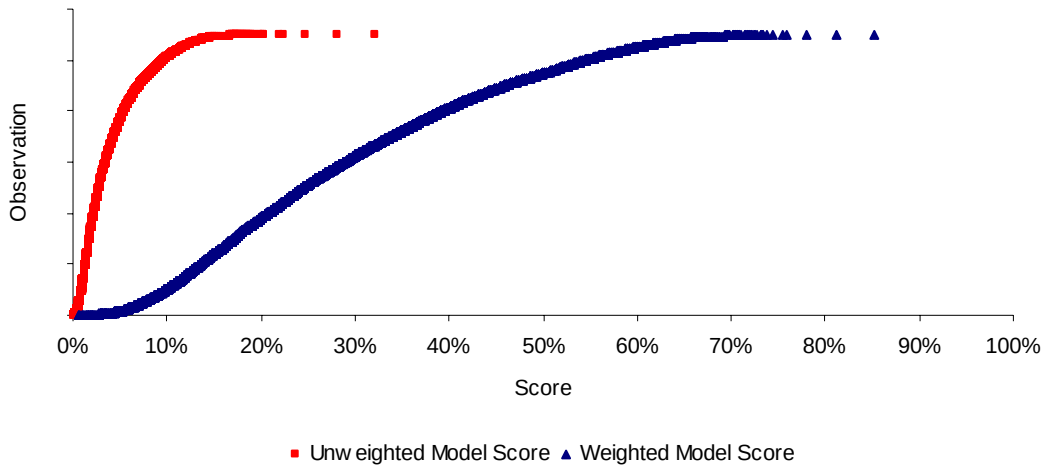

Figure 14 – Weighted vs. Unweighted Score
The figure presents the estimated scores for each observation in the development sample using both weighted and unweighted models.

The weighted model estimated considers a proportion of one default observation for two non-default observations. The weighed sample consists of 1,420 observations, of which 470 are defaults and the remaining 950 are non-default

observations[20]. The optimized model selects the same variables as the unweighted model though with different estimated coefficients.

Variable	$\hat{\beta}$	Wald Test P-Value
Current Ratio	-0.197	0.003
Liquidity / Assets	-0.223	0.006
Debt Service Coverage	-0.203	0.013
Interest Costs / Sales_1	1.879	0.050
Interest Costs / Sales_2	-0.009	0.000
Productivity Ratio	0.123	0.023
Constant	-0.841	0.000
Number of Observations		1,420
-2 LogLikelihood		1,608
H-L Test P-Value		0.465

Table 3 – Estimated Model Variables and Parameters, Model (C)

This table shows the estimation results for the weighted sample model. The selected variables are the same as for the unweighted model (B). All estimated parameters are significant at a 5% level and the suggested relationship with the dependent variable concurs with economic intuition. The high p-values for the Hosmer-Lemeshow test attest the overall significance of the regression.

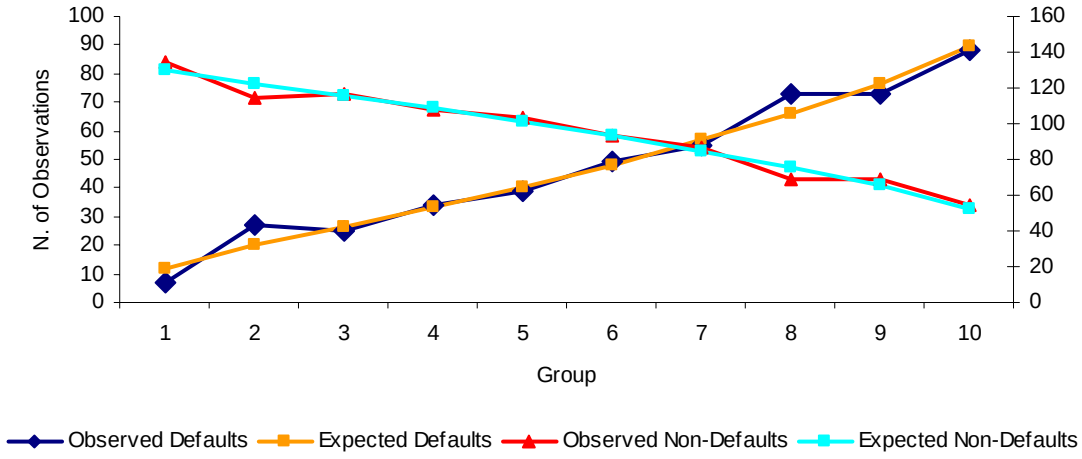

Figure 15 – Model C: Hosmer-Lemeshow Test

This figure presents the comparison between the observed and expected number of default and non-default observations for each of the 10 groups comprised in the Hosmer-Lemeshow test. The number of the default observations is represented on the left y axis, while the number of non-default observations is represented on the right y axis.

[20] Other proportions yield very similar results (namely the one default for one non-default and one default for three non-defaults proportions).

The following section analyses the estimation results in more detail and compares the different approaches in terms of efficiency.

4.5 Analysis of the Results

In Appendix 4, the final results of the estimations are presented for all three models: the two-equation model (Model A), the unweighted single-equation model (Model B) and the weighted single-equation model (Model C). The first step to obtain each model is to find the best linear combination through backward and forward selection procedures. The estimation equation that complies with both economic intuition and positive statistical diagnosis (described in steps i. to iii. of section 4.1.2), and had the higher discriminating power is considered the optimal linear model.

The second step is to check for non-linear relationships between the independent variables and the logit of the dependent. Results indicate that for all four selected linear regressions, there is a clear non-linear relationship between variable *Interest Costs / Sales* and the logit of the dependent variable. In order to account for this fact, the procedure described in step iv. of section 4.1.2 is implemented. The resulting non-linear relationship for the four regressions is illustrated in Figure 16. In order to depict graphically the relationship between the covariate and the binary dependent variable, a Locally Weighted Scatterplot Smoothing, or Lowess (Cleveland 1979), was created. In addition, the quality of the fit of this relationship for the three estimated models can be accessed by comparing the multivariate adjustment for each model with the lowess curve. For all three models, the quality of the adjustment is high but deteriorates for very high values of the explanatory variable.

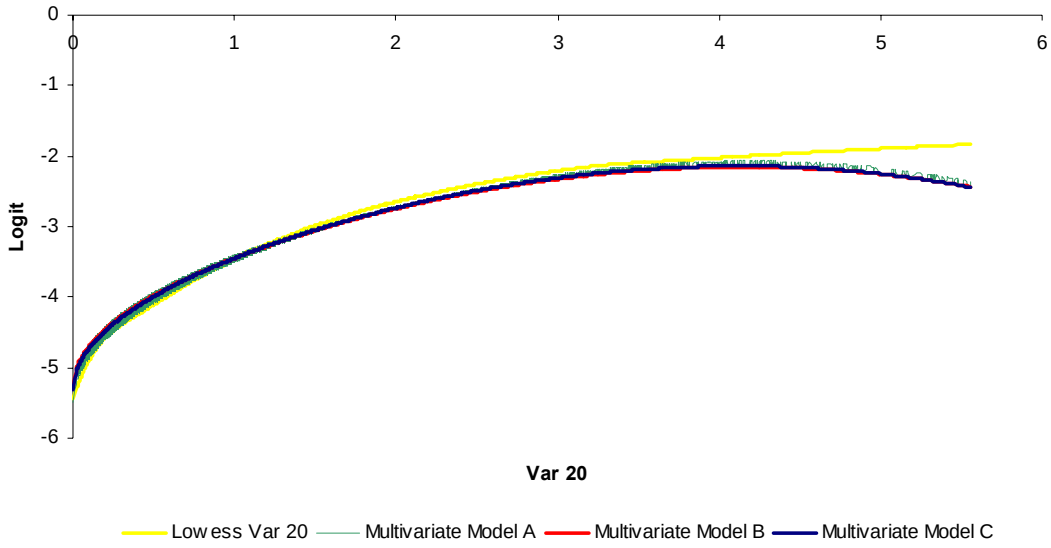

Figure 16 – Smoothed Lowess and Fractional Polynomial Adjustment for the *Interest Costs / Sales* Ratio

The figure compares the plot of the bivariate smoothed Lowess logit of the variable *Interest Costs / Sales* with the multivariate fractional polynomial adjustment for models A – Multiple Industry Equations Model, B – Standard Model and C – Weighted Sample Model.

After the optimal non-linear regressions are selected, a final test for multicolinearity is implemented. Only the *Trade & Services* regression of the Two-Equation Model presented signs of severe multicolinearity. Since there is no practical method to correct this problem, the model is discarded and the second best model suggested by the fractional polynomial procedure is selected. This alternative specification does not suffer from multicolinearity, as it can be observed in the results presented in Appendix 4[21]. In short, the modeling procedure consisted on selecting the best discriminating regression from a pool of possible solutions that simultaneously complied with economic and statistical criteria.

In terms of efficiency, all three models have a small number of selected variables: model A has five variables for each equation, while models B and C have six variables each. Analyzing Figures 17 and 18, we can conclude that all three models have significant discriminating power and have similar performances. Results

[21] In order to ensure stability of the final results, the whole modeling procedure is repeated with several random sub-samples of the main dataset. Across all sub-samples the variables selected for each model are the same, the values of the estimated coefficients are stable, and the estimated AR's are similar.

for Altman's Z'-Score Model for Private Firms (Altman, 2000) are also reported as a benchmark (Model D). Figure 17 displays the Receiver Operating Characteristics (ROC) curves. The ROC curve provides for each possible cut-off value the proportion of observations incorrectly classified as default by the model against the proportion correctly classified as default. The three suggested models have similar ROC curves, clearly above the Random Model and Z'-Score Model curves.

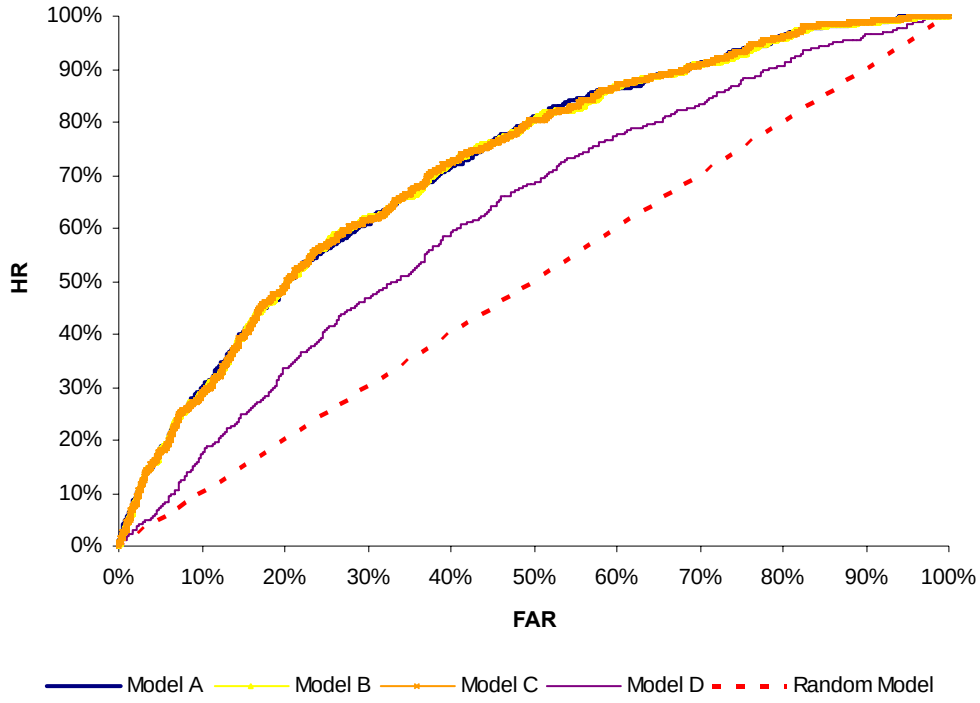

Figure 17 – Receiver Operating Characteristics Curves
The figure above displays the Receiver Operating Characteristics curves for the three estimated models (A – Multiple Industry Equations Model, B – Standard Model and C – Weighted Sample Model) and for the Z'-Score Model for Private Firms (Altman, 2000). The ROC curve provides for each possible cut-off value the proportion of observations incorrectly classified as default by the model, the False Alarm Ratio (FAR), against the proportion correctly classified as default, the Hit Ratio (HR).

Figure 18 displays the Cumulative Accuracy Profiles (CAP) curves. The CAP curve provides, for a given proportion of observations with the highest estimated scores, the proportion of correctly classified default observations. As with the ROC analysis, the curves for the three selected models are similar and clearly above the Random Model and Z'-Score Model curves.

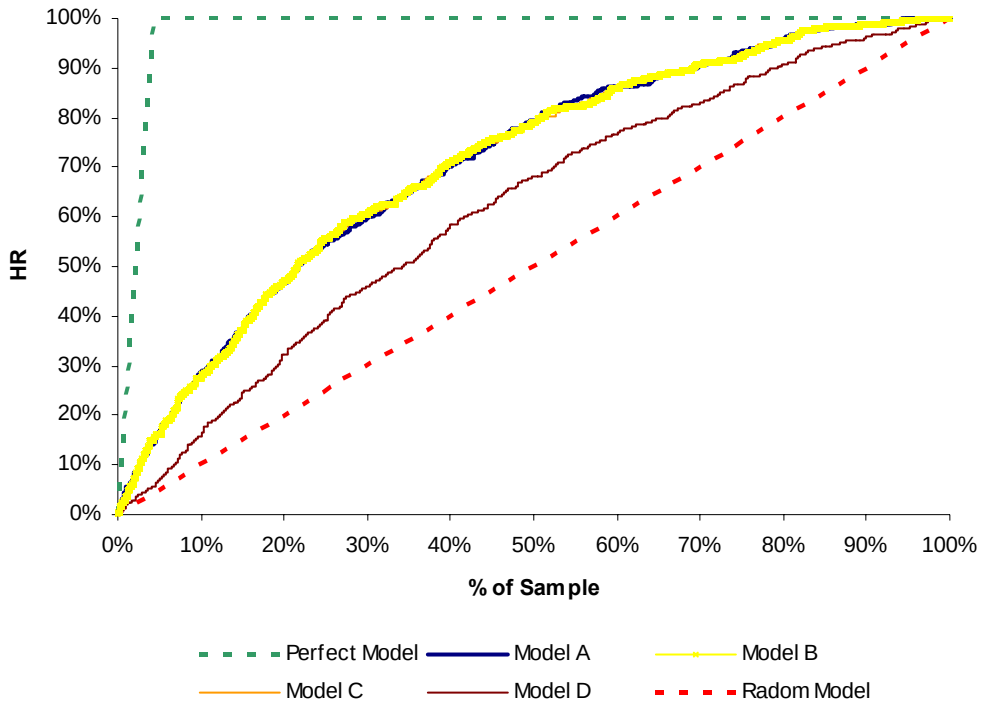

Figure 18 – Cumulative Accuracy Profiles Curves

This figure displays the Cumulative Accuracy Profiles curves for the three estimated models (A – Multiple Industry Equations Model, B – Standard Model and C – Weighted Sample Model) and for the Z'-Score model (Model D). The CAP curve provides, for a given proportion of observations with the highest estimated scores, the proportion of correctly classified default observations (the Hit Ratio, HR).

Figures 23-28 in Appendix 5 provide both the Kolmogorov-Smirnov (KS) analysis and Error Type curves. The KS analysis consists on evaluating for each possible cut-off point the distance between the Type I Error curve and the True Prediction curve. The higher the distance between the curves, the better the discriminating power of the model. The Error Type curves display for each cut-off point the percentages of Type I (incorrectly classifying an observation as non-default) and Type II (incorrectly classifying an observation as default) errors for each model.

Table 4 summarizes the results for both ROC/CAP analysis and KS analysis, under both the estimation and testing samples. All three measures of discriminating power, under both samples, indicate similar and positive values for the three models estimated, clearly above the Z'-Score model.

Model	Main Sample				Out-of-Sample		
	AUROC	σAUROC	AR	KS	AUROC	σAUROC	AR
A	71.88%	1.15%	43.75%	32.15%	73.04%	7.53%	46.07%
B	71.88%	1.15%	43.77%	32.97%	75.29%	6.55%	50.59%
C	71.87%	1.15%	43.74%	32.94%	74.15%	6.88%	48.29%
D	62.53%	1.25%	25.07%	19.77%	61.11%	6.87%	22.22%

Table 4 – AUROC, AR and KS Statistics

This table reports the Area Under the ROC curves, Accuracy Ratios and Kolmogorov-Smirnov statistics estimated for the three suggested models (A – Multiple Industry Equations Model, B – Standard Model and C – Weighted Sample Model) and for the Z'-Score model (Model D), under both the estimation and testing samples.

A more rigorous comparison of the discriminating power of the models can be obtained through a statistical test presented in DeLong et al. (1988) for the difference between the estimated AUROC's of the different models[22]. Table 5 presents the results of applying this test to the differences between all models for both samples.

Test	Main Sample			Out-of-Sample		
	θi - θj	σ (θi - θj)	P-Value	θi - θj	σ (θi - θj)	P-Value
A - B	-0.0089%	0.2225%	96.83%	-2.2571%	2.8844%	43.39%
A - C	0.0053%	0.2372%	98.23%	-1.1086%	2.7449%	68.63%
A - D	9.3425%	1.7807%	0.00%	11.9256%	7.7745%	12.50%
B - C	0.0141%	0.0476%	76.68%	1.1485%	0.5115%	2.47%
B - D	9.3514%	1.7788%	0.00%	14.1827%	6.7577%	3.58%
C - D	9.3372%	1.7751%	0.00%	13.0342%	7.0051%	6.28%

Table 5 – Testing the Differences between AUROC's

The table above provides the results of a statistical test for comparing the estimated AUROC curves between the different models. Model A is the Multiple Industry Equations model, Model B the Standard Model, Model C the Weighted Sample and Model D the Z'-Score model.

The results indicate that for both samples, Models A, B and C have similar discriminating power, and all three perform significantly better that the Z'-Score model.

[22] For a description of the test consult Appendix 2.

Regarding our first hypothesis that a setting with multiple equations could yield better results, both in-sample and out-of-sample results suggest there is no improvement from the standard approach. The estimated Accuracy Ratio for the two-equation model is 43.75%, which is slightly worse than the Accuracy Ratio of the single-equation model, 43.77%. The out-of-sample results confirm this tendency, the AR of the two-equation model is 46.07%, against 50.59% of the single-equation model, according to the test results presented in Table 5 none of these differences is statistically significant. Since the two-equation model involves more parameters to estimate and is not able to better discriminate to a significant extent the default and non-default populations of the dataset, the single-equation specification is considered superior in terms of scoring methodology for this dataset. Regarding the hypothesis that balancing the default and non-default populations could help the logistic regression to better discriminate them, again both in-sample and out-of-sample results do not provide positive evidence. The estimated Accuracy Ratio for the weighed model is 43.74%, marginally worse than the 43.77% of the unweighted model. Again, the out-of-sample results confirm that the weighted model does not have a higher discriminating power (AR of 48.29%) than the unweighted model (AR of 50.59%).

As reference, the private-firm model developed by Moody's to the Portuguese market has an in-sample AR of 61.1% (unfortunately no out-of-sample AR is reported)[23]. The selected variables are: *Equity / Total Accounts Payable, Bank Debt / Total Liabilities, Net P&L / Assets, (Ordinary P&L+ Depreciation)/ Interest and similar Expenses, (Ordinary P&L + Depreciation + Provisions) / Total Liabilities, Current Assets / Accounts Payable (due within 1 year)* and *Interest and similar Expenses / Turnover*. The sample data comprised financial statements of 18,137 unique firms, of which 416 had defaulted (using the "90 days past due" definition), with a time span from 1993 to 2000. Hayden (2003) reports an in-sample AR of 50.3% and an out-of-sample AR of 48.8% for a logistic regression model applied to the Austrian market, with the "90 days past due" default definition. The variables selected are *Equity / Assets, Bank Debt / Assets, Current Liabilities / Assets, Accounts*

[23] See Murphy et al. (2002)

Payable / Mat. Costs, Ordinary Business Income / Assets and *Legal Form*. The sample data included 16,797 observations, of which 1,604 were defaults, for a time period ranging from 1992 to 1999. Due to differences in the dataset, such as different levels of data quality or the ratio of default to non-default observations, the reported AR's for both studies presented above cannot be directly comparable to the AR's reported in our study. Despite this fact, they can still be regarded as references that attest the quality of the model presented in terms of discriminatory power.

The following chapter discusses possible applications of the scoring model presented. We start by discussing the creation of a quantitative rating system, followed by the estimation of probabilities of default and rating transition matrixes. Finally the capital requirements for a simulated portfolio are calculated under both the NBCA and current regulations.

5 Applications

5.1 Quantitative Rating System and Probability of Default Estimation

The scoring output provides a quantitative assessment of the credit quality of each firm. Rating classes can be built through a partition of the scoring scale into *k* groups. A default frequency can, in turn, be estimated for each partition, dividing the number of default observations by the total number of observations for each rating class. Furthermore, these default frequencies can be leveled in order to allow for the global default rate of the dataset to be similar to the projected default rate of the universe. These adjusted default frequencies represent the Probability of Default (PD) estimates of the quantitative rating system for each rating class. In light of the NBCA, these can be interpreted as an approximation to the long-run averages of one-year realized default rates for the firms in each rating class[24].

The quantitative rating system presented in this section is not directly comparable to the traditional rating approaches adopted by the rating agencies. The two main differences between the systems are the scope of the analysis and the volatility of the rating classes. Regarding the scope of the analysis, the system developed in this study is concerned with only one risk dimension, the probability of default. Ratings issued by the agencies address not just obligor risk but the facility risk as well. The other major difference is related to the time horizon, the quantitative system has a specific one-year time horizon, with high volatility subject to economic cycle fluctuations. The agencies approach is to produce through-the-cycle ratings, with unspecific, long-term time horizon. Cantor and Packer (1994) provide a description of the rating methodologies for the major rating agencies, while Crouhy et

[24] Basel Committee on Banking Supervision (2003), par. 409.

al. (2001) present the major differences between the internal rating system of a bank and the rating systems of two major credit rating agencies.

Regarding the quantitative rating system, two alternative methodologies are employed in order to obtain the optimal boundaries for each rating class. The goal is for the rating system to be simultaneously stable and discriminatory. A stable rating system is one with infrequent transitions, particularly with few ample transitions[25]. A discriminatory rating system is a granular system with representative and clear distinct classes, in terms of the frequency of default that should increase monotonically from high to low rating classes.

The first methodology employed consists in obtaining coherent rating classes through the use of cluster analysis on the scoring estimates. The second methodology is devised as an optimization problem that attempts to map the historical default frequencies of rating agency whole letter obligor ratings.

5.1.1 Cluster Methodology

Clustering can be described as a grouping procedure that searches for a "natural" structure within a dataset. It has been used thoroughly in a wide range of disciplines as a tool to develop classification schemes. The observations in the sample are reduced to *k* groups in a way that within each group, these observations are as close as possible to each other than to observations in any other group.

K-Means algorithm is implemented due to the large number of observations[26]. In order to determine the optimal number of clusters, the Calinski and Harabasz (1974) method is used. This index has been repeatedly reported in the literature as one of the best selecting procedures (Milligan and Cooper, 1985). The index is calculated as:

[25] An ample transition is a rating upgrade/downgrade involving several rating notches. For example, if a firm has a downgrade from Aaa to Caa in just one period.
[26] Refer to Appendix 6 for a description of the algorithm used.

$$CH(k) = \frac{\sum_{i=1}^{k} n_i (\overline{Y}_i - \overline{Y})^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} n_i (Y_{ij} - \overline{Y}_i)^2 / (n-k)} = \frac{BSS/(k-1)}{WSS/(n-k)}, \quad (7)$$

where *BSS* is the Between Sum-of-Squares; *WSS* the Within Sum-of-Squares; k the number of clusters; n the number of observations; Y_{ij} estimated score for observation j in cluster i.

The optimal k is the one that maximizes the value of *CH(k)*, since it will be at this point that the relative variance between groups respective to the variance within the groups will be higher.

The cluster analysis is performed on the scoring estimates of the three models estimated previously. Table 6 reports the *CH(k)* index for $k = 2$ up to $k = 20$.

k	Model A	Model B	Model C
2	25,092	25,644	28,240
3	30,940	32,046	36,176
4	35,105	36,854	44,639
5	39,411	42,252	50,774
6	43,727	45,889	58,179
7	48,015	51,642	65,751
8	54,666	49,930	72,980
9	55,321	56,529	77,201
10	61,447	62,321	86,546
11	55,297	57,629	93,152
12	62,620	63,913	95,021
13	69,788	71,726	104,821
14	65,603	78,093	110,153
15	73,152	73,530	116,503
16	78,473	75,129	126,060
17	74,141	84,335	129,162
18	**79,710**	82,801	138,090
19	75,293	78,527	**138,461**
20	79,154	**87,544**	134,544

Table 6 – Calinski-Harabasz CH(k) index for k = 2 up to k = 20

The table above reports the Calinski and Harabasz (1974) index for the three alternative specifications using 2 to 20 clusters. The optimal number of clusters is the one that maximizes the index, where the relative variance between groups respective to the variance within the groups will be higher. Model A is the Multiple Industry Equations model, B is the Standard Model, and C the Weighted Sample Model.

For Model A, the optimal number of clusters is 18, for Model B is 20, and for Model C is 19. In order to directly compare the resulting rating systems, classes are

aggregated into $k = 7$[27]. This class aggregation is performed taking in consideration both stability and discriminatory criteria. Figures 19 and 20 present the distribution of the default frequency and of the number of observations by rating class, for each model.

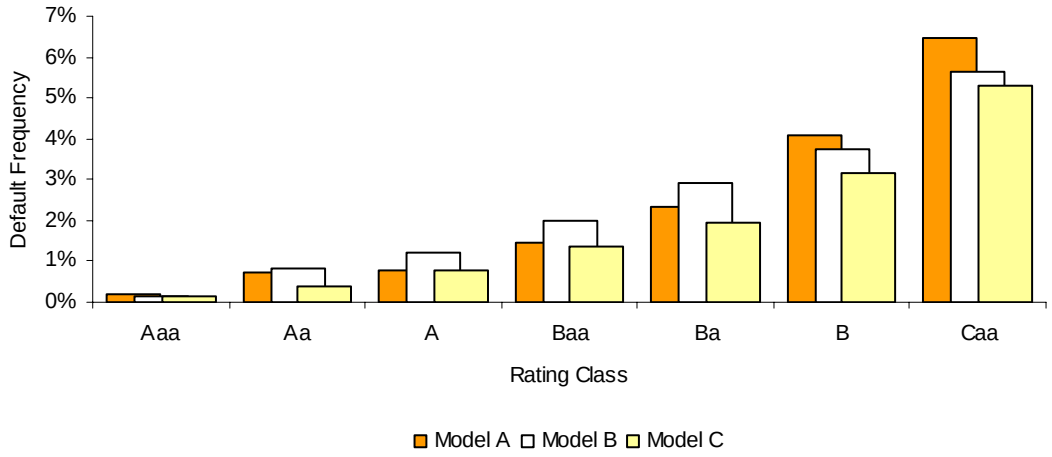

Figure 19 – Default Frequency by Rating Class (Cluster Method)
This figure provides the default frequency for each rating class determined by the cluster methodology. The default frequency corresponds to the proportion of default observations to the total number of observations in each rating class, adjusted by a calibration factor. Model A is the Multiple Industry Equations model, B is the Standard Model, and C the Weighted Sample Model.

Results in Figure 19 are similar across all three models, the default frequency rises from lower to higher risk ratings (only exception being the inflection point for Model A between classes Aa and A), although this rise is only moderate. The defaulted frequencies reported are calibrated frequencies that, as mentioned before, can be interpreted as the actual PD estimates for each rating class. Since the dataset is biased towards the default observations, the resulting default frequencies are leveled so that the overall default ratio would equal 1.5%[28].

[27] $K = 7$ is the minimum number of classes recommended in the NBCA (Basel Committee on Banking Supervision 2003, par. 366) and it is also the number of whole letter rating classes of the major rating agencies.
[28] The calibration value should be similar to the best estimate of the annual default ratio of the universe. For this study, it is estimated that this value should be equal to 1.5% for the non-financial private Portuguese firms.

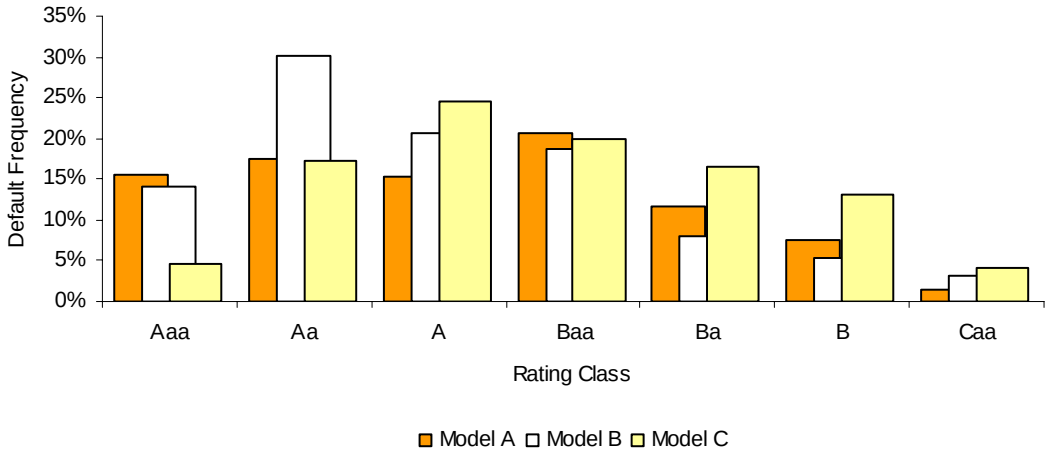

Figure 20 – Number of Observations Distribution by Rating Class (Cluster Method)
The figure above displays the distribution of the number of observations by rating class, with the rating classes determined by the cluster methodology. Model A is the Multiple Industry Equations model, B is the Standard Model, and C the Weighted Sample Model.

Regarding the distribution of observations (Figure 20), it is interesting to observe that the three models that have so far presented very similar results actually produce clearly distinct rating classes. Model A suggests a more uniformly distributed system, with only the lowest rating class having fewer observations. Model B presents a distribution more concentrated on the higher rating classes, while Model C presents a more orthodox distribution, with higher concentration on the middle ratings and lower weight on the extremes.

With the assumptions made, for the cluster methodology, Model B is the one that presents the less attractive rating system: it is not able to better discriminate between rating classes in terms of default frequency to a significant extent, and it assigns very high ratings too often. Models A and C rating systems have a similar discriminating power, although the rating distribution suggested by Model C is the one closer to what should be expected from a balanced portfolio. Thus, the empirical evidence seems to corroborate the hypothesis advanced in section 4.4, the weighting of the sample for the scoring model is helpful in order to identify coherent classes through a cluster methodology.

5.1.2 Historical / Mapping Methodology

The second methodology tested consists on defining the class boundaries in such a way that the resulting default frequencies for each class (after calibration) would approximate as best as possible a chosen benchmark. For this study, the benchmark is Moody's historical one-year default frequencies for corporate whole rating grades. Table 7 provides descriptive statistics for the Moody's ratings[29].

Rating	Min	1st Quartile	Median	Mean	StDev	3rd Quartile	Max
Aaa	0	0	0	0	0	0	0
Aa	0	0	0	0.06	0.18	0	0.83
A	0	0	0	0.09	0.27	0	1.7
Baa	0	0	0	0.27	0.48	0.37	1.97
Ba	0	0	0.64	1.09	1.67	1.29	11.11
B	0	0.38	2.34	3.71	4.3	5.43	20.78
Caa-C	0	0	7.93	13.74	17.18	20.82	100
Investment-Grade	0	0	0	0.15	0.28	0.21	1.55
Speculative-Grade	0	0.59	1.75	2.7	3.04	3.52	15.39
All Corporate	0	0.18	0.67	1.1	1.38	1.32	8.4

Table 7 – Annual Global Issuer-Weighted Default Rate Descriptive Statistics, 1920-2003

It is relevant to point out that this is not an attempt to create an alternative to Moody's ratings. The objective is to obtain a rating system whose default frequencies share some properties with an external reference. A downside of this mapping methodology is that implicitly we assume that our benchmark has the desired properties, and that the underlying structure of our population is similar to the one used to produce the benchmark statistics. The methodology is set up as an optimization problem that can be formalized as follows:

$$\min_{x_1,\ldots,x_{k-1}} \sum_{i=1}^{k} \left(y_i^b - y_i \right)^2, \tag{8}$$

$$\text{subject to } \; y_i = \frac{d_i}{x_i} \;,\; x_i > 0 \;,\; d_i > 0 \;,\; \forall i$$

[29] Source: Hamilton, 2004.

where y_i^b is the default frequency of the benchmark for class i, y_i is the default frequency of the model for class i, d_i is the number of default observations in class i and x_i is the number of observations in class i.

Figures 21 and 22 present the results of applying this methodology to the scoring models estimated previously.

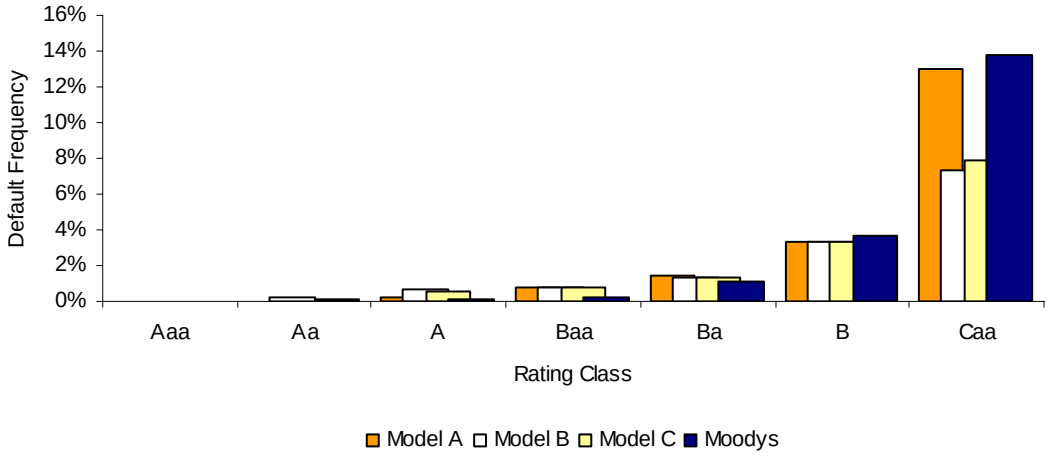

Figure 21 – Default Frequency by Rating Class (Historical Method)
This figure provides the default frequency for each rating class determined by the historical methodology. The default frequency corresponds to the proportion of default observations to the total number of observations in each rating class, adjusted by a calibration factor. Model A is the Multiple Industry Equations model, B is the Standard Model, and C the Weighted Sample Model. The average default rate for Moody's whole letter ratings in the period 1920-2003 is also presented.

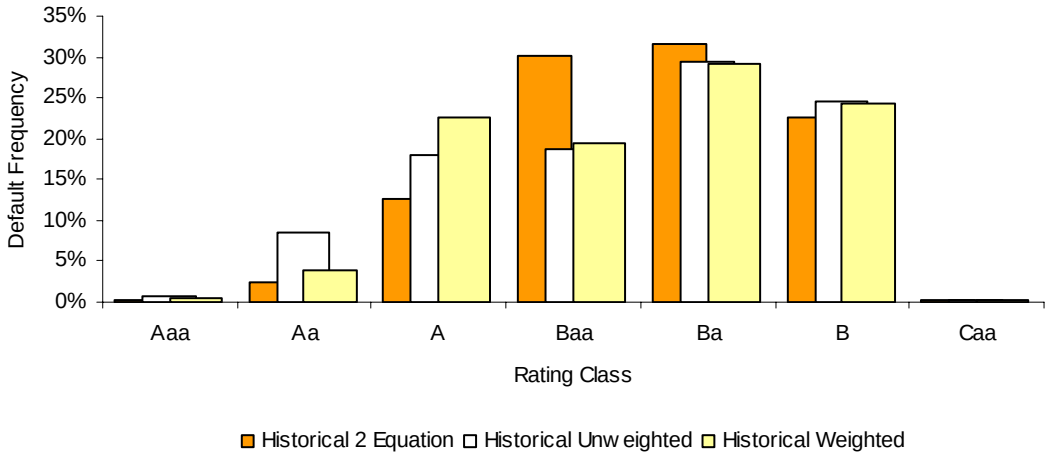

Figure 22 – Number of Observations Distribution by Rating Class (Historical Method)
The figure above displays the distribution of the number of observations by rating class, with the rating classes determined by the historical methodology. Model A is the Multiple Industry Equations model, B is the Standard Model, and C the Weighted Sample Model.

Figure 21 shows the default frequency by rating class for each model and selected benchmark. The default frequency presented are calibrated frequencies, the calibration is similar to the one described in the previous section. All three models can moderately approximate the benchmark, although only results for Model A provide a good fit for the default frequency in the lowest rating class. Even so, the results for the three models are clearly positive in terms of discriminatory power. When comparing the default frequencies between the two methodologies, it is clear that the historical methodology yields much steeper rating scales, starting at lower default rates for the higher rated classes, and ending at clearly higher default rates for the lower rated classes than the cluster methodology[30]. Consequently, the resulting distributions of observations for the rating systems based on the historical methodology (Figure 22, above) are less granular, with higher concentrations in the middle / lower classes. For all three models, only the very best firms belong to one of the two higher rating classes, and the worst class is reserved for the very worst performing firms.

[30] The default rates for the higher rating class, resulting from the historical methodology, are 0% because historically there are no observed one-year defaults for the benchmark, in the period considered.

Comparing the distributions of observations by rating class based on the three scoring models, there are no clear differences between them.

5.1.3 Rating Matrixes and Stability

Once the optimal boundaries for each rating class are determined, a rating classification can be attributed for each observation of the dataset. Tracking the evolution of the yearly observations of each firm enables the construction of one-year transition matrixes. If, for example, a firm is classified as Baa in the fist period considered, in the next period it could either have an upgrade (to Aaa, Aa or A), a downgrade to (Ba, B, Caa), remain at Baa, default, or have no information in the dataset (Without Rating – WR).

The analysis of the transition matrix is helpful in order to study the stability of the rating system. The fewer transitions, i.e., low percentages in the off-diagonal elements of the matrix, the more stable is the rating system. Furthermore, transitions involving jumps of several notches (for example, a transition from Aaa to Caa) are undesirable. Thus, a stable rating system is one whose rating transitions are concentrated in the vicinity of the main diagonal elements of the matrix.

Another relevant aspect of the transition matrix is the transition from each rating class to default. In terms of discriminatory power, a better rating system is one where the transitions to default rise at an exponential rate, from the higher rating to the lower rating classes.

Tables 8 – 10 present the transition matrices for the three models considered, with the class boundaries determined by the cluster methodology, while Tables 11 – 13 present the matrices based on the historical methodology:

	Aaa	Aa	A	Baa	Ba	B	Caa	D	WR
Aaa	41.15%	20.37%	4.71%	1.22%	1.06%	0.65%	0.00%	0.81%	30.03%
Aa	19.13%	29.82%	15.93%	6.11%	3.49%	1.16%	0.07%	2.98%	21.31%
A	7.15%	23.74%	25.84%	12.88%	8.25%	2.78%	0.08%	3.20%	16.08%
Baa	2.31%	14.08%	19.47%	19.25%	17.60%	6.93%	0.33%	5.39%	14.63%
Ba	1.21%	4.93%	10.85%	16.70%	29.69%	15.20%	0.43%	5.92%	15.06%
B	0.37%	1.76%	2.61%	4.27%	18.30%	42.96%	2.40%	11.85%	15.47%
Caa	0.00%	0.95%	0.95%	0.00%	3.81%	34.29%	9.52%	25.71%	24.76%

Table 8 – Model A 1 Year Transition Matrix (Cluster Method)

	Aaa	Aa	A	Baa	Ba	B	Caa	D	WR
Aaa	42.43%	23.57%	1.72%	0.36%	0.27%	0.00%	0.09%	0.63%	30.92%
Aa	13.42%	46.57%	14.09%	3.32%	0.38%	0.21%	0.04%	3.20%	18.76%
A	1.58%	26.52%	32.79%	14.78%	2.19%	0.43%	0.43%	4.81%	16.48%
Baa	0.46%	8.27%	24.22%	33.40%	8.27%	2.99%	1.04%	7.55%	13.80%
Ba	0.14%	2.32%	9.71%	29.28%	21.16%	8.12%	3.77%	10.72%	14.78%
B	0.00%	1.89%	2.95%	14.74%	21.47%	21.05%	7.58%	13.05%	17.26%
Caa	0.00%	1.54%	0.00%	7.34%	11.97%	16.22%	18.15%	21.62%	23.17%

Table 9 – Model B 1 Year Transition Matrix (Cluster Method)

	Aaa	Aa	A	Baa	Ba	B	Caa	D	WR
Aaa	26.67%	28.80%	7.20%	0.80%	0.00%	0.27%	0.00%	0.53%	35.73%
Aa	7.10%	41.27%	19.01%	3.85%	0.74%	0.44%	0.07%	1.63%	25.89%
A	1.15%	18.70%	40.28%	15.40%	3.72%	0.68%	0.10%	3.20%	16.76%
Baa	0.25%	3.84%	23.96%	32.14%	13.96%	2.89%	0.50%	5.41%	17.04%
Ba	0.07%	1.25%	9.42%	24.50%	29.29%	13.10%	1.62%	7.28%	13.47%
B	0.09%	0.35%	2.26%	9.03%	23.00%	32.38%	6.34%	11.37%	15.19%
Caa	0.00%	0.00%	1.15%	0.29%	5.48%	26.51%	23.34%	19.88%	23.34%

Table 10 – Model C 1 Year Transition Matrix (Cluster Method)

	Aaa	Aa	A	Baa	Ba	B	Caa	D	WR
Aaa	13.33%	26.67%	16.67%	6.67%	0.00%	0.00%	0.00%	0.00%	36.67%
Aa	1.44%	14.83%	25.84%	15.31%	0.48%	0.96%	0.00%	0.00%	41.15%
A	0.40%	5.23%	34.97%	26.83%	3.52%	0.60%	0.00%	1.01%	27.44%
Baa	0.04%	0.80%	13.43%	45.59%	16.65%	1.69%	0.00%	2.96%	18.84%
Ba	0.04%	0.16%	1.65%	21.18%	44.64%	11.38%	0.04%	5.69%	15.22%
B	0.00%	0.10%	0.26%	3.64%	22.44%	44.83%	0.26%	12.47%	16.00%
Caa	0.00%	0.00%	0.00%	0.00%	0.00%	22.22%	0.00%	55.56%	22.22%

Table 11 – Model A 1 Year Transition Matrix (Historical Method)

	Aaa	Aa	A	Baa	Ba	B	Caa	D	WR
Aaa	18.97%	36.21%	12.07%	0.00%	0.00%	0.00%	0.00%	0.00%	32.76%
Aa	2.85%	32.43%	24.32%	6.16%	1.05%	0.30%	0.00%	0.75%	32.13%
A	0.07%	12.79%	35.91%	16.80%	6.75%	0.98%	0.00%	2.46%	24.24%
Baa	0.00%	3.44%	22.66%	31.96%	21.35%	2.00%	0.00%	2.96%	15.63%
Ba	0.00%	0.42%	5.65%	17.58%	43.55%	11.80%	0.04%	5.31%	15.64%
B	0.00%	0.05%	0.76%	2.48%	20.88%	47.44%	0.43%	12.18%	15.77%
Caa	0.00%	0.00%	0.00%	0.00%	5.00%	25.00%	0.00%	30.00%	40.00%

Table 12 – Model B 1 Year Transition Matrix (Historical Method)

	Aaa	Aa	A	Baa	Ba	B	Caa	D	WR
Aaa	18.92%	29.73%	16.22%	0.00%	0.00%	0.00%	0.00%	0.00%	35.14%
Aa	2.24%	21.41%	35.14%	3.83%	0.96%	0.32%	0.00%	0.00%	36.10%
A	0.22%	5.17%	46.79%	14.62%	5.29%	0.84%	0.00%	2.08%	24.97%
Baa	0.00%	1.13%	24.27%	33.36%	20.49%	2.06%	0.00%	3.18%	15.52%
Ba	0.00%	0.21%	5.42%	18.13%	43.13%	11.77%	0.04%	5.33%	15.96%
B	0.00%	0.05%	0.81%	2.49%	20.90%	47.51%	0.34%	12.13%	15.77%
Caa	0.00%	0.00%	0.00%	0.00%	0.00%	22.22%	0.00%	38.89%	38.89%

Table 13 – Model C 1 Year Transition Matrix (Historical Method)

Results based on the historical methodology are more stable and display higher discriminatory power than the results based on the cluster methodology. In terms of stability, the historical based results have less high level transitions. For example, none of the three matrixes based on this methodology have transitions from the high classes Aa, A, Baa to lowest class Caa, while all of the three matrices based on the cluster methodology have such transitions.

In terms of discriminatory power, the matrixes based on the historical methodology also present better results, since the transitions to default start at lower percentages for the higher classes and increase continuously to considerable higher percentages than the transitions based on the cluster methodology.

Regarding the results for each model, within each methodology, none of them produces a clearly better rating matrix.

5.2 Regulatory Capital Requirements

Under the New Basel Capital Accord (NBCA), financial institutions will be able to use their internal risk assessments in order to determine the regulatory capital requirements[31]. In the first pillar of the Accord – Minimum Capital Requirements – two broad methodologies for calculating capital requirements for credit risk are proposed. The first, the Standardized Approach, is similar to the current capital accord, where the regulatory capital requirements are independent of the internal assessment of the risk components of the financial institutions. Conversely, in the second methodology – the Internal Ratings-Based Approach – banks complying with certain minimum requirements can rely on internal estimates of risk components in order to determine the capital requirements for a given exposure. Under this methodology, two approaches are available: a Foundation and an Advanced approach. For the Foundation Approach, credit institutions will be able to use their own estimates of the PD but rely on supervisory estimates for the other risk components. For the Advanced Approach, banks will be able to use internal estimates for all risk components, namely the PD, Loss-Given-Defaults (LGD), Exposure-At-Default (EAD) and Maturity (M). These risk components are transformed into Risk Weighted Assets (RWA) through the use of risk weight functions[32].

Up to this point we have devised six alternative methodologies for determining one of the risk components, the PD. Assuming fixed estimates for the other risk components we are able to estimate capital requirements under the IRB Foundation approach, and compare them to the capital requirements under the current accord. The parameters assumed are LGD = 45%, M = 3 years, EAD for SME = 0.3 Million Eur and EAD for large firms = 1.5 Million Eur. The PD used for firm *i* corresponds to the maximum PD estimated for the rating class where *i* belongs. For the calculations under the current capital accord, it is considered that all exposures have the standard risk weight of 100%. Table 14 provides results for all six models.

---

[31] Basel Committee on Banking Supervision (2003).
[32] Appendix 7 provides a description of the formulas used to compute the RWA for corporate exposures.

Rating Methodology	Model	Average RWA %	Capital Requirements, EUR Bal II IRB Fnd	Bal I / Bal II Std	Difference, EUR
Historical	A	90.95%	272,396,629.19	299,496,000.00	27,099,370.81
	B	91.45%	273,885,689.68	299,496,000.00	25,610,310.32
	C	91.59%	274,315,296.00	299,496,000.00	25,180,704.00
Cluster	A	96.24%	288,220,548.59	299,496,000.00	11,275,451.41
	B	91.11%	272,881,107.86	299,496,000.00	26,614,892.14
	C	91.79%	274,906,654.20	299,496,000.00	24,589,345.80

Table 14 – Average RWA and Total Capital Requirements

The table above provides the average risk weighted assets for a standard portfolio using the rating classifications obtained through the three scoring model specifications and for each rating methodology. In addition, capital requirements under the current capital accord and the NBCA (IRB-Foundation) are also calculated and compared. Model A is the Multiple Industry Equations model, B is the Standard Model, and C the Weighted Sample Model.

Results are similar for all models, the capital requirements under the IRB Foundation approach are lower than those that would be required under the current capital accord. For the Historical rating methodology, the two-equation scoring specification (Model A) is the one that provides the highest capital difference, but for the Cluster rating methodology it is the one that provides the lowest.

Figures 29 – 34 in Appendix 8 provide the distribution if the relative RWA for each rating class of all six methodologies, weighted by the number of observations attributed to each class by each rating methodology. The results based on the Historical Methodology are more concentrated on the middle classes, and typically only the two lowest rating classes have a risk weight above the standard Basel I weight. Results under the Cluster Methodology are more evenly spread out through the different classes, with the three up to five lowest rating classes having a risk weight above Basel I requirements.

6 Conclusion

The first and main result from this research is that it is possible to build a relatively simple but powerful and intuitive rating system for privately-held corporate firms, with few data requirements. In order to set up a similar system, it is only necessary to retrieve for a given time frame (at very least 4 years, better would be a full economic cycle) yearly default data and the accounting reports used to concede these loans. This purely quantitative system is enough to provide a scoring rule that, for this dataset, is able to discriminate to a very satisfactory extent the defaulting and non-defaulting populations, both in and out-of-sample. It is also capable of classifying the various firms into meaningful and coherent rating classes. Meaningful in the sense that firms belonging to a certain rating class have distinct probabilities of default from firms belonging to other classes, and to lower ratings correspond significantly higher probabilities of default. Coherent in the sense that rating transitions are stable: if a firm has a given rating for a given year, the probability that in the following period it would be either upgraded or downgraded several notches is very small. Furthermore, the probabilities of default associated to each rating class are calibrated to the estimated real average default frequency of the portfolio, and can therefore be used to access the potential impact of introducing the IRB – Foundation approach of the NBCA, for a given portfolio.

In terms of the scoring methodology, two alternatives to the classical regression are presented. The first alternative is a two-equation specification that allows for industry differentiation. The second is a weighted model that balances the proportion of defaulting and non-defaulting observations. In terms of the discriminating power of the scoring model, both in-sample and out-of-sample results indicate that neither of the two alternative specifications provide significant improvement to the classical regression. However, both alternatives have proven useful later when building the rating classes. The weighted model provides the best results when using a cluster methodology to group individual observations into rating classes, while the two-equation specification provides the most discriminating system

when rating classes are built through a mapping methodology. Comparing the two rating methodologies, the mapping methodology yields more discriminating systems but, on the other hand, the cluster methodology provides more granular rating distributions. Regarding the rating matrixes, the mapping methodology provides more discriminatory power with considerably less ample rating transitions.

There are, however, important extensions to the basic setup that should be considered. The first one derives from the fact that the scoring model only considers a subset of all the variables that can potentially help to discriminate the defaulting and non-defaulting populations. A more complete setup would then consider alternative explanatory variables (such as the reputation of management, the quality of the accounting reports or the relationship of the client to the bank), but more importantly, it should incorporate the subjective opinion or expertise of the credit analyst. A desirable feature of a rating system is giving the possibility for the credit analyst to override the rating decision provided by the mechanical score. This is particularly relevant in the corporate segment, since a wide array of idiosyncrasies (such as creative accounting) could distort the results of the quantitative assessment.

Another potentially useful extension would be to develop a system that provides ratings based not just on the most current available information, but also on the information available on the previous periods. This would result in a more stable system: for a firm to have a very good / bad classification, it would have to present very good / bad indicators for several periods. There is however, a trade-off between stability and discriminatory power: for example, if a firm has in the past produced consistently good indicators, but in the present is rapidly becoming on the verge of bankruptcy, such a system may not downgrade the rating classification of such a firm fast enough.

One final point worth mentioning is that the system developed only provides borrower ratings. In order to use such a system to concede loans, the variables specific to each loan (such as collateral) should be taken in consideration together with the borrower rating. In other words, the final rating assigned to a certain loan is a function of the borrower rating and the Loss-Given-Default (LGD).

Bibliography

Allen, L., 2002, Credit Risk Modeling of Middle Markets, *presented at Conference on Credit Risk Modeling and Decisioning, Wharton FIC, University of Pennsylvania*.

Altman, E., 2000, Predicting Financial Distress Of Companies: Revisiting The Z-Score And Zeta® Models, *Working Paper, Stern School of Business, New York University*.

Altman, E., 1968, Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance*, Sep, 589–609.

Altman, E., G. Marco, and F. Varetto, 1994, Corporate Distress Diagnosis: Comparison Using Linear Discriminant Analysis and Neural Networks (the Italian Experience), *Journal of Banking and Finance* 18, 505-529.

Bamber, D., 1975, The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Graph, *Journal of Mathematical Psychology* 12, 387–415.

Braga, A., 2000, Curvas Roc: Aspectos Funcionais e Aplicações, *PhD Thesis, Universisdade do Minho*.

BarNiv, R. and J. McDonald, 1999, Review of Categorical Models for Classification Issues in Accounting and Finance, *Review of Quantitative Finance and Accounting* 13, 39-62.

Basel Committee on Banking Supervision, 2003, The New Basel Capital Accord, *Bank for International Settlements Consultative Document,* April.

Beaver, W., 1966, Financial ratios as predictors of bankruptcy, *Journal of Accounting Research (Supplement),* 71–102.

Black, F. and J. Cox, 1976, Valuing corporate securities: Some effects of bond indenture provisions, *Journal of Finance,* 351–367.

Black, F. and M. Scholes, 1973, The pricing of options and corporate liabilities, *Journal of Political Economy* 81, 637–54.

Box, G. and P. Tidwell, 1962, Transformation of independent variables, *Technometrics* 4, 531-550.

Boyes, W., D. Hoffman and S. Low, 1989, An econometric analysis of the bank credit scoring problem, *Journal of Econometrics* 40, 3–14.

Caiazza, S., 2004, The comparative performance of credit scoring models: an empirical approach, in M. Bagella, et al., ed.: *Monetary Integration, Markets and Regulation, Research in Banking and Finance* 4 (Elsevier Science Ltd, Oxford).

Calinski, T. and J. Harabasz, 1974, A dendrite method for cluster analysis, *Communications in Statistics* 3, 1–27.

Cantor, R. and F. Packer, 1994, The Credit Rating Industry, *Federal Reserve Bank of New York Quarterly Review,* Summer/Fall.

Cleveland, W., 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* 74, 829-836.

CMVM, 2005, *Relatórios de Auditoria às Contas das Sociedades com Valores Mobiliários Cotados* (Comissão do Mercado de Valores Mobiliários, Lisbon).

Conover, W., 1999, *Practical Nonparametric Statistics* (John Wiley & Sons, New-York).

Crook, J. and J. Banasik, 2004, Does reject inference really improve the performance of application scoring models?, *Journal of Banking & Finance* 28, 857–874.

Crouhy, M., D. Galai and R. Mark, 2001, Prototype risk rating system, *Journal of Banking & Finance* 25, 47-95.

DeLong, E., D. DeLong, and D. Clarke-Pearson, 1988, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, *Biometrics* 44, 837–845.

Duffie, D. and K. Singleton, 1997, An econometric model of the term structure of interest-rate swap yields, *The Journal of Finance* 52(4), 1287–1322.

Dwyer, D., A. Kocagil and R. Stein, 2004, Moody's Kmv Riskcalc™ v3.1 Model, *Moody's KMV Company*.

Engelmann, B., E. Hayden and D. Tasche, 2003, Measuring the Discriminative Power of Rating Systems, *Deutsche Bundesbank Discussion Paper Series 2: Banking and Financial Supervision* 01/2003.

Galindo, J. and P. Tamayo, 2000, Credit risk assesment using statistical and machine learning: basic methodology and risk modelling applications, *Computational Economics* 15, 107-143.

Hamilton, D., 2004, Default and recovery rates of corporate bond issuers, *Moody's Special Comment,* Jannuary.

Hartigan, J., 1975, *Clustering Algorithms* (John Wiley and Sons, New-York).

Hayden, E., 2003, Are Credit Scoring Models Sensitive With Respect to Default Definitions? Evidence from the Austrian Market, *Working Paper, Department of Business Administration, University of Vienna*.

Hosmer, D. and S. Lemeshow, 2000, *Applied Logistic Regression, 2nd Edition* (John Wiley & Sons, New-York).

INE, 2003, *Anuário Estatístico de Portugal 2003* (Instituto Nacional de Estatística, Lisbon).

Jacobson, T. and K. Roszbach, 2003, Bank lending policy, credit scoring and value-at-risk, *Journal of Banking & Finance* 27, 615-633.

Jarrow, R. and S. Turnbull, 1995, Pricing derivatives on financial securities subject to credit risk, *Journal of Finance* 50, 53–85.

Kraft, H., G. Kroisandt and M. Müller, 2004, Redesigning Ratings: Assessing the Discriminatory Power of Credit Scores under Censoring, *Working Paper, Fraunhofer Institut für Techno- und Wirtschaftsmathematik*.

Laitinen, E., 1999, Predicting a corporate credit analyst's risk estimate by logistic and linear models, *International Review of Financial Analysis* 8, 97-121.

Longstaff, F. and E. Schwartz, 1993, A simple approach to valuing risky fixed and floating rate debt, *Working Paper* 22–93*, Anderson Graduate School of Management, University of California*.

Mann, H. and D. Whitney, 1947, On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics* 18, 50–60.

Menard, S., 2002, *Applied logistic regression analysis, 2nd Edition* (Sage Publications, Thousand Oaks, CA).

Merton, R., 1974, On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, *Journal of Finance* 29, 449-470.

Metz, C., 1978, Basic Principles of ROC Analysis, *Seminars in Nuclear Medicine* VIII (4), 283–298.

Murphy, A., A. Kocagil, P. Escott and F. Glormann, 2002, Moody's RiskCalc™ For Private Companies: Portugal, *Moody's Investors Service Rating Methodology*.

Milligan, G. and M. Cooper, 1985, An examination of procedures for determining the number of clusters in a data set, *Psychometrika* 50, 159-179.

Ohlson, J., 1980, Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research* 18, 109–131.

Platt, H. and M. Platt, 1990, Development of a class of stable predictive variables: the case of bankruptcy prediction, *Journal of Business Finance and Accounting* 17, 31-51.

Royston, P. and D. Altman, 1994, Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling, *Applied Statistics* 43, 429–467.

Sobehart, J. and S. Keenan, 2001, Measuring default accurately, *Risk – Credit Risk Special Report,* March, S31–S33.

Yang, Z., M. Platt and H. Platt, 1999, Probabilistic Neural Networks in Bankruptcy Prediction, *Journal of Business Research,* Feb, 67-74.

Appendix 1 – Description of Financial Ratios and Accuracy Ratios

Type	Name	Definition	Expected Effect on PD	AR
Profitability	**Net P&L / Assets**	[Net Profit & Loss] / [Total Assets]	-	15.60%
	Current Earnings / Assets	[Current Earnings] / [Total Assets]	-	16.60%
	Current Earnings and Depreciation / Turnover	[Current Earnings + Depreciation] / [Turnover]	-	-1.00%
	P&L / Assets	[Net Profit & Loss + Depreciation + Provisions] / [Total Assets]	-	5.80%
	Gross Earnings / Production	[EBT + Depreciation + Provisions] / [Production]	-	-3.20%
	EBITDA / Production	[EBITDA] / [Production]	-	-15.80%
Liquidity	**Liquidity / Current Liabilities**	[Bank Deposits & Cash + Marketable Securities] / [Short-Term Liabilities]	-	10.60%
	Current Ratio	[Current Assets] / [Short-Term Liabilities]	-	9.20%
	Liquidity / Assets	[Bank Deposits & Cash + Marketable Securities] / [Total Assets]	-	12.00%
Leverage / Gearing	**Equity / Assets**	[Equity] / [Total Assets]	-	3.80%
	Equity / Accounts Payable	[Equity] / [Accounts Payable]	-	4.40%
	Bank Debt / Accounts Payable	[Bank Debt] / [Accounts Payable]	+	1.80%
	Accounts Payable / Assets	[Accounts Payable] / [Total Assets]	+	5.20%
	Liabilities / Assets	[Total Liabilities] / [Total Assets]	+	3.80%
	Net Current Accounts Payable / Assets	[Short-Term Accounts Payable - Bank Deposits & Cash] / [Total Assets]	+	-0.40%
Debt Coverage	**Gross Earnings / Liabilities**	[Current Earnings + Depreciation + Provisions] / [Total Liabilities]	-	9.80%
	Debt Service Coverage	[Current Earnings + Depreciation] / [Interest & Similar Costs]	-	25.20%
	P&L / L-T Liabilities	[Net Profit & Loss + Depreciation + Provisions] / [Long-Term Liabilities]	-	0.00%
	Operating Earnings / Debt Service	[Operating Earnings] / [Interest & Similar Costs]	-	16.60%
Activity	**Interest Costs / Sales**	[Interest & Similar Costs] / [Turnover]	+	40.20%
	Inventories / Turnover	[Inventories] / [Turnover]	+	12.00%
	Turnover / Assets	[Turnover] / [Total Assets]	-	20.40%
Productivity	**Productivity Ratio**	[Personnel Costs] / [Turnover]	+	15.80%

Note: Turnover = Total Sales + Services Rendered

Appendix 2 – Estimating and Comparing the Area Under the ROC curves

The estimated ROC curve and, consequently, the AUROC are outcomes of random variables, since we only have one sample of the scoring of the borrowers and their realized defaults. Following DeLong et al. (1988), the area under the population ROC curve can be defined as the probability that, when the estimated scoring is observed for a randomly selected borrower from the default population and a randomly selected borrower from the non-default population, the resulting scores will be in the correct order (the scoring of the default observation is higher than the scoring of the non-default observation). For a given sample, the AUROC can be estimated either through parametric or nonparametric methods. A parametric approach would involve distributional assumptions on the observed variable, although these distributions cannot be uniquely determined from the ROC curve (see, for example, the binormal model used in Metz 1978). The nonparametric approach used in this study relates the estimation of the AUROC to the Mann-Whitney (1947) U-statistic[33]. Let d_i ($i = 1,\ldots, m$) be the estimated scores for the default observations and r_j ($j = 1,\ldots, n$) be the estimated scores for the non-default observations. An unbiased estimator of the probability of correctly classifying two randomly chosen subjects from the default and non-default populations is given by the average over a kernel ψ:

$$\widehat{AUROC} = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\left(d_i, r_j\right)$$

where,

$$\psi\left(d,r\right) = \begin{cases} 1 & d > r \\ ½ & d = r \\ 0 & d < r \end{cases}$$

[33] Bamber (1975) originally developed this relationship. For more details see, for example, Braga (2000).

The variance of this estimator can be computed through the use of placement values. Let $V(d_i)$ be the placement of the estimated score d_i in the distribution of r scores (i.e., the fraction of r scores that it exceeds). In addition, let $V(r_j)$ be the placement of the estimated score r_j in the distribution of d scores:

$$V\left(d_i\right)=\frac{\sum_{j=1}^{n}\psi\left(d_i,r_j\right)}{n} \qquad \text{and} \qquad V\left(r_j\right)=\frac{\sum_{i=1}^{m}\psi\left(d_i,r_j\right)}{m}$$

The variance of the estimator for large samples can then be computed as the sum of the scaled variances for the placement values of d and r:

$$\operatorname{var}\left(\widehat{AUROC}\right)=\frac{m\sum_{j=1}^{m}V\left(r_j\right)^2-\left[\sum_{j=1}^{m}V\left(r_j\right)\right]^2}{m^2\left(m-1\right)}+\frac{n\sum_{i=1}^{n}V\left(d_i\right)^2-\left[\sum_{i=1}^{n}V\left(d_i\right)\right]^2}{n^2\left(n-1\right)}$$

If we wish to build a test to compare the AUROC estimates for two alternative models, A and B based on the same dataset it is also relevant to compute the covariance of the estimates:

$$\operatorname{cov}\left(\widehat{AUROC_A},\widehat{AUROC_B}\right)=\frac{\sum_{j=1}^{m}\left[V\left(r_{Aj}\right)-\overline{V}\left(r_A\right)\right]\left[V\left(r_{Bj}\right)-\overline{V}\left(r_B\right)\right]}{m\left(m-1\right)}+$$
$$+\frac{\sum_{i=1}^{n}\left[V\left(d_{Ai}\right)-\overline{V}\left(d_A\right)\right]\left[V\left(d_{Bi}\right)-\overline{V}\left(d_B\right)\right]}{n\left(n-1\right)}$$

The test statistic for testing $H_0:\widehat{AUROC_A}=\widehat{AUROC_B}$ is given by:

$$T=\frac{\left(\widehat{AUROC_A}-\widehat{AUROC_B}\right)^2}{\operatorname{var}\left(\widehat{AUROC_A}-\widehat{AUROC_B}\right)}$$

where,

$$\operatorname{var}\left(\widehat{AUROC_A}-\widehat{AUROC_B}\right)=\operatorname{var}\left(\widehat{AUROC_A}\right)+\operatorname{var}\left(\widehat{AUROC_B}\right)-2\operatorname{cov}\left(\widehat{AUROC_A},\widehat{AUROC_B}\right)$$

The test statistic T is asymptotically χ^2-distributed with one degree of freedom.

Appendix 3 – Binomial Logistic Regression Estimation and Diagnostics[34]

a) Binomial Logistic Regression

Binomial (or binary) logistic regression is a type of regression useful to model relationships where the dependent variable is dichotomous (only assumes two values) and the independent variables are of any type. Logistic regression estimates the probability of a certain event occurring, since it applies maximum likelihood estimation after transforming the dependent variable into a logit variable (the natural log of the odds of the dependent occurring or not). Unlike OLS regression, it estimates changes in the log odds of the dependent variable, not changes in the dependent itself.

Let y_i be a binary discrete variable that indicates whether firm *i* has defaulted or not in a given period of time, and let x_i^k represent the values of the *k* explanatory variables for the firm *i*. The conditional probability that firm *i* defaults is given by $P\left(y_i = 1 \mid x_i^k\right) = \pi\left(x_i^k\right)$, while the conditional probability that the firm does not default is given by $P\left(y_i = 0 \mid x_i^k\right) = 1 - \pi\left(x_i^k\right)$. Thus, the odds that this firm defaults is simply: $odds_i = \pi\left(x_i^k\right) \Big/ 1 - \pi\left(x_i^k\right)$. The estimated regression relates a combination of the independent variables to the natural log of the odds of the dependent outcome occurring:

$$g\left(x, \beta\right) = \ln\left[\frac{\pi\left(x\right)}{1-\pi\left(x\right)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

or,

[34] This Appendix is based on Menard (2002) and Hosmer and Lemeshow (2000).

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}$$

Assumptions

i. Each y_i follows a Bernoulli distribution with parameter $\pi\left(x_i^k\right)$. Which is equivalent to saying that each y_i follows a Binomial distribution with 1 trial and parameter $\pi\left(x_i^k\right)$;
ii. The error terms are independent;
iii. No relevant variables are omitted, no irrelevant variables are included, and the functional form is correct;
iv. There is a linear relationship between the logit of the independent variables and the dependent;
v. There is no significant correlation between the independent variables (no multicolinearity).

Estimation

Estimation of the binomial logistic regression is made through the maximum likelihood methodology. The expression of the likelihood function of a single observation is given by:

$$l_i = \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

Since independence between the observations is assumed, the likelihood function will be the product of all individual likelihoods:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

The log-likelihood function to be maximized will be:

$$L(\beta) = \ln\left[l(\beta)\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\}$$

The ML estimators correspond to the values of β that maximize the previous expression.

b) Residual Analysis

For the logistic regression, the residuals in terms of probabilities are given by the difference between the observed and predicted probabilities that default occurs:

$$e_i = P(y_i = 1) - \hat{P}(y_i = 1) = \pi(x_i) - \hat{\pi}(x_i)$$

Since these errors are not independent of the conditional mean of *y*, it is useful to adjust them by their standard errors, obtaining the Pearson or Standardized residuals:

$$r_i = \frac{\pi(x_i) - \hat{\pi}(x_i)}{\sqrt{\hat{\pi}(x_i)\left[1 - \hat{\pi}(x_i)\right]}}$$

These standardized residuals follow an asymptotically standard normal distribution. Cases that have a very high absolute value are cases for which the model fits poorly and should be inspected.

In order to detect cases that may have a large influence on the estimated parameters of the regression, both the Studentized residuals and the Dbeta statistic are used. The studentized residual corresponds to the square root of the change in the -2 Log-Likelihood of the model attributable to deleting the case from the analysis:

$$s_i = \sqrt{d_i^2 - \frac{r_i^2 h_i}{1 - h_i}}$$

The dbeta is an indicator of the standardized change in the regression estimates obtained by deleting an individual observation:

$$dbeta_i = \frac{r_i^2 h_i}{(1 - h_i)^2}$$

In the previous two expressions, h_i corresponds to the leverage statistic and d_i to the deviance residual. The leverage statistic is derived from the regression that expresses the predicted value of the dependent variable for case *i* as a function of the observed values of the dependent for all cases (for more information see Hosmer and

Lemeshow (2000), 168-171). The deviance residual corresponds to the contribution of each case to the -2 Log-Likelihood function (the deviance of the regression).

c) Testing Coefficient Significance: the Wald Chi-Square Test

For the purpose of testing the statistical significance of the individual coefficients, the Wald Chi-Square test is implemented. Under the hypothesis that $\beta_i = 0$, the test statistic bellow follows a chi-square distribution with one degree of freedom:

$$W_i = \frac{\hat{\beta}_i^2}{\widehat{SE}\left(\hat{\beta}_i\right)^2}$$

d) Testing Regression Significance: the Hosmer & Lemeshow Test

In order to evaluate how effectively the estimated model describes the dependent variable the Hosmer & Lemeshow goodness-of-fit test is applied. The test consists in dividing the ranked predicted probabilities into deciles ($g=10$ groups) and then computing a Pearson chi-square statistic that compares the predicted to the observed frequencies in a 2x10 contingency table. Let o_i^0 be the observed count of non-defaults for group *i* and p_i^0 be the predicted count. Similarly, let o_i^1 be the observed count of defaults for group *i* and p_i^1 be the predicted count. Then the *HL* test statistic following a chi-square distribution with *g-2* degrees of freedom is:

$$HL = \sum_{i=1}^{g}\left[\frac{\left(o_i^0 - p_i^0\right)^2}{p_i^0} + \frac{\left(o_i^1 - p_i^1\right)^2}{p_i^1}\right]$$

Lower values of *HL*, and non-significance indicate a good fit to the data and, therefore, good overall model fit.

e) Testing for Non-Linear Relationships: the Box-Tidwell Test

If the assumption of linearity in the logit is violated, then logistic regression will underestimate the degree of relationship of the independents to the dependent and will lack power, thus generating Type II errors (assuming no relationship when there actually is). A simple method to investigate significant non-linear relationships is the Box-Tidwell (1962) Transformation Test. It consists on adding to the logistic model interaction terms corresponding to the cross-product of each independent variable with its natural logarithm *(x)ln(x)*. If any of these terms are significant, then there is evidence of nonlinearity in the logit. This procedure does not provide the type of nonlinearity, thus if present further investigation is necessary.

f) Fitting Non-Linear Logistic Regressions: the Fractional Polynomial Methodology

Whenever evidence of significant non-linear relationship between a given independent variable and the logit of the dependent is detected, the Fractional Polynomial methodology (Royston and Altman, 1994) is implemented, in order to detect the best non-linear functional form that describes the relationship. Instead of trying to directly estimate a general model, where the power parameters of the non-linear relationship is estimated simultaneously with the coefficients of the independents, this methodology searches for the best functional form from a given set of possible solutions.

As presented before, our logistic regression expression is given by:

$$g(x,\beta) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

For this study, only one of the independent variables had a potentially non-linear relationship with the logit, let this variable be represented by x_k. In order to accommodate the non-linear relationship, the logistic regression expression could be generalized to:

$$g(x,\beta) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \sum_{j=1}^{J} \beta_{j+k-1} H_j(x_k)$$

where, for $j = 1,\ldots,J$:

$$H_j(x_k) = \begin{cases} x_k^{p_j} & \text{if } p_j \neq p_{j-1} \\ H_{j-1}(x_k)\ln(x_k) & \text{if } p_j = p_{j-1} \end{cases}$$

Under this setting, p represents the power and j the number of polynomial functions. For example, a quadratic relationship would have $J=2$, $p_1=1$ and $p_2=2$:

$$g(x,\beta) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \beta_k x_k + \beta_{k+1} x_k^2$$

In practice, as suggested by Royston and Altman (1994), it is sufficient to restrict J to 2 and p to the set $\Omega = \{-2,-1,-0.5,0,0.5,1,2,3\}$, where $p=0$ denotes the natural log of the variable. The methodology is implemented through the following steps:

i. Estimate the linear model;

ii. Estimate the general model with $J=1$ and $p \in \Omega$, and select the best $J=1$ model (the one with lower deviance);

iii. Estimate the general model with $J=2$ and $p \in \Omega$, and select the best $J=2$ model;

iv. Compare the linear model with the best $J=1$ and the best $J=2$ models. This comparison is made through a likelihood ratio test, asymptotically chi-square distributed. The degrees of freedom in the test increases by 2 for each additional term in the fractional polynomial, one degree for the power, and another for the extra coefficient. The selected model is the one that represents a significant better fit than that of next lower degree, but not a significant worse fit than that of next higher degree;

v. Graphically examine the fit estimated by the model selected in the previous stage, in order to validate the economic intuition of the non-linear relationship suggested by the model. This is achieved by comparing the lowess[35] function of the relationship between the dependent and the

[35] The Lowess is the Locally Weighted Scatterplot Smoothing (Cleveland 1979) between two variables. Since the dependent is a binary variable, it is convenient to use this smoothed function to be able to graphically access the relationship in question.

independent variable in question, and the multivariable adjusted function that results from the model selected in the previous stage.

g) Testing for Multicolinearity: the Tolerance Statistic

As for linear regression, high colinearity between the independent variables in a logistic regression results in loss of efficiency, with unreasonably high estimated coefficients and large associated standard errors. Detection of multicolinearity can be made through the use of the Tolerance statistic, defined as the variance of each independent variable that is not explained by all of the other independent variables. For the independent variable X_i, the tolerance statistic equals $1 - R^2_{X_i}$, where $R^2_{X_i}$ is the R^2 of a linear regression using variable X_i as the dependent variable and all the remaining independents as predictors.

If the value of the statistic for a given independent is close to 0, it indicates that the information the variable provides can be expressed as a linear combination of the other independent variables. As a rule of thumb, only tolerance values lower than 0.2 are cause for concern.

Appendix 4 – Estimation Results

Linear Regressions General Results

Regression	Nº Obs	Obs Y=0	Obs Y=1	Deviance	Hosmer & Lemeshow χ^2	df	P-Value	AUROC
A - 2 Eq. Model / Sectors 1 & 2	5,044	4,819	225	1,696	8.74	8	36.51%	71.30%
A - 2 Eq. Model / Sector 3	5,951	5,706	245	1,928	6.79	8	55.89%	
B - Standard Model	10,995	10,525	470	3,626	7.07	8	52.94%	71.28%
C - Weighted Model	1,420	950	470	1,623	11.73	8	16.38%	71.44%

Linear Regressions Estimated Coefficients

Variable	A - 2 Eq. Model / Sectors 1 & 2				A - 2 Eq. Model / Sector 3				B - Unweighted Model				C - Weighted Model			
	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value
R7	-0.39246	0.12878	9.29	0.2307%	-	-	-	-	-	-	-	-	-	-	-	-
R8	-	-	-	-	-0.19705	0.07590	6.74	0.9427%	-0.16455	0.05230	9.90	0.1653%	-0.18762	0.06564	8.17	0.4258%
R9	-	-	-	-	-0.18184	0.08514	4.56	3.2691%	-0.22849	0.06887	11.01	0.0907%	-0.23442	0.08127	8.32	0.3923%
R17	-0.28779	0.09241	9.70	0.1843%	-0.24115	0.08659	7.76	0.5356%	-0.28909	0.06361	20.66	0.0005%	-0.26327	0.07845	11.26	0.0791%
R20	0.46940	0.06164	58.00	0.0000%	0.45161	0.05664	63.57	0.0000%	0.44002	0.04283	105.55	0.0000%	0.50697	0.06564	59.66	0.0000%
R23	0.23328	0.06380	13.37	0.0255%	-	-	-	-	0.15280	0.04436	11.86	0.0572%	0.15948	0.06234	6.54	1.0520%
K	-3.35998	0.08676	1,499.67	0.0000%	-3.33521	0.07658	1,896.73	0.0000%	-3.33613	0.05688	3,440.16	0.0000%	-0.94820	0.06586	207.30	0.0000%

Box-Tidwell Final Backward Stepwise Regression Coefficients

Variable	A - 2 Eq. Model / Sectors 1 & 2				A - 2 Eq. Model / Sector 3				B - Unweighted Model				C - Weighted Model			
	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value
R7	-0.38011	0.12830	8.78	0.3049%	-	-	-	-	-	-	-	-	-	-	-	-
R8	-	-	-	-	-0.21276	0.07622	7.79	0.5247%	-0.17143	0.05241	10.70	0.1073%	-0.19597	0.06552	8.95	0.2782%
R9	-	-	-	-	-0.15921	0.08632	3.40	6.5114%	-0.21020	0.06940	9.17	0.2454%	-0.22388	0.08196	7.46	0.6301%
R17	-0.22552	0.09719	5.38	2.0317%	-0.18249	0.09026	4.09	4.3184%	-0.23063	0.06677	11.93	0.0552%	-0.20282	0.08150	6.19	1.2824%
R20	1.68533	0.36083	21.82	0.0003%	1.58508	0.31588	25.18	0.0001%	1.57265	0.23829	43.56	0.0000%	1.57769	0.29246	29.10	0.0000%
R23	0.19889	0.06597	9.09	0.2570%	-	-	-	-	0.12254	0.04590	7.13	0.7586%	0.12243	0.06302	3.77	5.2037%
BT20*	-0.66208	0.19297	11.77	0.0601%	-0.63780	0.17459	13.34	0.0259%	-0.62538	0.12917	23.44	0.0001%	-0.62506	0.16384	14.55	0.0136%
K	-2.96198	0.13987	448.47	0.0000%	-2.91367	0.13336	477.33	0.0000%	-2.93971	0.09630	931.87	0.0000%	-0.53335	0.12495	18.22	0.0020%

R7 = Liquidity / Current Liabilities; R8 = Current Ratio; R9 = Liquidity / Assets; R17 = Debt Service Coverage; R20 = Interest Costs / Sales; BT20 = R20*LN(R20); R23 = Productivity Ratio;

Fractional Polynomial Model Comparisons (Best J=1,2,3 Models)

R20	df	A - 2 Eq. Model / Sectors 1 & 2				A - 2 Eq. Model / Sector 3				B - Unweighted Model				C - Weighted Model			
		Deviance	Gain	P-Value	Powers	Deviance	Gain	P-Value	Powers	Deviance	Gain	P-Value	Powers	Deviance	Gain	P-Value	Powers
Not in model	0	1750.177	-	-	-	1986.467	-	-	-	3724.482	-	-	-	1687.181	-	-	-
Linear	1	1696.043	0.000	0.000	1	1927.857	0.000	0.000	1	3626.025	0.000	0.000	1	1623.173	0	0.000	1
J = 1	2	1684.842	11.201	0.001	0	1915.064	12.793	0.000	0	3603.782	22.243	0.000	0	1610.633	12.54	0.000	0
J = 2	4	1682.437	13.605	0.301	.5 3	1913.080	14.778	0.371	1 1	3599.921	26.105	0.145	.5 3	1608.129	15.044	0.286	.5 3
J = 3	6	1681.540	14.503	0.639	-1 2 2	1911.768	16.089	0.519	2 3 3	3599.042	26.983	0.644	-1 1 2	1607.349	15.824	0.677	-1 1 2

Reported Deviances for Fractional Polynomial Search

Model #	Power 1	Power 2	Deviance			
			Model A1	Model A2	Model B	Model C
1	-2	-	1750.175	1986.353	3724.480	1687.179
2	-1	-	1699.910	1937.404	3636.893	1633.960
3	-0.5	-	1689.693	1922.565	3614.541	1618.907
4	0	-	1684.842	1915.064	3603.782	1610.633
5	0.5	-	1687.719	1918.091	3609.449	1613.104
6	1	-	1696.043	1927.857	3626.025	1623.173
7	2	-	1715.213	1949.074	3662.488	1646.956
8	3	-	1728.820	1962.952	3686.848	1663.446
9	-2	-2	1750.175	1986.353	3724.480	1687.179
10	-1	-2	1699.911	1937.404	3724.480	1687.179
11	-0.5	-2	1689.694	1922.565	3614.542	1618.908
12	0	-2	1684.842	1915.066	3603.784	1610.634
13	0.5	-2	1687.718	1918.071	3609.449	1613.103
14	1	-2	1696.040	1927.808	3626.023	1623.170
15	2	-2	1715.210	1948.992	3662.485	1646.953
16	3	-2	1728.817	1962.857	3686.846	1663.444
17	-1	-1	1750.175	1935.171	3724.480	1687.179
18	-0.5	-1	1689.685	1922.555	3614.528	1618.898
19	0	-1	1684.842	1915.064	3603.782	1610.633
20	0.5	-1	1685.583	1916.271	3605.742	1611.041
21	1	-1	1687.582	1919.556	3610.443	1613.928
22	2	-1	1692.009	1925.960	3620.050	1620.917
23	3	-1	1695.230	1930.067	3626.581	1626.092

Reported Deviances for Fractional Polynomial Search (Cont.)						
Model #	Power 1	Power 2	Deviance			
			Model A1	Model A2	Model B	Model C
24	-0.5	-0.5	1688.517	1920.858	3612.048	1617.124
25	0	-0.5	1684.839	1915.060	3603.776	1610.627
26	0.5	-0.5	1685.272	1915.696	3604.853	1610.790
27	1	-0.5	1686.189	1917.169	3606.940	1612.193
28	2	-0.5	1687.903	1919.588	3610.549	1615.131
29	3	-0.5	1688.928	1920.884	3612.589	1617.001
30	0	0	1684.776	1914.838	3603.591	1610.240
31	0.5	0	1684.827	1914.977	3603.738	1610.376
32	1	0	1684.838	1915.058	3603.778	1610.541
33	2	0	1684.661	1914.992	3603.492	1610.619
34	3	0	1684.353	1914.867	3603.072	1610.458
35	0.5	0.5	1684.297	1914.262	3602.552	1609.827
36	1	0.5	1683.755	1913.681	3601.482	1609.245
37	2	0.5	1682.890	1913.159	3600.148	1608.379
38	3	0.5	1682.437	1913.385	3599.921	1608.129
39	1	1	1683.014	1913.080	3600.178	1608.436
40	2	1	1682.485	1913.609	3600.057	1608.280
41	3	1	1682.816	1915.364	3601.938	1609.432
42	2	2	1684.157	1917.984	3605.421	1611.925
43	3	2	1687.085	1923.590	3613.244	1617.255
44	3	3	1692.467	1932.259	3626.168	1626.117

Non-Linear Regressions General Results								
Regression	Nº Obs	Obs Y=0	Obs Y=1	Deviance	Hosmer & Lemeshow			AUROC
					χ^2	df	P-Value	
A - 2 Eq. Model / Sectors 1 & 2	5,044	4,819	225	1,682	8.20	8	41.46%	71.88%
A - 2 Eq. Model / Sector 3	5,951	5,706	245	1,913	6.29	8	61.49%	
B - Standard Model	10,995	10,525	470	3,600	2.23	8	97.32%	71.88%
C - Weighted Model	1,420	950	470	1,608	7.68	8	46.53%	71.87%

Non-Linear Regressions Estimated Coefficients

Variable	A - 2 Eq. Model / Sectors 1 & 2				A - 2 Eq. Model / Sector 3				B - Unweighted Model				C - Weighted Model			
	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value	β^	σ^	Wald	P-Value
R7	-0.38053	0.12831	8.80	0.3020%	-	-	-	-	-	-	-	-	-	-	-	-
R8	-	-	-	-	-0.21229	0.07617	7.77	0.5321%	-0.17136	0.05241	10.69	0.1078%	-0.19728	0.06560	9.04	0.2637%
R9	-	-	-	-	-0.16045	0.08631	3.46	6.3017%	-0.21111	0.06940	9.25	0.2353%	-0.22341	0.08196	7.43	0.6414%
R17	-0.22465	0.09710	5.35	2.0686%	-0.18418	0.09013	4.18	4.1003%	-0.23136	0.06668	12.04	0.0521%	-0.20304	0.08142	6.22	1.2638%
R23	0.20007	0.06590	9.22	0.2398%	-	-	-	-	0.12378	0.04587	7.28	0.6964%	0.12343	0.06299	3.84	5.0039%
R20_1	2.01146	0.31598	40.52	0.0000%	1.79215	0.27152	43.56	0.0000%	1.84306	0.21015	76.92	0.0000%	1.87907	0.26051	52.03	0.0000%
R20_2	-0.00933	0.00424	4.83	2.7966%	-0.00873	0.00421	4.30	3.8206%	-0.00876	0.00297	8.72	0.3145%	-0.00907	0.00400	5.13	2.3451%
K	-3.25891	0.08887	1,344.58	0.0000%	-3.42640	0.08329	1,692.28	0.0000%	-3.24970	0.05921	3,012.06	0.0000%	-0.84100	0.07034	142.94	0.0000%

R7 = Liquidity / Current Liabilities; R8 = Current Ratio; R9 = Liquidity / Assets; R17 = Debt Service Coverage; R20 = Interest Costs / Sales; R23 = Productivity Ratio;

Multicolinearity Test

Unweighted Reg.		Weighted Reg.		Sectors 1&2 Reg.		Sector 3 Reg.		Sector 3 Reg.	
Variable	Tolerance	Variable	Tolerance	Variable	Tolerance	Variable	Tolerance	Variable	Tolerance
R8	0.989	R8	0.988	R7	0.989	R8	0.9880	R8	0.9878
R9	0.964	R9	0.963	R17	0.763	R9	0.9700	R9	0.9685
R17	0.762	R17	0.722	R23	0.868	R17	0.8130	R17	0.8128
R23	0.854	R23	0.853	R20_1	0.379	R20_1	0.4200	R20_1	0.0646
R20_1	0.375	R20_1	0.336	R20_2	0.477	R20_2	0.4890	R20_2	0.0685
R20_2	0.477	R20_2	0.440	-	-	-	-	-	-
Model # 38		Model # 38		Model # 38		Model # 38		Model # 39	

Appendix 5 – Kolmogorov-Smirnov and Error Type Analysis

This section provides both the Kolmogorov-Smirnov (KS) analysis and Error Type curves for models A – Multiple Industry Equations, Model B – Single Equation, Unweighted Sample and C – Weighted Sample. The KS analysis consists on evaluating for each possible cut-off point the distance between the Type I Error curve and the True Prediction curve. The higher the distance between the curves, the better the discriminating power of the model. The Error Type curves display for each cut-off point the percentages of Type I (incorrectly classifying an observation as non-default) and Type II (incorrectly classifying an observation as default) errors for each model.

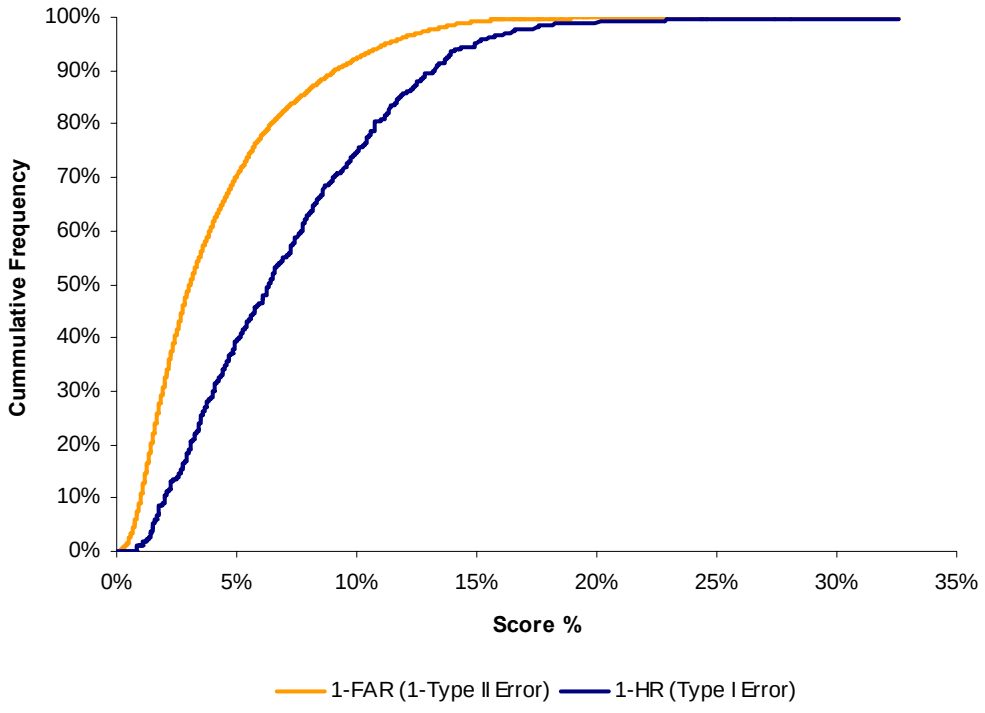

Figure 23 – Model A: Kolmogorov-Smirnov Analysis

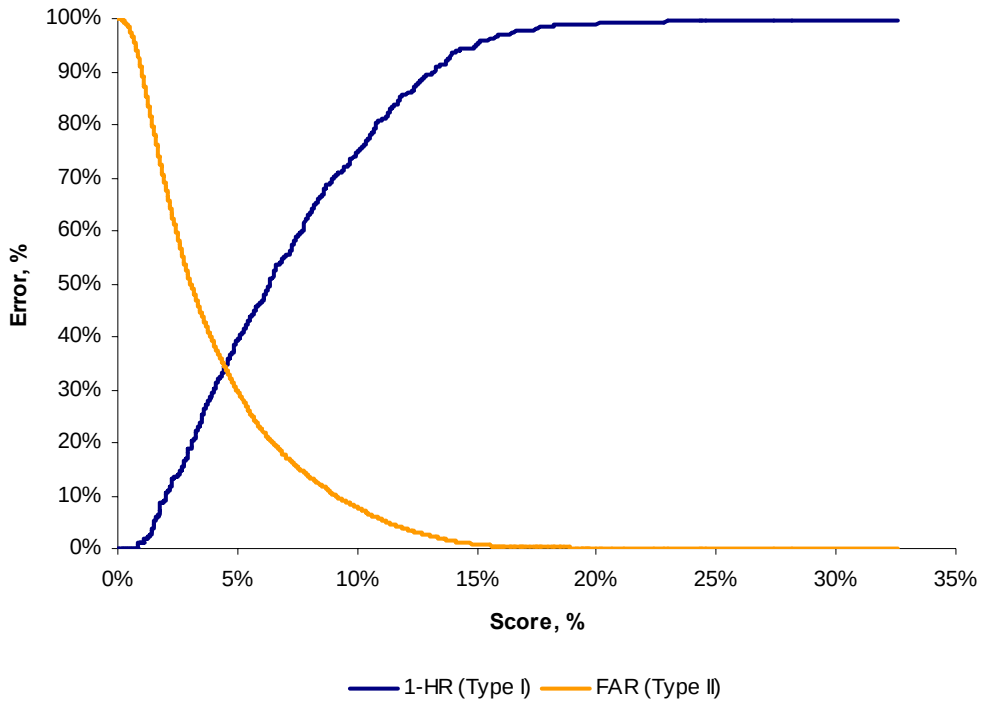

Figure 24 – Model A: Types I & II Errors

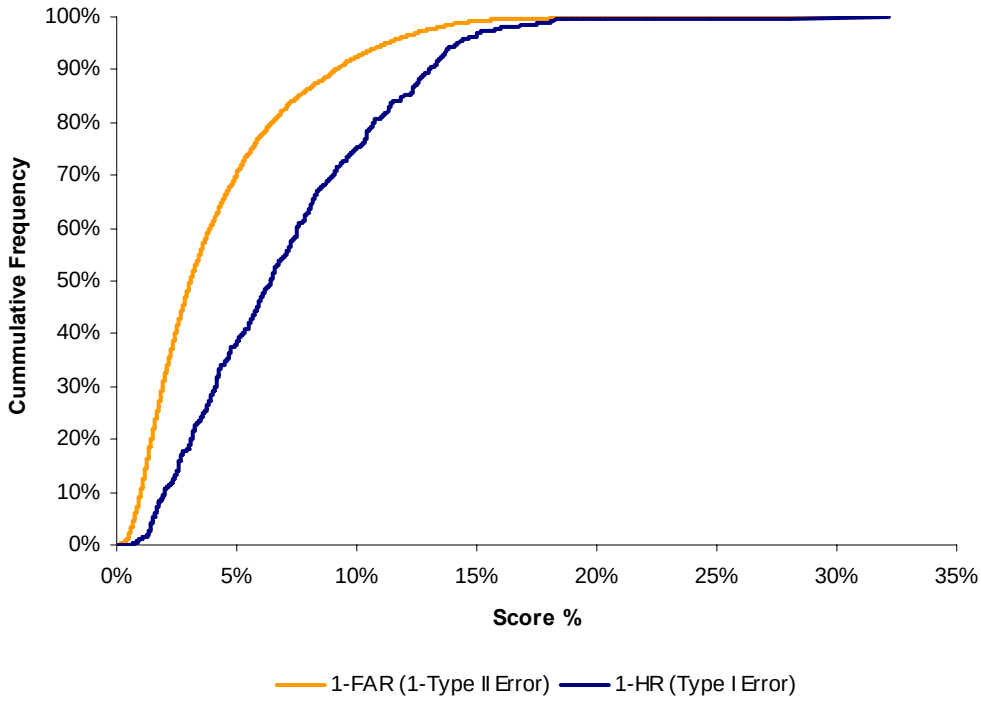

Figure 25 – Model B: Kolmogorov-Smirnov Analysis

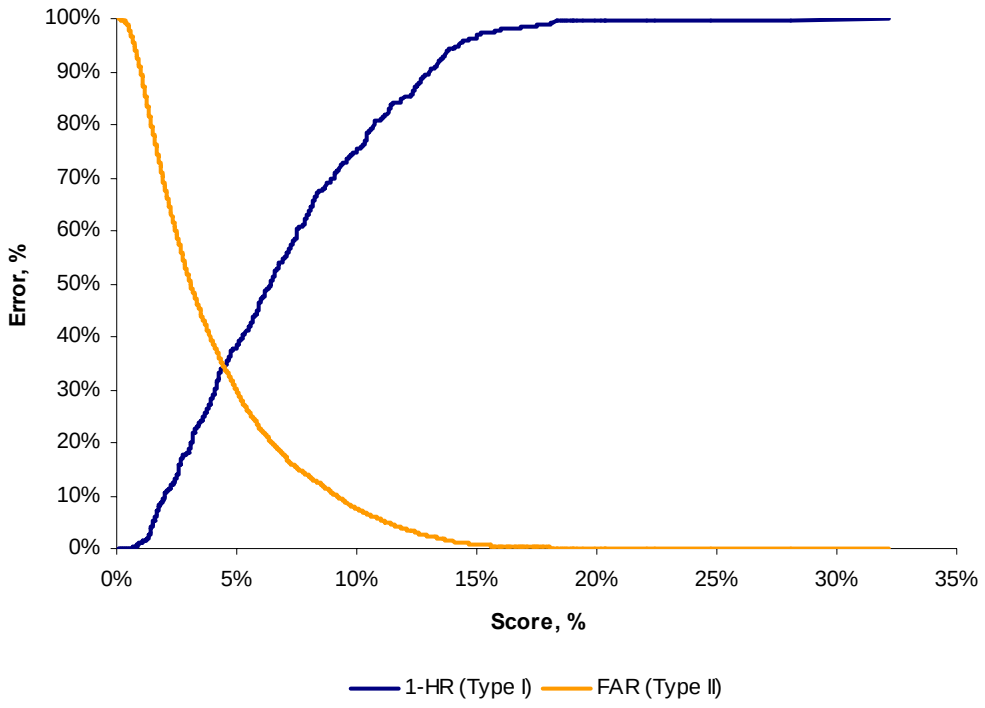

Figure 26 – Model B: Types I & II Errors

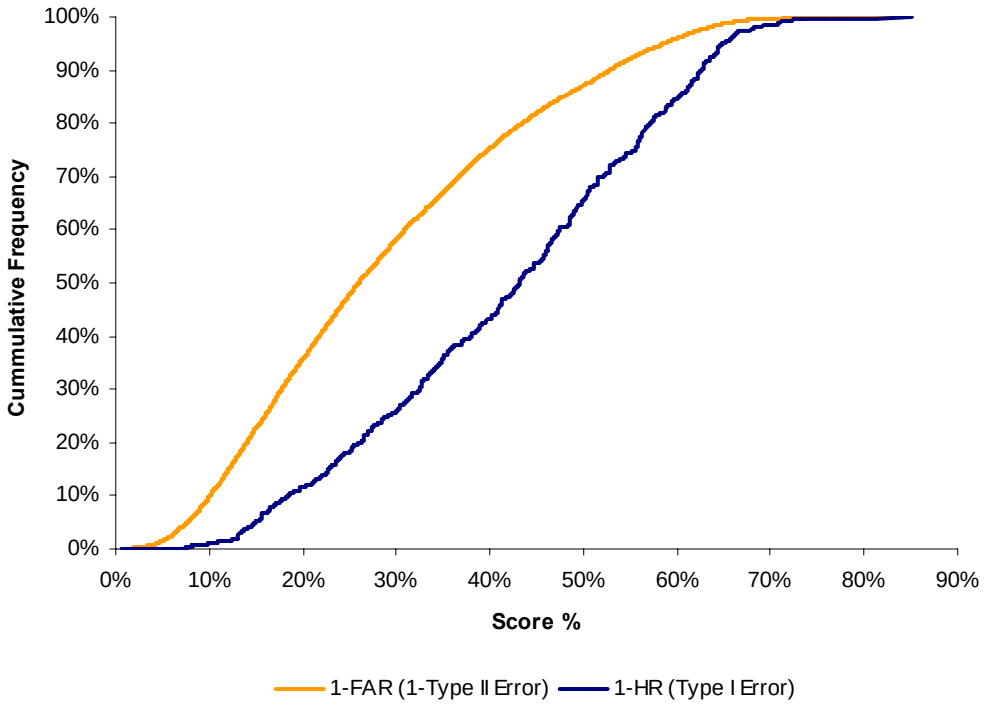

Figure 27 – Model C: Kolmogorov-Smirnov Analysis

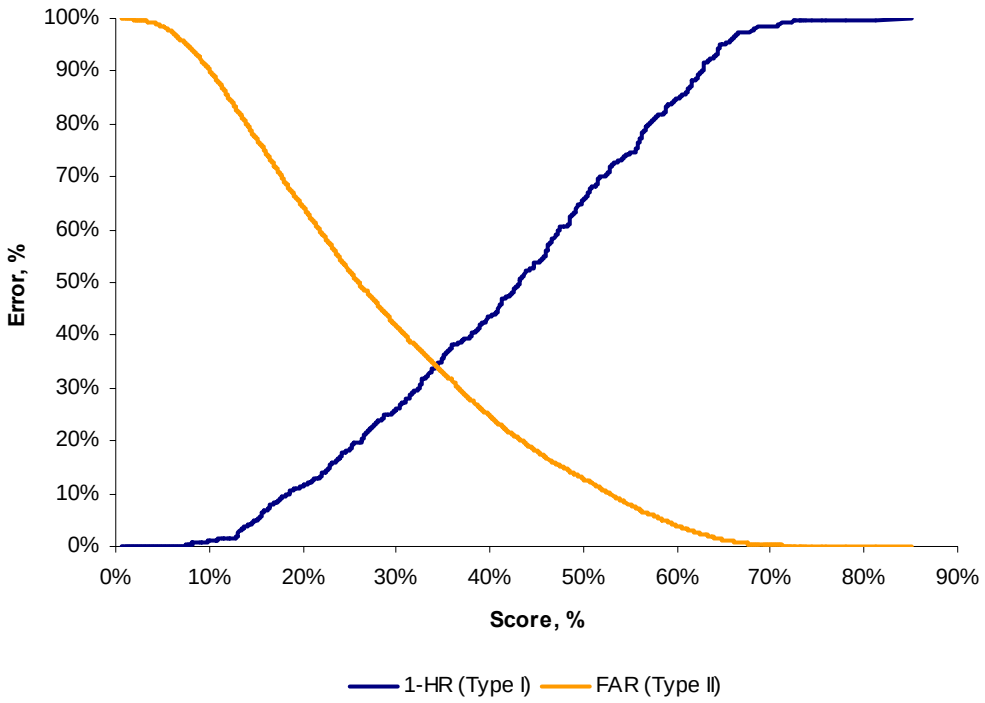

Figure 28 – Model C: Types I & II Errors

Appendix 6 – K-Means Clustering

K-Means Clustering[36] is an optimization technique that produces a single cluster solution that optimizes a given criteria or objective function. In the case of the methodology applied in this study, the criteria chosen is the Euclidean Distance between each case, c_i and the closest cluster centre C_k:

$$d\left(c_i, C_k\right) = \sqrt{\left(c_i - C_k\right)^2}$$

Cluster membership is determined through an iterative procedure involving two steps:

i. The first step consists on selecting the initial cluster centers. Two conditions are checked for all cases: first, if the distance between a given case c_i and its closest cluster mean C_k is greater than the distance between the two closest means, C_n and C_m, then that case will replace either C_n or C_m, whichever is closer to it. If case c_i does not replace any cluster mean, a second condition is applied: if c_i is further from the second closest cluster's centre than the closest centre if from any other cluster's centre, then that case will replace the closest cluster centre. The initial k cluster centers are set after both conditions are checked for all cases;

ii. The second step consists of assigning each case to the nearest cluster, where the distance is the Euclidean Distance between each case and the cluster centers determined in the previous step. The final cluster means are then computed as the average values of the cases assigned to each cluster. The algorithm stops when the maximum change of cluster centers in two successive iterations is smaller than the minimum distance between initial cluster centers times a convergence criterion.

[36] For more information see, for example, Hartigan (1975).

Appendix 7 – IRB RWA and Capital Requirements for Corporate Exposures

The formulas for calculating the RWA for corporate exposures under the IRB approach are:

$$RWA = k * 12,5 * EAD$$

where *k* is the Capital Requirement, computed as:

$$k = LGD * \Phi\left[\frac{\Phi^{-1}(PD)}{\sqrt{1-R}} + \sqrt{\frac{R}{1-R}} * \Phi^{-1}(0,999)\right] * \frac{1+(M-2,5)*b(PD)}{1-1,5*b(PD)}$$

b(PD) is the Maturity Adjustment:

$$b = \left(0,08451 - 0,05898 * \log(PD)\right)^2$$

and *R* is the Default Correlation:

$$R = 0,12 * \frac{1-\exp(-50*PD)}{1-\exp(-50)} + 0,24 * \left[1 - \frac{1-\exp(-50*PD)}{1-\exp(-50)}\right]$$

PD and LGD are measured as decimals[37], Exposure-At-Default (EAD) is measured as currency, Maturity (M) is measured in years, and Φ denotes the cumulative distribution function for a standard normal random variable.

The Default Correlation (R) formula has a firm-size adjustment of $\left[0,04 * 1 - \left(\frac{S-5}{4}\right)\right]$ for SME borrowers, where *S* is the total annual sales in Millions of Eur, and $5 \leq S \leq 50$. SME borrowers are defined as "Corporate exposures where the reported sales for the consolidated group of which the firm is a part is less than 50 Millions of Eur" (Basel Committee on Banking Supervision 2003, par. 242). It is possible for loans to small business to be treated as retail exposures, provided that the

[37] The PD for corporate exposures has a minimum of 0.03%.

borrower, on a consolidated basis, has a total exposure to the bank of less than one Million Eur, and the bank has consistently treated these exposures as retail. For the purpose of this study it is assumed that all exposures are treated as corporate exposures.

Thus, ignoring both Market and Operational risks, we have:

$$\text{Capital Ratio} = \frac{\text{Regulatory Capital}}{\text{Total RWA}}$$

If the minimum value for the capital ratio (8%) is assumed, then:

$$\text{Regulatory Capital} = 8\% * \text{Total RWA}.$$

Appendix 8 - IRB Capital Requirements Figures

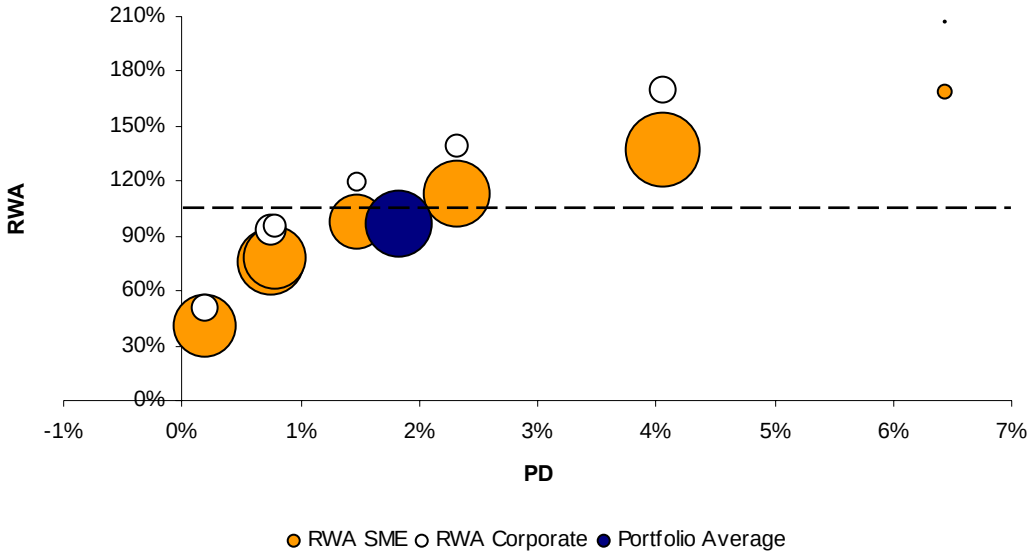

Figure 29 – Model A - IRB Capital Requirements (Cluster Method)

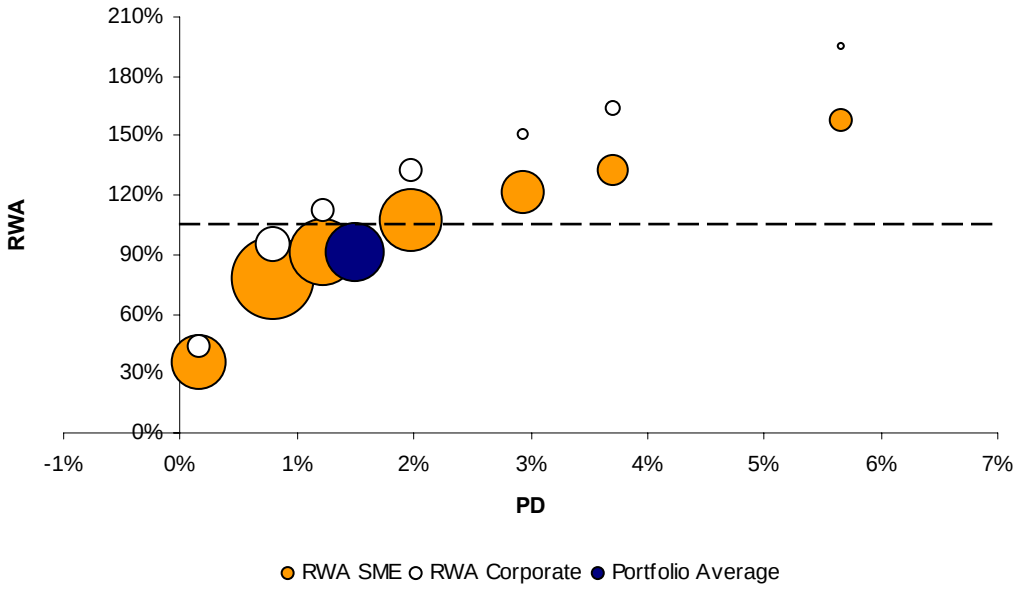

Figure 30 – Model B - IRB Capital Requirements (Cluster Method)

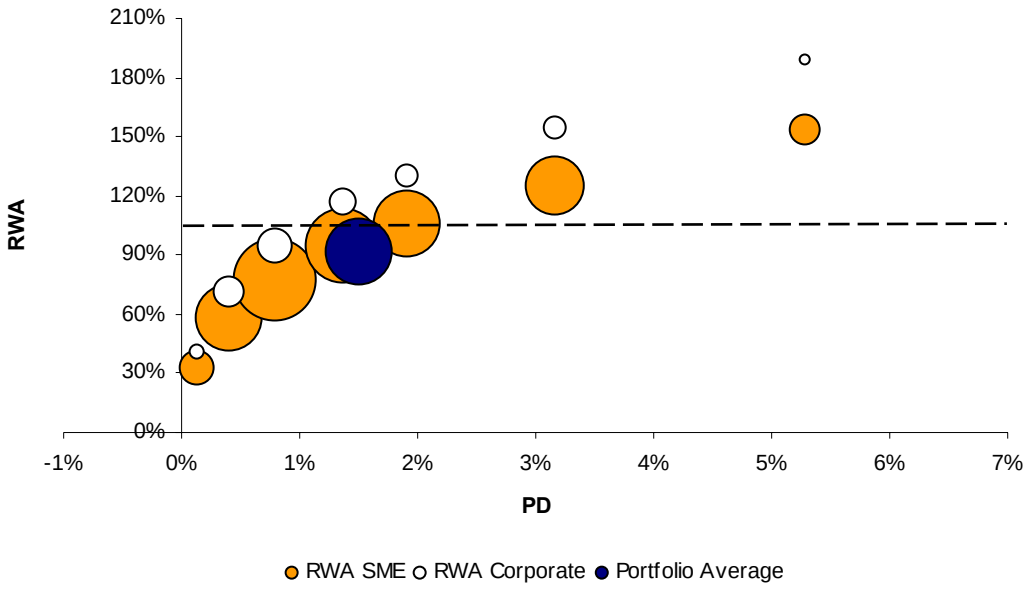

Figure 31 – Model C - IRB Capital Requirements (Cluster Method)

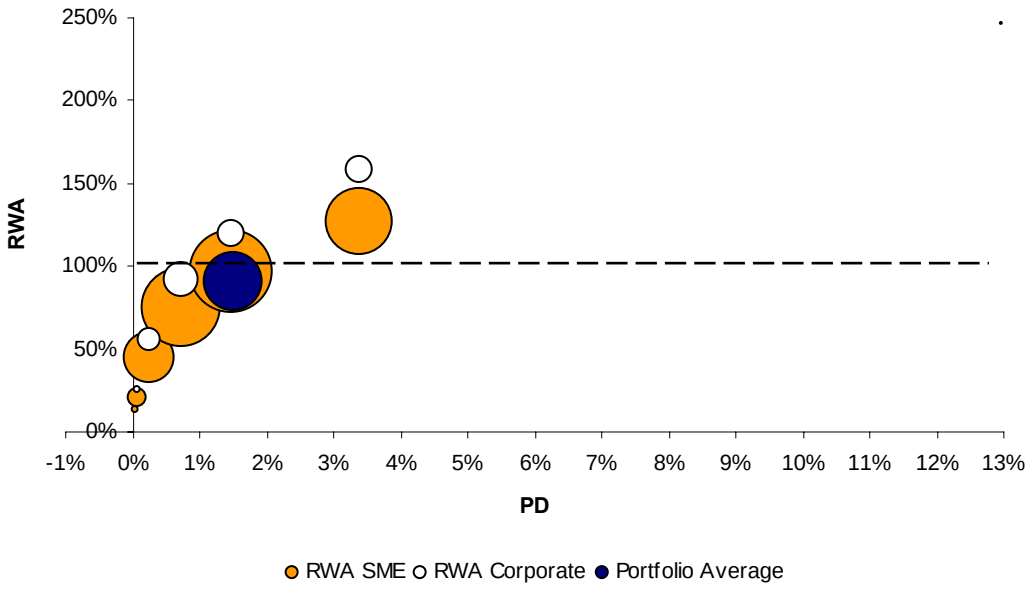

Figure 32 – Model A - IRB Capital Requirements (Historical Method)

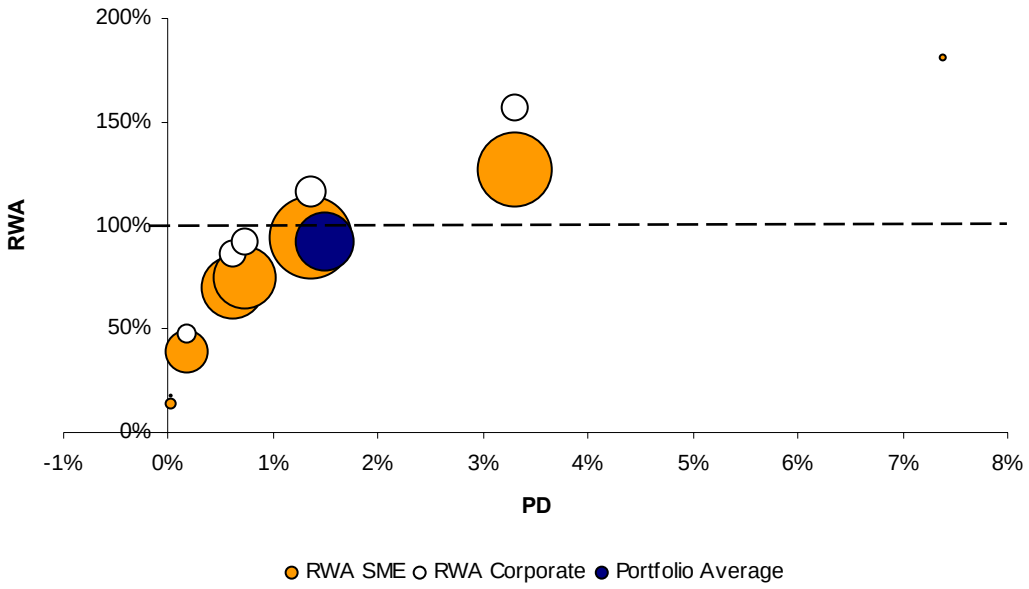

Figure 33 – Model B - IRB Capital Requirements (Historical Method)

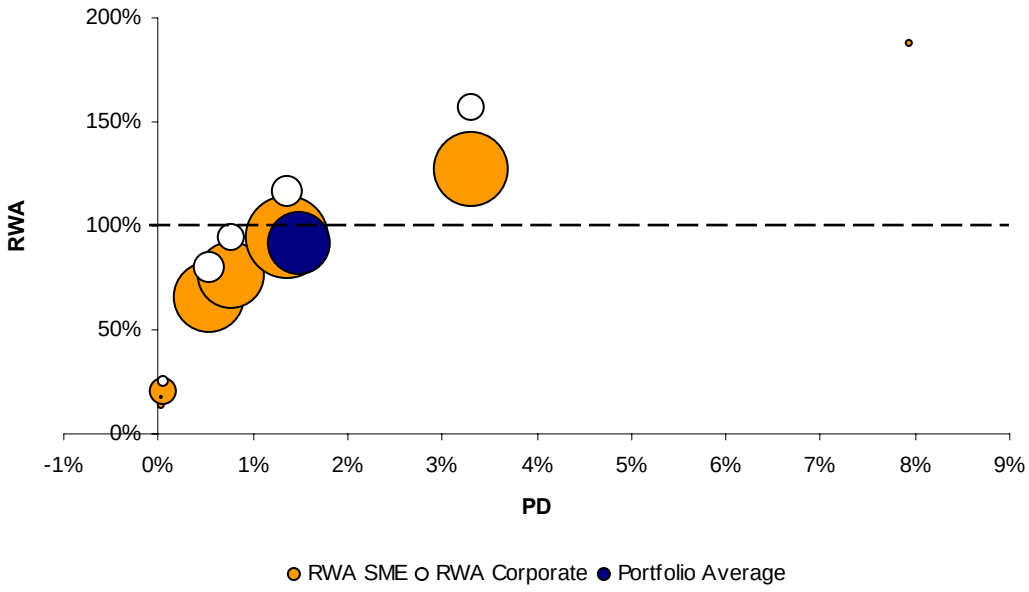

Figure 34 – Model C - IRB Capital Requirements (Historical Method)

Chapter II

Loss Given Default Estimation

1 Introduction

The ongoing revision of the regulatory capital requirement framework has motivated an increased interest on the measurement and modeling of key credit risk inputs[38]. In particular, for corporate bank loan portfolios, under the proposed Advanced Internal Ratings Based approach (AIRB) banks will be required to provide, among others, internal estimates of the Loss-Given-Default (LGD) for each exposure on the portfolio. Due to lack of publicly available data, most of the literature is concerned with the analysis of credit risk on corporate bonds. The scarce research using private bank data is usually limited to the estimation of individual probabilities of default. LGD modeling on bank loan portfolios for privately held corporate customers is rare due to high data requirements. It involves a wealth of knowledge on long recovery processes motivated by a default event, which is by itself a rare event.

Although most of the available research focuses on the estimation of the Probability of Default (PD), LGD has a potentially higher impact on the proposed regulatory capital requirements. For example, under the proposed regulation two loans with the same Expected Loss (EL), but with the first having a PD double than the second, but an LGD half of the other, then the Risk-Weighted-Asset (RWA) for the first loan will clearly be lower than that of the second loan. For example, using the formula for corporate exposures with M=3, if Loan 1 has PD=2.00% and LGD=20%, then EL=0.4% and the Risk Weighted Assets (RWA) will be 58.67%. If Loan 2 has PD=1.00% and LGD=40%, then EL will be the same 0.4% but the RWA=92.11%. Thus, this paper aims to build on the limited literature on the empirical estimation of a crucial risk factor, for the corporate market segment of a bank credit portfolio.

Schuermann (2004) and Altman et al. (2005) provide good surveys on LGD modeling literature. The main issues analyzed are LGD measurement and its role on the proposed capital accord, identifying the key LGD drivers, analyzing the

[38] See Basel Committee on Banking Supervision (2001).

relationship between PD and LGD, and LGD modeling methodologies. Earlier literature was restricted to the measurement of historical LGD for a given bank loan portfolio. Asarnow and Edwards (1995), Hurt and Felsovalyi (1998) survey the historical recovery experience from Citibank's loan portfolios, using databases of over 20 years of data for the US and Latin America, respectively. Franks et al. (2004) and Araten et al. (2004) survey the historical recovery rates experience for corporate bank loans on the European and US markets, respectively. Franks et al. (2004) use data from 10 banks in three different countries: France, Germany and the UK. They conclude that the main LGD drivers are the country jurisdiction, collateral, bank recovery procedures, and the loan structure of the firm. Araten et al. (2004) use 18 years of LGD data from a major US bank and demonstrate the importance of the economic cycle for LGD measurement, especially for unsecured loans.

Regarding LGD modeling, several alternatives have recently been proposed. Glößner et al. (2006) suggest an LGD score model were the score is a function of empirical derived haircuts for collateralized exposures, and a loss rate for uncollateralized exposures. This LGD score is then calibrated to reflect the banks' internal loss history. Grunert and Weber (2005) develop an LGD model considering borrower, loan specific and macroeconomic factors as regressors, using data from a large German bank. They apply logistic regression, coding the dependent variable as a dummy variable that assumes 1 if the observed recovery rate is higher than a given percentile.

More sophisticated approaches directly model the LGD variable, which is commonly regarded has a non-normal distributed continuous variable, bounded on the [0,1] interval. Gupton and Stein (2005) and Gupton (2005) or Singh (2003) propose a Beta transformation of LGD data, in order to be able to apply standard regression techniques. Dermine and Carvalho (2006) suggest the application of Generalized Linear Models (GLM) methodology to LGD data extracted from the database of a major European bank, following Papke and Wooldrigdge (1996) that first suggested an application of the GLM methodology to fractional response variables.

The first main contribution of this study is to apply the Kaplan-Meier survival analysis to the historical recovery rate sample, in order to overcome the common

problem of insufficient recovery history. The other main contribution is to estimate and compare for the first time LGD models using both the Beta transformation and the GLM methodologies, applied to a dataset comprising 7 years of recovery experience from the corporate loan portfolio of a European bank. Considering loan, guarantee and customer characteristics, the models can be used to predict long-term LGDs for corporate bank loans.

This paper is structured as follows. Section 2 describes the randomly selected dataset, extracted from the banks' internal database. The derivation of the LGD variable is presented in Section 3, based on the historical recovery experience of the bank. Section 4 presents the variables considered as regressors, as well as their one-to-one relationship with the historical recovery rates. Section 5 develops the LGD models and compares them using both GLM and Beta transformation methodologies. Section 6 concludes.

2 Sample Description

The dataset contains 1,200 observations of defaulted loans randomly selected from the default pool database of a European bank. A loan is considered in default if a given interest or principal installment is overdue by more than 90 days. The sample ranges from January 1999 to December 2005 and considers only loans from corporate customers. Table 15 presents the distribution by original loan maturity, about 60% are medium-term loans with maturities ranging from 2 to 5 years, 30% are short-term loans (maturity lower than 1 year), while only about 10% are long-term loans:

Maturity, Yrs	Number of Obs.	%
1	358	29.83%
2	304	25.30%
3	190	15.82%
4	97	8.12%
5	122	10.15%
6	20	1.69%
7	28	2.37%
8	17	1.44%
9	4	0.34%
10	38	3.13%
More	22	1.86%
Total	**1200**	**100.00%**

Table 15 – Sample Distribution by Original Loan Maturity

This table presents the distribution of the 1.200 default observations on the sample by original loan maturity.

Figure 35 displays the sample distribution by the year of the loan and the year of default. Although seven years are considered most of the loans are from the 2001-2004 period, while most of the defaults occur in the 2003-2005 period:

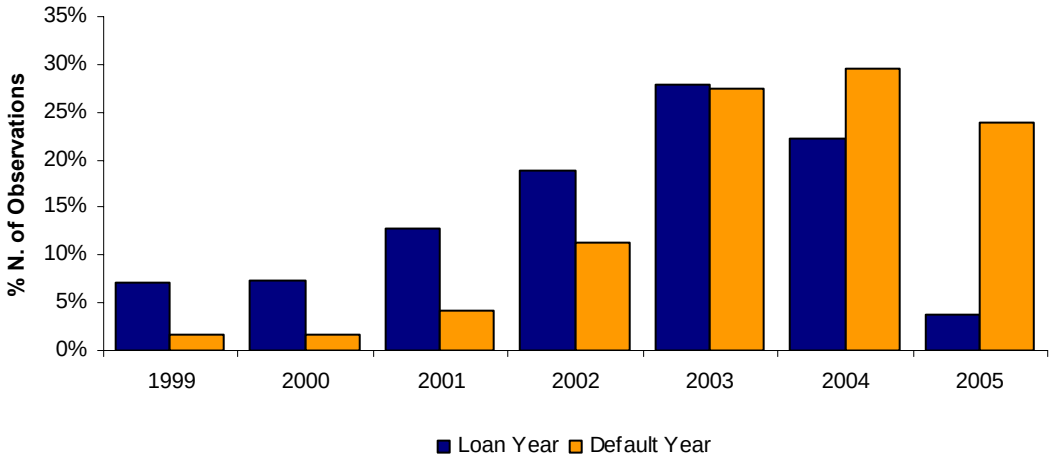

Figure 35 – Sample Distribution by Year of Loan and Year of Default
This figure displays the distribution of observations by the year each loan started and by the year each default occurred. Seven years of data are represented in the sample, ranging from January 1999 to December 2005.

Figure 36 presents the sample distribution by the loan size, there is a concentration of small loans (up to 30 thousand Eur) and on large loans (above 150 thousand Eur):

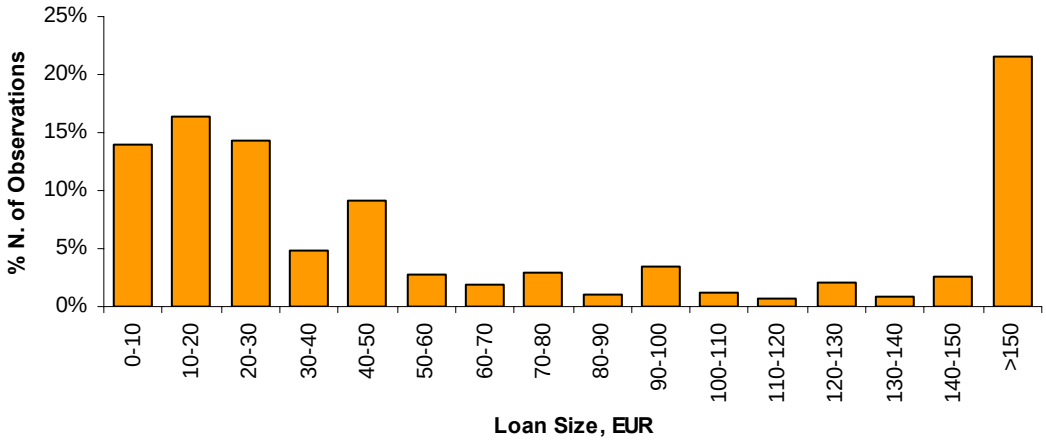

Figure 36 – Sample Distribution by Original Loan Size, thousand EUR
The figure shows the distribution of the defaulted loans considered by original size in thousand of Euros.

3 LGD Definition

LGD is defined as the complement of the Recovery Rate (RR), which in turn can be defined as the ratio between the present value of the sum of recoveries associated with a given loan, net of costs supported by the bank and the total outstanding amount at the time of default (EAD)[39]:

$$LGD = 1 - \frac{PV\left(R - C\right)}{EAD}, \tag{9}$$

where,

R represents cash recoveries received by the bank; both recoveries before and after legal contention were considered; if a given defaulted loan is restructured and replaced by a new loan, the former is considered fully recovered if all the overdue installments at the time of default are paid, and the new loan has not defaulted. C represents costs supported by the bank; these include specific costs such as legal expenses, collateral liquidation and/or insolvency expenses; furthermore, generic costs that affect indirectly all recovery processes are also considered; these generic costs represent mainly costs associated with internal recovery departments; total costs are estimated to represent 2.45% of the total recovered amount; EAD is the total amount outstanding at the time of default; only principal installments are considered, interest due not paid at the time of default is not included.

The annual discount rate chosen to compute the net present value of the recoveries is 15%. This rate represents the opportunity cost to the bank of holding capital against defaulted assets. Thus, the discount rate equals the average Return On Equity (ROE) of the bank on the whole sample period. The use of a high discount rate is consistent with the prudent view suggested by the new Basel capital accord

[39] For the remainder of this study we will work with the Recovery Rate since the interpretation of the results is more intuitive, although they can all be reinterpreted in terms of LGD.

regarding the risk parameters estimations. Other common alternatives for the discount rate are a risk free rate or the interest rate of the loan. Maclachlan (2004) provides a survey of several alternatives. Following a proposal by the Financial Services Authority, FSA (2003), it is suggested that the appropriate discount rate should correspond to the rate the bank would apply to an asset of similar risk. Since this rate is unknown for the dataset used in this study, the opportunity cost approach is used.

Figure 37 shows the distribution of the recovery rates. These recovery rates follow a bimodal distribution, with high concentration of observations at the extreme 0% and 100% rates. This result has been reported frequently in the literature[40].

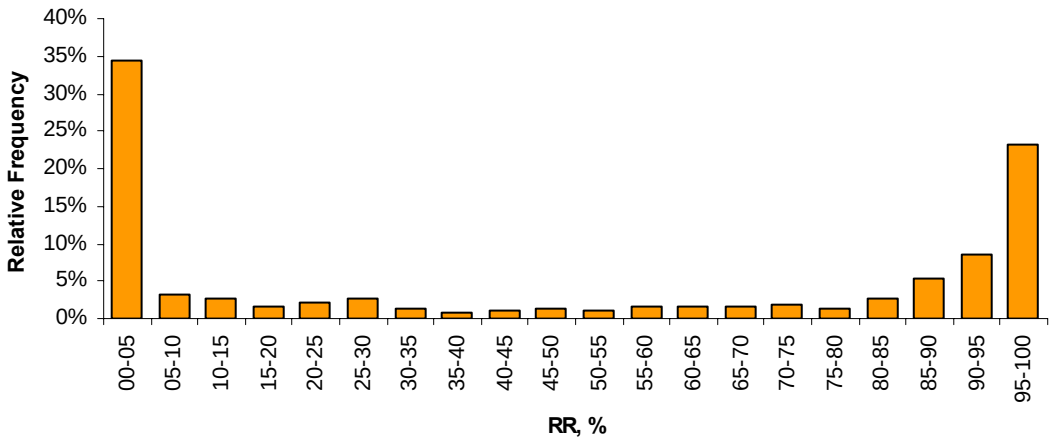

Figure 37 – Empirical Recovery Rate Distribution
The figure presents the historical recovery rate distribution. The bimodal distribution has high concentration on the extreme 0% and 100% recovery rates.

In order to study the evolution of the historical recovery rates with time, the Kaplan-Meier (1958) survival analysis is applied to monthly recovery data. This analysis allows to determine how much of the defaulted amount in a given year is recovered *x* months after the default, taking in account the fact that not all possible recovery periods are observed (see Appendix 1 for more details). Figure 38 and Figure 39 present the Kaplan-Meier analysis for each year and for the average curve for all periods:

[40] See, for example, Asarnow and Edwards (1995), Hurt and Felsovalyi (1998), Schuermann (2004) and Dermine and Carvalho (2005).

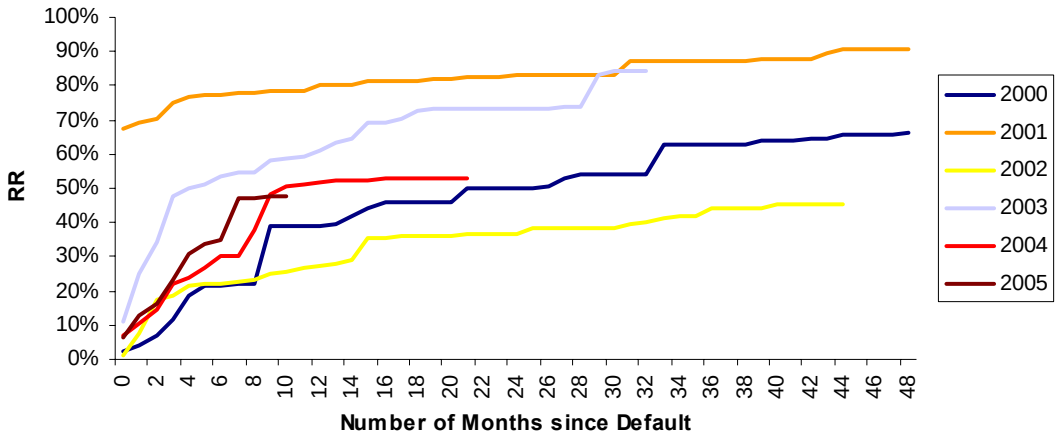

Figure 38 – Cumulative Recovery Rate Growth by Year

The figure shows the cumulative recovery rate for each annual vintage, x months after the default date. Results are obtained using the Kaplan-Meier survival analysis.

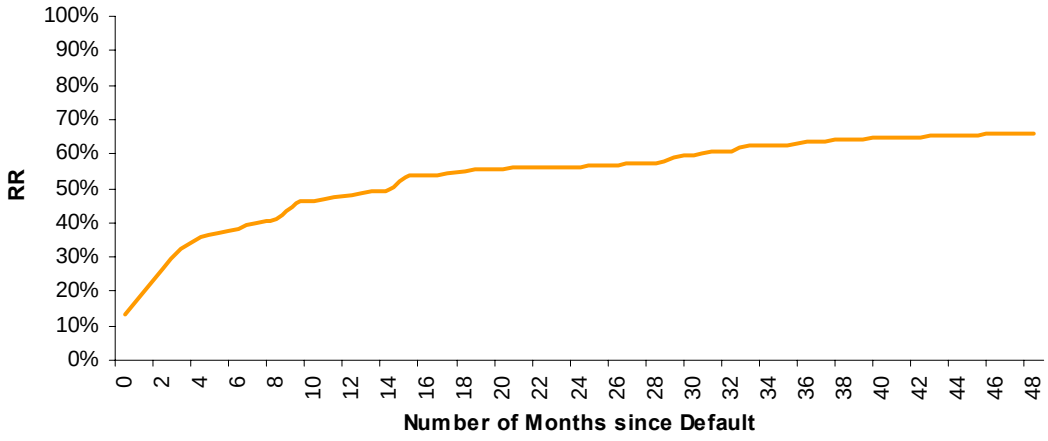

Figure 39 – Average Cumulative Recovery Rate Growth

This figure presents the average of the annual vintages for the cumulative recovery rates, x months after the default date.

The cumulative average recovery rate curve grows at a diminishing rate through time. Most of the recoveries occur during the first months after the default, and the recovery rate stabilizes after 48 months at 67%.

Given that the main goal of the LGD model is to provide long-term estimates, and that not all default observations have had at least 48 months to be recovered, it is

useful to use the average recovery curve in order to project all of the censored recovery observations until the 48 month after default period. This allows to consider more observations for the long-term recovery model (only 220 observations have had 48 month recovery period), insuring that all default observations have the same period of recovery.

As a result we will have the estimated cumulative recovery rate for 1,200 defaulted loans, 48 months after the default. Table 16 presents the summary statistics for the Recovery Rate variable. Applying the 15% discount rate, the average recovery rate is 48.59%. Franks et al. (2004) report, for the same discount rate, average recovery rates of 64.6%, 36.2% and 51.1% for the UK, France and German samples respectively, although the loans considered have different recovery periods and no survival analysis is applied. Dermine and Carvalho (2006) report an undiscounted average recovery rate of 71% for the subset of loans with 48 month recovery period.

Summary Statistics	
Mean	48.59%
Median	50.70%
StDev	42.66%
Min	0.00%
Max	100.00%

Table 16 – Cumulative Recovery Rate Summary Statistics

This table presents summary statistics for the cumulative recovery rates, 48 months after the default, using a 15% discount rate. The statistics presented are the Mean, Median, Standard Deviation, Minimum and Maximum.

4 Recovery Drivers

The main objective of this study is to develop an econometric model that is useful to predict the LGD associated with bank loans granted to private corporate customers. Given this objective, it is only relevant to consider as potential explanatory variables those that are known ex-ante to the default event. Thus variables such as the year of default or control dummy's for organizational changes in internal recovery departments can useful to explain historical LGDs but cannot be used in a predictive model.

Table 17 below lists the 10 variables that are considered in this study, they represent three relevant dimensions: guarantee, loan and customer characteristics[41]:

Section	Type	Variable
4.1	Guarantee Characteristics	Guarantee Type: Financial, Mortgage, Personal, Other Collateral
4.2	Loan Characteristics	Loan Value
		Loan Maturity
		Default Value / Loan Value
		Seasoning / Maturity
		Interest Rate
4.3	Customer Characteristics	Industry
		Age of Relationship
		Firm Age
		Geographic Location

Table 17 – List of Explanatory Variables

The table lists the 10 variables used as explanatory variables for the recovery rate models. The variables represent guarantee, loan and customer characteristics.

In order to study the relationship between each variable and the recovery rate, the CHAID – Chi-squared Automatic Interaction Detector (Kass 1980) methodology is applied. Through this classification tree methodology it is possible to classify the

[41] If data was available, other variables such as the ratio between the value of the loan and the value of the collateral, the borrower rating, borrower size and business line could have been tested.

observations of a given categorical or continuous variable into homogeneous groups. It is based on an iterative algorithm that attempts at each step to find the optimal split of the predictor variable through a chi-squared test. Thus, in this study, this methodology is applied for each independent variable, in order to find homogeneous groups of observations in terms of the average recovery rate for each group. The advantage of this procedure can be illustrated by considering the Industry variable as an example. Instead of directly considering the 43 different industries represented in the sample, only four groups of industries are considered, each group containing industries with similar recovery experiences. Given the low number of observations of the sample (1,200), estimating 3 parameters for this variable instead of 42 represents an important increase in the degrees of freedom of the model.

Next, we provide a description of the variables selected and the relationship between each variable and the recovery event.

4.1 Guarantee Characteristics

4.1.1 Guarantee Type

Four types of guarantees are considered: personal guarantees, mortgage collateral (residential or commercial), financial collateral and other collateral. For the selected loans, 89% have a personal guarantee associated and 38% have some type of collateral. Overall, the average RR is 48.59%, collateralized loans have an average 55.25% RR, while uncollateralized loans have an average RR of 45.12%. Araten et al. (2004) report a global 56.7% RR, with 59.1% RR for secured loans and 49.5% for unsecured loans.

Since the bank demands a high degree of coverage for the loans with collateral, this variable will be introduced in the multivariate model as dummy variables for each type considered. Analyzing Table 18, loans with higher historical recovery rates have financial collateral, followed by loans with other non-financial

collateral. Finally, loans with personal guarantees have lower recovery rates. Araten et al. (2004) report a similar result, loans associated with financial collateral have the highest historical recovery rate, followed by loans with other non-financial collateral: accounts receivable, inventory, fixed assets, and mortgages.

Guarantee Type	% Number Guarantees	Average RR	Std. Deviation
Financial	2.72%	72.14%	28.96%
Other Coll.	6.90%	65.77%	39.97%
Mortgage	20.17%	49.20%	41.98%
Personal	70.21%	46.69%	42.89%

Table 18 – Guarantee Weight and Recovery Rate by Guarantee Type
The table presents both the weight of each type of guarantee, the average historical recovery rate and the standard deviation for the defaulted loans associated with each type of guarantee. The guarantee types considered are: Personal Guarantee, Mortgage, Financial and Other Collateral.

4.2 Loan Characteristics

4.2.1 Loan Value

The size of the loan at the time of default can be an indicator of the recovery rate since banks have a higher propensity to dedicate more resources to monitor and recover high value loans. Figure 40 presents this relationship with the loans ranked in the highest percentiles, in terms of loan value having the highest recovery rates. Hurt and Felsovalyi (1998) and Dermine and Carvalho (2006) report the opposite result, stating that larger loans are more complex and take longer periods of time to recover than smaller loans.

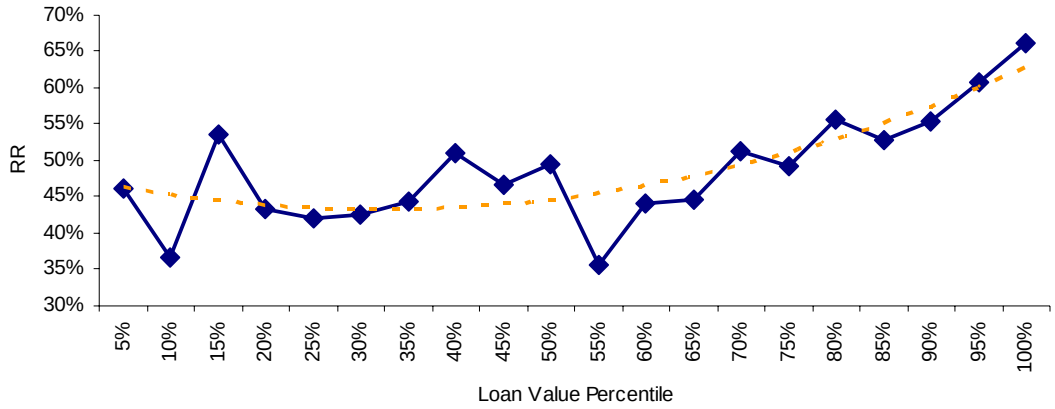

Figure 40 – Average Recovery Rate by Loan Value Percentile

This figure illustrates the relationship between Loan Value and the historical Recovery Rate. Loans are ranked in ascending order, in terms of their value at the time of default. For each percentile the average recovery rate is calculated.

Summary Statistics	
Mean	201,245.47
Median	42,000.00
StDev	956,098.75
Min	949.72
Max	19,951,915.88

Table 19 – Loan Value Summary Statistics, in EUR

This table presents summary statistics for the Loan Value variable. The statistics presented are the Mean, Median, Standard Deviation, Minimum and Maximum.

4.2.2 Loan Maturity

The maturity of the loan can be interpreted as a recovery rate regressor since long-term loans have low amortization schedules, thus if the defaulted loan becomes performing again, i.e. the customer restarts payment according to the original plan, the sum of the discounted recovery payments will be lower for longer-termed loans.

Table 20 displays this relationship, short-term loans with maturity less or equal to 1 year have an average recovery rate lower than middle-term loans (with maturities between 2 and 5 years), which in turn have a lower average recovery rate than the long-term loans (maturity higher than 5 years).

To the best of our knowledge, and up to this point, no study has reported empirical recovery rates for corporate bank loans by loan maturity.

Maturity, Years	% Number Observations	Average RR	Std. Deviation
0<Y<=1	32.83%	50.88%	46.38%
2<Y<=5	56.60%	48.63%	41.04%
5<Y	10.58%	41.24%	38.35%
Total	**100.00%**	**48.59%**	**42.66%**

Table 20 – Loan Frequency and Recovery Rate by Loan Maturity

This table presents the relative loan frequency, average historical recovery rate and standard deviation for short (maturity up to 1 year), medium (maturity between 2 and 5 years) and long-term loans (maturity higher than 5 years).

4.2.3 Default-to-Loan Value and Seasoning-to-Loan Maturity Ratios

Two dynamic measures of the sunk-in effect from the loan amortization over time on the recovery rate are the relative weight of the seasoning of a loan over its original maturity, and the weight of the default value over the original value of the loan. Both variables provide alternative evidence of the higher propensity to recover on more seasoned loans. The lower the ratio between the default amount and the original value of the loan, the higher the recovery rate. Conversely, high recovery rates are also usually associated with high seasoning to maturity ratios. Due to the diversity of amortization schedules of the loans considered, both measures are not highly correlated.

Up to this point and to the best of our knowledge, no study has considered these variables as corporate bank loan recovery drivers. Figure 41 and Figure 42 provide evidence of the relationships described, presenting the average recovery rates for each percentile of the value of each ratio, ranked in ascending order, while Table 21 provides the summary statistics for both variables.

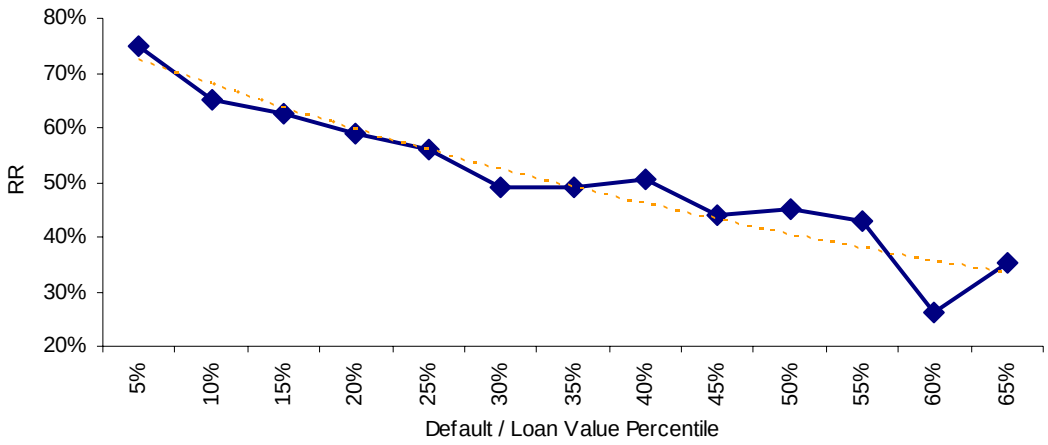

Figure 41 – Average Recovery Rate by Default-to-Loan Value Ratio Percentile

The figure shows the relationship between the Default-to-Loan Value Ratio and the historical Recovery Rate. Loans are ranked in ascending order, in terms of the value of the ratio at the time of default. For each percentile the average recovery rate is calculated.

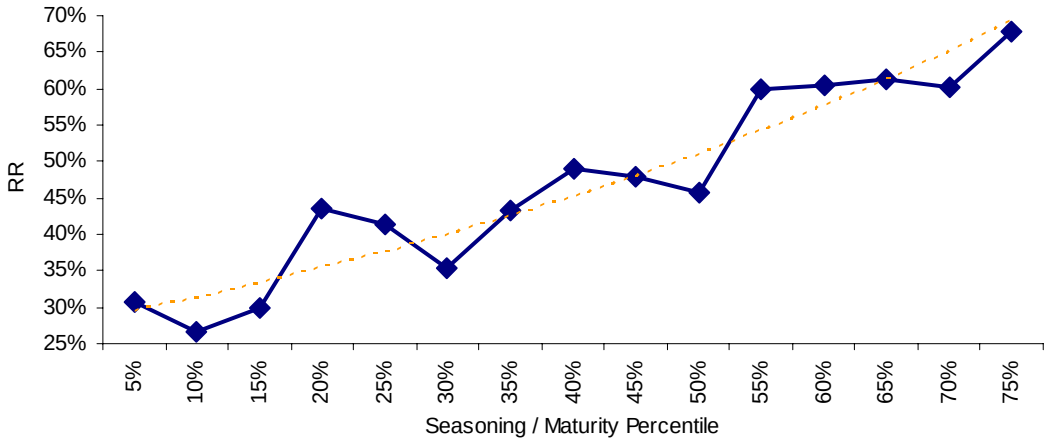

Figure 42 – Average Recovery Rate by Seasoning-to-Maturity Ratio Percentile

This figure illustrates the relationship between the Seasoning-to-Maturity Ratio and the historical Recovery Rate. Loans are ranked in ascending order, in terms of the value of the ratio at the time of default. For each percentile the average recovery rate is calculated.

Summary Statistics	D-L Ratio	S-M Ratio
Mean	78.37%	54.54%
Median	88.84%	51.76%
StDev	25.60%	33.30%
Min	1.34%	2.08%
Max	100.00%	100.00%

Table 21 – Default-to-Loan Value and Seasoning-to-Maturity Ratios Summary Statistics
This table presents the summary statistics for the Default-to-Loan Value and Seasoning-to-Maturity Ratios. The statistics presented are the Mean, Median, Standard Deviation, Minimum and Maximum.

4.2.4 Interest Rate

The original interest rate of the loan is also used to predict recovery rates. For this dataset, loans with higher interest rates are associated with lower recovery rates. The inclusion of this variable can be interpreted has a proxy for the risk of the firm if the bank has a risk based pricing for the spreads, and also as a reflection of the market conditions at the time the loan is granted. Table 22 below provides evidence of this effect, using the CHAID methodology a dummy is constructed for loans with low (<= 4.25%) or high (> 4.25%) interest rates.

Interest Rate	% Number Observations	Average RR	Std. Deviation
<= 4.25%	20.81%	58.47%	43.29%
> 4.25%	79.19%	45.99%	42.13%
Total	**100.00%**	**48.59%**	**42.66%**

Table 22 – Loan Frequency and Recovery Rate by Interest Rate
This table presents the relative loan frequency, average historical recovery rate and standard deviation for low and high interest rate groups.

4.3 Customer Characteristics

4.3.1 Firm Industry

The firm industry variable is constructed by grouping together level 3 NACE codes (classification of economic activities in the European community) with similar historical recovery experiences. Applying the CHAID methodology, 4 groups are considered. Appendix 2 provides the details on how the 4 groups are formed and Table 23 provides the sample distribution and average recovery rate for each industry group.

Group A, were the most representative industry is the NACE code 70 – Real Estate Activities, has the highest average recovery rate. Group B represented by industry 45 – Construction has the second highest average recovery rate. Next, Groups C and D have the lowest average recovery rates. The most representative industries in Group C are services and manufacturing activities, while wholesale trade is the major industry in Group D.

Dermine and Carvalho (2006) report a similar result, stating that the manufacturing and trade sectors have lower recoveries than the real sector, while Franks et al. (2004) found little systematic relationship between industry groups and recovery rates.

Group	% Number Observations	Average RR	Std. Deviation
A	7.61%	78.92%	32.49%
B	35.28%	58.53%	39.92%
C	20.14%	47.74%	41.59%
D	36.97%	33.32%	41.64%
Total	**100.00%**	**48.59%**	**42.66%**

Table 23 – Loan Frequency and Recovery Rate by Industry Group
This table presents the sample distribution by industry group and the historical recovery experience for each group.

4.3.2 Age of Relationship

The age of the relationship between the customer and the bank can be a recovery rate indicator in view of the fact that if the relationship is strong, the firm could have a higher incentive to repay the defaulted loans.

Franks et al. (2004) report higher average recovery rates for older customers with more than 5 years of relationship, than for new customers with less than 5 years.

Table 25 provides the relative number of observations and the average recovery rates for defaulted loans with short to long-term customer-bank relationships. The average RR does not increase consistently with the age of relationship, but for very long-term relationships (over 20 years), the historical recovery rate is clearly higher than for shorter relationships.

Summary Statistics	
Mean	7.4
Median	6.0
StDev	4.8
Min	0.0
Max	23.0

Table 24 – Age of Relationship Summary Statistics, in Years
This table presents summary statistics for the Age of Relationship variable. The statistics presented are the Mean, Median, Standard Deviation, Minimum and Maximum.

Years	% Number Observations	Average RR	Std. Deviation
0<Y<=2	10.58%	44.33%	37.97%
2<Y<=10	66.33%	50.66%	42.45%
10<Y<=20	19.63%	42.99%	45.05%
20<Y	3.47%	62.02%	40.66%
Total	**100.00%**	**48.59%**	**42.66%**

Table 25 – Loan Frequency and Recovery Rate by Age of Relationship
This table presents the sample distribution and the historical recovery experience by the age of relationship between the customer and the bank.

4.3.3 Firm Age

Loan recovery can occur at several stages, first the customer is able to repay according to the original plan, second the customer is able to pay but only with an

alternative payment plan, or finally, the customer is not able to repay and either the guarantees provided can cover the outstanding debt or recovery can only occur through legal dispute. Considering firm age we hypothesize that defaulted loans on older firms have higher likelihood of being recovered in the earlier stages. The earlier the recovery, the higher the sum of the discounted cash-flows.

Evidence from our data (see Table 27) suggests that although the RR does not increase consistently with firm age, the historical RR associated with defaulted loans on very young firms (up to two years) is clearly lower than the RR on older firms' loans.

Dermine and Carvalho (2006) report a similar result, stating that for older firms it is easier to evaluate management and asset quality.

Summary Statistics	
Mean	14.4
Median	11.0
StDev	12.7
Min	1.0
Max	103.0

Table 26 – Firm Age Summary Statistics, in Years

This table presents summary statistics for the Firm Age variable. The statistics presented are the Mean, Median, Standard Deviation, Minimum and Maximum.

Years	% Number Observations	Average RR	Std. Deviation
0<Y<=2	2.96%	33.83%	36.56%
2<Y<=10	39.85%	49.99%	41.69%
10<Y<=20	29.36%	50.31%	43.79%
20<Y	18.61%	43.98%	43.79%
#N/A	9.22%	--	--
Total	**100.00%**	**48.59%**	**42.66%**

Table 27 – Loan Frequency and Recovery Rate by Firm Age

This table presents the sample distribution and the historical recovery experience by the age of the customer.

4.3.4 Geographic Location

Economic conditions in a given region may affect the recovery rates for the loans granted to customers in those areas. For example, if an economic crisis affects a given region, the value of mortgage collateral and other firms' assets would most likely

decrease, affecting recovery rates. Another important factor could be the bank structure. If the bank has a decentralized structure, with regional departments using different policies to evaluate and monitor loans, there could be a regional bias affecting recovery rates.

Several studies have considered geographic location has a recovery driver, but at a country level: Hurt and Felsovalyi (1998) compare recovery experiences in 27 different Latin American countries, while Franks et al. (2004) report results for three European countries.

Table 28 provides the relative number of observations and the historical average RR for each major region in Portugal. The regions with better and worse average RRs are the island groups *Açores* and *Madeira*, respectively. This might be a reflection of the fact that both regions have administrative autonomy that could result, for example, in different timings to solve legal disputes in courts. Another interesting result is that all regions in the center and south of mainland Portugal have RRs higher than the overall average, while regions in the North have lower RRs than the average (with the exception of *Beira Interior*).

Location	% Number Observations	Average RR	Std. Deviation
Açores	1.18%	58.22%	28.21%
Beira Interior	4.74%	56.74%	41.94%
Algarve	2.62%	52.40%	41.79%
Lisboa	26.82%	51.51%	42.75%
Estremadura	6.68%	49.25%	42.62%
Alentejo	4.40%	48.64%	39.81%
Beira Litoral	18.10%	47.00%	43.20%
Trás-os-Montes	2.12%	46.82%	45.13%
Porto	11.93%	46.53%	42.68%
Minho	20.56%	44.69%	43.69%
Madeira	0.76%	39.28%	38.79%
#N/A	0.08%	--	--
Total	**100.00%**	**48.59%**	**42.66%**

Table 28 – Loan Frequency and Recovery Rate by Geographic Location
This table displays the sample distribution and the historical recovery experience for the major regions in Portugal.

5 LGD Modeling

The modeling of the LGD differs from standard regression models in two fundamental ways: first, the dependent variable is usually restricted to the [0,1] interval; second, the dependent variable clearly does not follow a normal distribution, with heavy concentrations on the extremes. Classic OLS regression would yield a poor fit to the data and would predict recovery rates outside the [0,1] interval.

Two alternative methodologies will be employed in order to overcome these issues. The first, the Beta Transformation methodology consists on a two-step transformation of the dependent variable. First, the Beta distribution is fitted to the observed recovery rates, if the fit is satisfactory, a second transformation is performed in order to map the range of outcomes from the probability space to the normalized space. After this second transformation it is now possible to apply the standard regression techniques.

The second alternative consists on applying the Generalized Linear Model (GLM) methodology to our data. This methodology allows to directly overcome the two issues mentioned above by estimating a regression where the dependent variable is mapped to the real space and, at the same time, allowing for an alternative for the normal distribution of the errors. The two methodologies are discussed in detail below, and the results from applying each of them are presented.

5.1 Beta Transformation Methodology

The Beta Transformation methodology attempts to map the recovery values to the normalized space through the use of a parametric distribution[42]. If this empirical match between the historical recovery rates and the Beta distribution is reasonable then it will be possible to apply the standard regression techniques to the transformed dependent variable.

The probability function of the Beta distribution with domain [0,1] is given by:

$$b(x,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-x)^{\beta-1}x^{\alpha-1}, \qquad (10)$$

where Γ is the gamma function, and α and β are the shape parameters. These shape parameters can be derived from the mean μ and standard deviation σ of the population:

$$\alpha=\mu\left(\frac{\mu(1-\mu)}{\sigma^2}-1\right) \qquad \text{and} \qquad \beta=(1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2}-1\right)$$

By matching the population moments to the sample moments ($\hat{\mu}$= 0.486, $\hat{\sigma}$=0.425), a Beta Distribution with parameters α = 0.186 and β = 0.197 is fitted:

[42] See for example: Gupton and Stein (2005), or Miu and Ozdemir (2005).

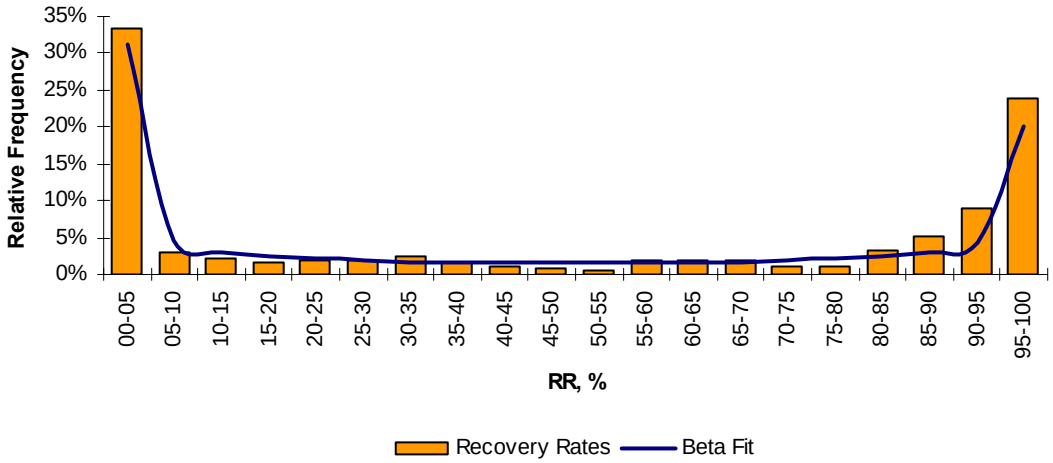

Figure 43 – Beta Distribution Fit to Empirical Recovery Rate Distribution
This figure illustrates the fit of a Beta distribution with parameters α = 0.186 and β = 0.197 to the empirical recovery rate distribution.

Descriptive Statistics	
Mean	49.27%
Median	56.32%
Std Deviation	42.54%
Min	0.00%
Max	100.00%

Table 29 – Beta Fit Descriptive Statistics
This table provides the descriptive statistics mean, median, standard deviation, minimum and maximum for the beta fit to the Recovery Rate variable.

Given that the Beta distribution seems to provide a satisfactory fit to the sample recovery data, the Beta transformation of the dependent variable is given by:

$$Y_i = N^{-1}\left[\mathrm{B}\left(RR_i, \alpha, \beta\right)\right], \tag{11}$$

where N^{-1} is the inverse normal cumulative distribution, B is the Beta cumulative distribution and RR_i is the observed recovery rate. Since the inverse normal cumulative distribution function is not defined for the extreme values 0% and 100%, the observed recovery rates need to be adjusted by a small value ε at the extremes.

With the normalized dependent variable it is now possible to apply the standard regression techniques. In order to be able to directly compare the results from the Beta and GLM methodologies, the model is estimated using Maximum-

Likelihood procedure. Applying backward and forward variable selection techniques, the final model is[43]:

Explanatory Variable	Estimated Coefficient	Standard Error	Wald Test P-Value
S/M	0.1569	0.0377	0.000
D/L	-0.3142	0.0483	0.000
irate_d	-0.0572	0.0290	0.048
ind2	-0.2345	0.0472	0.000
ind3	-0.3321	0.0498	0.000
ind4	-0.5907	0.0467	0.000
g_per	-0.1312	0.0377	0.001
g_oth	0.1802	0.0414	0.000
g_fin	0.1865	0.0637	0.003
const	1.0205	0.0770	0.000

Number of observations	1133
Wald Chi-Squared Test	418.6
P-Value Wald Test	0.000
AIC	0.909
BIC	-7736
McFaden R-squared	0.262

Table 30 – Maximum-Likelihood estimates of long-term cumulative recovery rates, Beta Methodology

This table shows the Beta Methodology maximum-likelihood estimates for the cumulative recovery rate model. A positive relationship between the variables Seasoning-to-Maturity Ratio (S/M) and Collateral Dummies (g_oth and g_fin) and the Recovery Rate is suggested. Conversely, a negative relationship is estimated between the Recovery Rate and the other variables, the Default-to-Loan Value Ratio (D/L), the high Interest Rate Dummy (irate_d), and Industry Groups 2, 3 and 4 (ind2, ind3, ind4) relative to Group 1. The statistical significance of each individual coefficient is provided by the p-value of the individual Wald test. A Wald test to the overall significance of the model, the McFaden R^2 and the AIC/BIC information criteria are also provided.

where,
S/M = Seasoning to Maturity Ratio;
D/L = Default to Loan Value Ratio;
Irate_d = Interest Rate Dummy;
Ind_i = Dummy for industry belonging to group *i*;
G_per = Dummy for personal guarantee;
G_other = Dummy for collateral than mortgage or financial;
G_fin = Dummy for financial collateral;

The sign of the estimated coefficients is consistent with the economic intuition, as discussed in the previous section: higher recovery rates are estimated for more seasoned loans (measured by the Seasoning-to-Maturity Ratio – S/M), and for loans

[43] Appendix 3 provides a definition of the statistics used.

with collateral (measured by the dummies g_oth and g_fin). Conversely, lower recovery rates are estimated for loans that have low repayment rates by the time of default (measured by the Default-to-Loan Value Ratio – D/L), for loans with high interest rates (measured by the dummy irate_d), and for loans granted to customers of industry groups 2, 3 and 4 (dummies ind2, ind3, ind4). The other variables mentioned in the previous section, Loan Value, Loan Maturity, Age of Relationship, Firm Age and Geographic Location were also tested but are not significant.

Regarding statistical significance, each individual coefficient and the overall regression are significant at a 5% level. As to residual analysis, model misspecification can be detected if the estimated residuals clearly are not normal distributed. Figure 44 below provides a plot of the estimated recovery rates against the Pearson residuals of the regression, no gross misspecification is detected, although there is a slight bias towards low estimated recoveries:

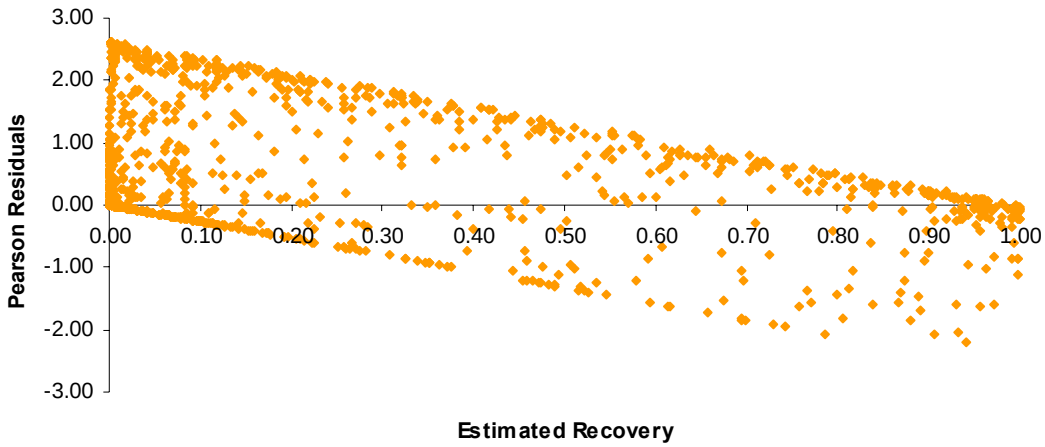

Figure 44 – Pearson Residuals Plot by Estimated Recovery Rate, Beta Methodology
This figure presents the scatter between the estimated recovery rates, using the Beta Methodology, and the Pearson Residuals of the estimate.

5.2 GLM Methodology

The GLM models represent a generalization of the classical regression models in the sense that the relationship between the dependent variable *y* and the covariates *x* need not to be linear, and the dependent variable can follow a distribution other than the normal distribution[44]. GLM requires that the relationship of a transformation of *y*, given by a link function *g( )*, and x to be linear, with *y* following a given distribution *F*:

$$g\{E(y)\} = x\beta\ , \qquad y \sim F \tag{12}$$

In the case of this study, we are interested in specifications for *g( )* and *F* that take in consideration the fact that the dependent variable is a proportion, and that it follows a bimodal distribution. One possible solution would then be to consider the binomial family, with denominator 1, and the loglog link function. With the binomial family, other link functions that map *y* from the probability space to the real space are feasible, such as the logit and cloglog functions[45].

The loglog link is used since it is asymmetric, biased towards the lower values, which fits well with the high concentration on the 0% to 5% observed recoveries in our sample (see Figure 37). The logit function is symmetric while the cloglog function is asymmetric but biased towards positive values:

[44] See Hardin and Hilbe (2001) for more details on GLM models.
[45] Results using the 3 proposed link functions yield similar results, although the loglog regressions seems to provide the best fit.

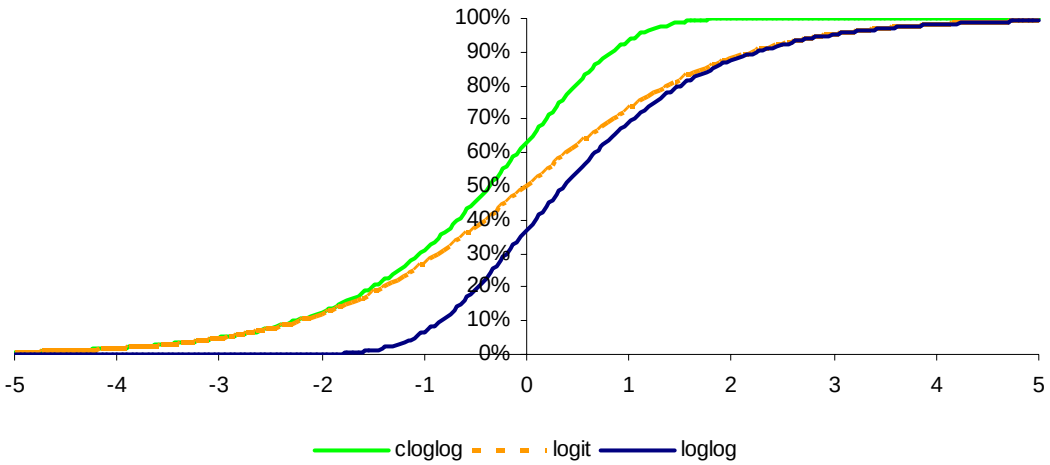

Figure 45 – Cloglog, Logit and Loglog Link Functions
The figure presents the three link functions considered. With the Loglog function there is a bias towards lower recoveries, which is closer to the empirical recovery rate distribution of the sample. The Cloglog function is biased towards high recoveries, while the Logit function lies between the other two.

where,

$$\text{cloglog} = \ln\left\{-\ln\left(1-y\right)\right\},\ \text{logit} = \ln\left\{y/\left(1-y\right)\right\},\ \text{loglog} = -\ln\left\{-\ln\left(y\right)\right\}$$

Model estimation is then developed using the Maximum Likelihood method. The log-likelihood function for observation *i* is given by:

$$\ln L_i = -\ln\left[\Gamma\left(y_i+1\right)\right] - \ln\left[\Gamma\left(2-y_i\right)\right] + y_i \ln\left[g\left(x_i,b\right)\right] + \left(1-y_i\right)\ln\left[1-g\left(x_i,b\right)\right] \quad (13)$$

The estimated results for our sample are:

Explanatory Variable	Estimated Coefficient	Standard Error	P-Value
S/M	0.9268	0.1167	0.0000
D/L	-1.1460	0.1766	0.0000
irate_d	-0.2344	0.1001	0.0190
ind2	-1.2999	0.2003	0.0000
ind3	-1.5642	0.2058	0.0000
ind4	-2.3143	0.1964	0.0000
g_per	-0.4703	0.1359	0.0010
g_oth	0.6879	0.1507	0.0000
g_fin	0.7009	0.2225	0.0020
const	2.9608	0.2998	0.0000

Number of observations	1133
Wald Chi-Squared Test	519.6
P-Value Wald Test	0.000
AIC	1.015
BIC	-7125
McFaden R-squared	0.203

Table 31 – Maximum-Likelihood estimates of long-term cumulative recovery rates, GLM Methodology

This table shows the GLM Methodology maximum-likelihood estimates for the cumulative recovery rate model. The same variables are selected as for the Beta Methodology model. A positive relationship between the variables Seasoning-to-Maturity Ratio (S/M) and Collateral Dummies (g_oth, g_fin) and the Recovery Rate is suggested. Conversely, a negative relationship is estimated between the Recovery Rate and the other variables, the Default-to-Loan Value Ratio (D/L), the high Interest Rate Dummy (irate_d), and Industry Groups 2, 3 and 4 (ind2, ind3, ind4) relative to Group 1. The statistical significance of each individual coefficient is provided by the p-value of the individual Wald test. A Wald test to the overall significance of the model, the McFaden R^2 and the AIC/BIC information criteria are also provided.

where,
S/M = Seasoning to Maturity Ratio;
D/L = Default to Loan Value Ratio;
Irate_d = Interest Rate Dummy;
Ind_i = Dummy for industry belonging to group *i*;
G_per = Dummy for personal guarantee;
G_other = Dummy for collateral than mortgage or financial;
G_fin = Dummy for financial collateral;

The same explanatory variables are used as in the Beta Transformation methodology. Again, the sign of the estimated coefficients agrees with economic intuition. Regarding statistical significance, each individual coefficient and the overall regression are significant with a 5% level. A plot of the estimated recovery rates against the Pearson residuals again shows no evidence of gross misspecification:

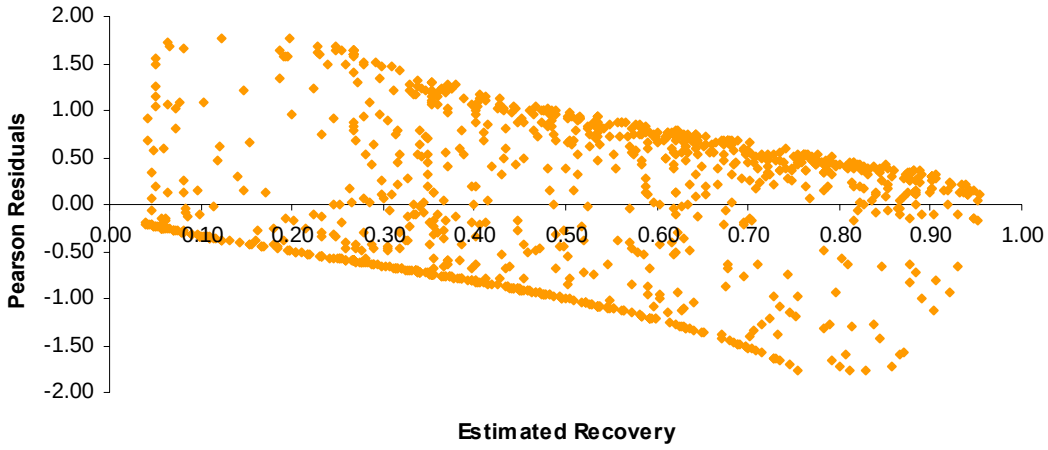

Figure 46 – Pearson Residuals Plot by Estimated Recovery Rate, GLM Methodology
This figure presents the scatter between the estimated recovery rates, using the GLM Methodology, and the Pearson Residuals of the estimate.

Comparing the results from both methodologies, we can conclude that both are able to suitably model the LGD of the portfolio considered in this study. The resulting models are similar in the sense that the same variables are selected, and both can be used to predict long-run LGDs, producing predictions on the]0,1[interval.

Although the GLM methodology is simpler to apply, since the Beta methodology involves a two-step transformation of the dependent variable, it is the Beta methodology that provides the best overall fit: it has a higher McFadden R^2 (26.22% against 20.25%), and lower values for both AIC (0.9089 against 1.0146) and BIC (-7736 against -7125) information criteria. As reference, Dermine and Carvalho (2006) report Pseudo R^2 of 20% and 18% for two alternative GLM models, using the loglog link function.

Furthermore, a random sample of 60 observations, representing 5% of the overall sample is used for out-of-sample testing, providing similar results for both methodologies. A Root Mean Squared Error (RMSE) of 35.41 is estimated for the Beta methodology, while the RMSE for predictions using the GLM methodology is 34.54[46].

[46] See Appendix 3 for more details on the RMSE.

6 Conclusion

LGD modeling of bank loans for privately-held firms is usually hampered by the lack of available data. Unlike probability of default (PD) the LGD sample does not comprise the whole corporate bank loan portfolio, but only those loans where there has been a default. Starting from this considerable smaller base, a considerable amount of information is then required for these observations over a long period of time, that should encompass at least one full economic cycle. All the cash-flows generated during the recovery processes should be retrieved, typified (capital, interest, and cost), quantified and dated. The present value of these cash-flows should then be computed using the appropriate discount rate that can change over time, across institutions and over different business segments.

These issues have been addressed in this study using simplifying assumptions. First of all, since it is not possible to retrieve interest and cost cash-flows for most observations only capital recoveries are considered and an overall cost of 2.45% of total recoveries is distributed per recovery process according to the size of each loan. Additionally, a constant annual discount rate of 15% is chosen, representing the opportunity cost to the bank of holding capital against defaulted assets. Furthermore, since only a limited time frame of data is available (1999-2005), a considerable number of observations have limited recovery time. Given the objective of modeling long-term LGD, survival analysis is applied to these observations in order to simulate cumulative recoveries over longer periods.

Regardless of these data issues we are able to provide two alternative specifications that provide simple but efficient models for long-term LGD prediction. Relationships suggested by each model are stable across samples and different recovery horizons. The models presented encompass exclusively explanatory variables that are known at the time the loans are granted and thus can be used to predict long-term LGDs for new bank loans. Overall 10 explanatory variables are

considered, representing three relevant dimensions: guarantee, loan and customer characteristics.

Two alternative modeling methodologies are considered that attempt to overcome the fact that LGD has a bimodal distribution with heavy concentrations on the [0,1] interval. The Beta Transformation methodology maps the recovery values to the normalized space through the use of a parametric distribution. It relies on the fact that the empirical match between the historical recovery rates and the Beta distribution is reasonable. Alternatively, the GLM methodology represents a generalization of the classical regression models allowing the dependent variable to follow a distribution other than the normal distribution and for a nonlinear relationship between the dependent variable and the covariates. Although both methodologies suggest similar valid models, the Beta Transformation methodology provides the best fit to the data considered in this study.

Further work would consider a larger dataset comprising interest due not paid recoveries, a more detailed study on the indirect and direct costs of recovery incurred by the bank and the firm during the recovery process, and alternative specifications for the discount rate. In terms of the modeling methodology, a non-parametric approach could be tested that could potentially provide a better fit over the methodologies presented in this study, at the expense of increasing the model complexity.

Bibliography

Altman, E., A. Resti and A. Sironi, 2003, The Link between Default and Recovery Rates: Theory, Empirical Evidence and Implications, *Working Paper*.

Altman, E., A. Resti and A. Sironi, 2005. *Recovery Risk: The Next Challenge in Credit Risk Management* (Risk Books).

Araten, M., M. Jacobs Jr. and P. Varshney, 2004, Measuring LGD on Commercial Loans: An 18-year Internal Study, *RMA Journal*, May.

Asarnow, E. and D. Edwards, 1995, Measuring Loss on Defaulted Bank Loans: a 24-Year Study, *The Journal of Commercial Lending* 77, 11-23.

Basel Committee on Banking Supervision, 2001, The New Basel Capital Accord, *Bank for International Settlements*.

Basel Committee on Banking Supervision, 2004, Background Note on LGD Quantification, *Bank for International Settlements*.

Dermine, J. and C. Neto de Carvalho, 2006, Bank Loan Losses-Given-Default: A Case Study, *Journal of Banking & Finance* 30, 1219-1243.

Franks, J., A. de Servigny and S. Davydenko, 2004, A Comparative Analysis of the Recovery Process and Recovery Rates for Private Companies in the U.K., France, and Germany, *Standard & Poors Risk Solutions* May.

FSA - Financial Services Authority, 2003, Report and first consultation on the implementation of the new Basel and EU Capital Adequacy Standards, *Consultation Paper* 189.

Glößner, P., A. Steinbauer and V. Ivanova, 2006, Internal LGD Estimation in Practice, *Wilmott* January, 86-91.

Grunert, J. and M. Weber, 2005, Recovery Rates of Bank Loans: Empirical Evidence for Germany, *University of Mannheim Working Paper*.

Gupton, G., 2005, Advancing Loss Given Default Prediction Models: How the Quiet Have Quickened, *Economic Notes* 34, 185-230.

Gupton, G. and R. Stein, 2002, LossCalc™: Model for Predicting Loss Given Default LGD, *Moody's Investors Service*.

Gupton, G. and R. Stein, 2005, LossCalc V2: Dynamic Prediction of LGD, *Moodys KMV Modeling Methodology*.

Hardin, J. and J. Hilbe, 2001. *Generalized Linear Models and Extensions* (Stata Press).

Hurt, L. and A. Felsovalyi, 1998, Measuring Loss on Latin American Defaulted Bank Loans: A 27-Year Study of 27 Countries, *The Journal of Lending and Credit Risk Management* 80, 41-46.

Kaplan, E. and P. Meier, 1958, Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* 53, 457-481.

Kass, G., 1980, An exploratory technique for investigating large quantities of categorical data, *Applied Statistics* 29, 119-127.

Maclachlan, I., 2004, Choosing the Discount Factor for Estimating Economic LGD, *Working Paper*.

Miu, P. and B. Ozdemir, 2005, Basel Requirement of Downturn LGD: Modeling and Estimating PD & LGD Correlations, *Working Paper*.

Papke, L. and J. Wooldrigdge, 1996, Econometric Methods for Fractional Response Variables with an Application to 401 K Plan Participation Rates, *Journal of Applied Econometrics* 11, 619-632.

Schuermann, T., 2004, What do we know about LGD? *Forthcoming in D. Shimko ed., Credit Risk Models and Management 2nd Edition* (Risk Books, London, UK).

Singh, M., 2003, Recovery Rates from Distressed Debt - Empirical Evidence from Chapter 11 Filings, International Litigation, and Recent Sovereign Debt Restructurings, *IMF Working Paper* 03/161.

Appendix 1 – Kaplan Meier Survival Analysis

The Kaplan-Meier (1958) survival analysis is a nonparametric, actuarial technique for estimating time-related events. It can be used to measure the length of time required to recover from a given default event. An important feature of this analysis is that it takes into account censored data, the losses from the sample before the final outcome is observed. The Kaplan-Meier estimate for the survival function is given by:

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{r_i}{d_i} \right),$$

where,

r_i = amount recovered in period t_i;

d_i = amount that can be recovered in period t_i.

Appendix 2 – Industry Groups by Economic Activity

Group	NACE	Description
A	37	Recycling
	30	Manufacture of office machinery and computers
	71	Renting of machinery and equipment without operator…
	14	Other mining and quarrying
	02	Forestry, logging and related service activities
	35	Manufacture of other transport equipment
	70	Real estate activities
	80	Education
B	31	Manufacture of electrical machinery and apparatus n.e.c.
	05	Fishing, fish farming and related service activities
	22	Publishing, printing and reproduction of recorded media
	92	Recreational, cultural and sporting activities
	15	Manufacture of food products and beverages
	55	Hotels and restaurants
	45	Construction
	52	Retail trade, except of motor vehicles and motorcycles; repair of personal …
	01	Agriculture, hunting and related service activities
	13	Mining of metal ores
C	60	Land transport; transport via pipelines
	72	Computer and related activities
	74	Other business activities
	33	Manufacture of medical, precision and optical instruments, watches and clocks
	93	Other service activities
	34	Manufacture of motor vehicles, trailers and semi-trailers
	19	Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, …
	91	Activities of membership organizations n.e.c.
	28	Manufacture of fabricated metal products, except machinery and equipment
	36	Manufacture of furniture; manufacturing n.e.c.
	85	Health and social work
	25	Manufacture of rubber and plastic products
D	51	Wholesale trade and commission trade, except of motor vehicles and motorcycles
	20	Manufacture of wood and of products of wood and cork, except furniture; …
	18	Manufacture of wearing apparel; dressing and dyeing of fur
	29	Manufacture of machinery and equipment n.e.c.
	40	Electricity, gas, steam and hot water supply
	17	Manufacture of textiles
	26	Manufacture of other non-metallic mineral products
	50	Sale, maintenance and repair of motor vehicles and motorcycles; retail sale …
	63	Supporting and auxiliary transport activities; activities of travel agencies
	98	Unknown Activity
	24	Manufacture of chemicals and chemical products
	27	Manufacture of basic metals
	64	Post and telecommunications

Table 32 – Industry Groups by Economic Activity Classification

The table lists the level 3 NACE codes (classification of economic activities in the European community) associated with each industry group. The groups are formed using the CHAID methodology, linking together industries with similar recovery experiences.

Appendix 3 – Regression Fit Analysis

This appendix provides a technical description of the several metrics used in this study in order to assess the overall fit of the estimated regressions and to compare results from different methodologies. Four types of fit measures are considered: residual analysis, information criteria, R^2 and Mean Squared Error measures:

Residual Analysis – Pearson Residuals

Residual analysis provides a description of the divergence between the observed and fitted values for each individual observation. Model misspecification can be detected if the estimated residuals clearly are not normal distributed. The Pearson residuals are defined as:

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{V(\hat{y}_i)}},$$

where,

y_i = observed values for observation i;
$\hat{y}_i$ = fitted values for observation i;
$V(\hat{y}_i)$ = variance function of the distribution family considered.

Information Criteria

Model comparison for models estimated with the maximum likelihood procedure can be performed through the use of information criteria that balance likelihood results with penalty terms based on the degrees of freedom of a model. Two popular criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), the lower the value of each criterion, the better the fit of the model:

$$\text{AIC} = \frac{-2\ln L + 2k}{n},$$

$$\mathrm{BIC} = D - (n-k)\ln(n),$$

where,

L = overall likelihood of the model;
k = number of predictors;
n = number of observations;
D = overall deviance of the model.

Pseudo-R^2

For likelihood estimated models, the classical R^2 measure cannot be directly computed. Instead several alternatives have been suggested that attempt to measure the percent variance explained. One popular alternative is the McFadden likelihood-ratio index:

$$R_M^2 = 1 - \frac{\ln L(M_1)}{\ln L(M_0)},$$

where,

M_1 = model with intercept and predictors;
M_0 = model with intercept only.

Root Mean Squared Error (RMSE)

An indicator of model prediction performance is obtained by taking the root of the average of the squared errors of the prediction, the lower the value of the RMSE, the better the fit:

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-1}}$$

where,

y_i = observed values for observation i;
$\hat{y}_i$ = predicted values for observation i;
n = number of observations;