APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind. 2008; **24**:291–306
Published online 18 February 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.705

Assessing the default risk by means of a discrete-time survival analysis approach

Daniele De Leonardis[1,*,†] and Roberto Rocci[2]

[1]*Intesa SanPaolo SpA*, *Corporate Banking*, *Milan*, *Italy*
[2]*Department SEFeMEQ*, *University of Tor Vergata*, *Rome*, *Italy*

SUMMARY

In this paper, the problem of company distress is assessed by means of a multi-period model that exploits the potentialities of the survival analysis approach when both survival times and regressors are measured at discrete points in time. The discrete-time hazards model can be used both as an empirical framework in the analysis of the causes of the deterioration process that leads to the default and as a tool for the prediction of the same event. Our results show that the prediction accuracy of the duration model is better than that provided by a single-period logistic model. It is also shown that the predictive power of the discrete-time survival analysis is enhanced when it is extended to allow for unobserved individual heterogeneity (frailty). Copyright © 2008 John Wiley & Sons, Ltd.

Received 3 March 2006; Revised 11 June 2007; Accepted 19 November 2007

KEY WORDS: Cox's proportional hazards model; (baseline) hazard function; unobserved heterogeneity (frailty)

1. INTRODUCTION

Over the past 40 years, an impressive amount of theoretical and empirical research concerning the analysis and the prediction of the default risk has been produced. Three main approaches can be distinguished. A first approach, usually relying on accounting-based indicators, deals with the identification of an appropriate regression technique. The aim of these studies is to define a limited set of explicative variables together with a classification procedure that should be able to discriminate between safe and risky debtors. Among these studies, we cite the seminal work of Altman [1] and Altman *et al.* [2] who proposed the well-known *Z-score* and *Zeta* models, both based on the use of multivariate discriminant analysis, and those of other authors (e.g. [3–5]) based on binomial regression techniques, such as logit and probit models.

*Correspondence to: Daniele De Leonardis, Intesa SanPaolo SpA, Via del Corso 226, 00186 Rome, Italy.
†E-mail: daniele.deleonardis@intesasanpaolo.com

A second approach that relies on the option-pricing theory stemmed from the work of Merton [6] who was the first to exploit the analogy of corporate equity with options. These 'structural' models assume that the market value of total assets is observable in principle. Furthermore, the capital structure is explicitly considered and default happens if the value of total assets is lower than a certain proportion of the value of liabilities.

A third approach derives a relationship between the credit spreads of risky securities and the probability of default of their issuers. An essential part of the theory of these 'reduced-form models' (e.g. [7, 8]) is the risk neutral valuation under the absence of arbitrage opportunities.

Following a statistical approach, this paper deals with the use of survival analysis techniques in the assessment of the risk of default. Survival analysis studies the length of time that elapses from the beginning of some process to its end or the waiting time before an event occurs. In the present context, the event we are interested in is the company's default.

Since the early empirical applications in bankruptcy prediction, survival analysis has been proposed as an alternative to the well-known single-period approaches based on multivariate discriminant analysis or on qualitative-response models. Among these pioneering studies, we cite those of Lane *et al*. [9] and Luoma and Laitinen [10] that first used the Cox's proportional hazards model to assess the problem of banks' and companies' failure prediction, respectively. More recently, Shumway [11] pointed out that a sample selection bias problem arises from using single-period models based on one, non-randomly selected observation per firm. A systematic inconsistency of the model estimator is the consequence of this selection bias. The author proposes a multi-period logit regression as a discrete approximation of a duration model with time-dependent regressors based on traditional accounting measures. Following a similar approach, Hillegeist *et al*. [12] find that a structural model based on option-pricing theory provides a higher information content than models based on traditional accounting measures. A multi-period logit regression is also employed by Chava and Jarrow [13] who compare a model relying only on accounting variables with one based also on publicly available data related to traded equity. Their evidence supports the hypothesis that market variables play a leading role in failure prediction. They also find that indicators related to the specific industry to which the observed firms belong are statistically significant. A similar approach is used by Beaver *et al*. [14] to check the evolution over the past 40 years of the predictive ability of some ratios based on accounting measures.

In all these works, the problem of the unobserved heterogeneity due to omitted variables is ignored.

Survival models can be extended to take into account different kinds of independent events that can cause the termination of a process (competing risks models). A number of recent studies (e.g. [15–19]) have used these tools to analyze the dynamics of mortgages and personal loans whose termination can be caused not only by the occurrence of the borrower's default but also by the exercise of the prepayment option.

In this paper, we propose the application of the discrete-time counterpart of the Cox's proportional hazards model to analyze and predict the default risk empirically. The same model is also extended to allow for the unobserved heterogeneity. The coherence of both models is tested using a large sample of small and middle-sized Italian companies. The classification performance, which is assessed by means of an out-of-sample validation test, is compared with that of a logit model. This paper is organized as follows: Section 2 deals with the description of the model and its extension incorporating unobserved heterogeneity. Section 3 describes the data used in the application. Section 4 presents the parameter estimates. Section 5 studies the inclusion of system-level variables and reports the results of such an extension. Section 6 follows with a discussion about

DOI: 10.1002/asmb

the classification performance of the models and a comparison with a single-period approach. Concluding remarks are drawn in Section 7.

2. THE MODEL

In this section, we model the probability that a company will default in a given year, conditional upon whether it was solvent up to that point in time. Following Prentice and Gloeckler [20], we approach this problem adapting the Cox's [21] proportional hazards model in order to take account of the discrete structure of data concerning the survival times of firms.

Let T_i be the time when the ith firm defaults. The instantaneous default risk or hazard rate $\lambda_i(t)$ for the ith company at time t is defined as

$$\lambda_i(t) = \lim_{\Delta_t \to 0} \frac{P(t \leqslant T_i < t + \Delta_t | T_i \geqslant t)}{\Delta_t} \tag{1}$$

The hazard rate is assumed to take on the Cox's proportional hazards form

$$\lambda_i(t; \mathbf{x}_i) = \lambda_0(t) \exp[\mathbf{x}_i'(t)\boldsymbol{\beta}] \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard at time t, $\mathbf{x}_i(t)$ is the vector of time-dependent explanatory variables for the ith company, and $\boldsymbol{\beta}$ is a vector of parameters. The baseline hazard $\lambda_0(t)$ is the instantaneous default risk of a firm with $\mathbf{x}_i(t) = 0$. If the covariates are deviations from the mean, then $\lambda_0(t)$ could be interpreted as the hazard rate of an 'average' firm. Thus, model (2) claims an underlying hazard function upon which the individual deviations of the explanatory variables from their mean values act multiplicatively on the instantaneous default risk of an average firm.

Even if the time to default could be viewed in principle as a continuous variable, data concerning a company's survival or default are available on discrete-time basis, usually yearly. Indeed, what we can typically observe is whether a specific company survives or defaults in a given time interval. That means that model (2) should be adapted in order to treat time as grouped into disjoint intervals $[t_1, t_2), [t_2, t_3), \ldots, [t_K, t_{K+1} = \infty)$.

The discrete-time counterpart of (2) is the probability that the ith firm defaults in the interval $[t_k, t_{k+1})$ conditional on its survival upon the beginning of the same interval

$$\lambda_i(t_k; \mathbf{x}_i) = P\{T_i \in [t_k, t_{k+1}) | T_i \geqslant t_k\} \tag{3}$$

Equation (3) could be expressed as the complement of the conditional probability of surviving the interval given that the company was alive at the beginning and can be rewritten as

$$\begin{aligned} \lambda_i(t_k; \mathbf{x}_i) &= 1 - \exp\left(- \int_{t_k}^{t_{k+1}} \lambda_i(u)\, \mathrm{d}u \right) \\ &= 1 - \exp\left(- \int_{t_k}^{t_{k+1}} \mathrm{e}^{\mathbf{x}_i'(t_k)\boldsymbol{\beta}} \lambda_0(u)\, \mathrm{d}u \right) \\ &= 1 - \exp\left(-\mathrm{e}^{\mathbf{x}_i'(t_k)\boldsymbol{\beta}} \int_{t_k}^{t_{k+1}} \lambda_0(u)\, \mathrm{d}u \right) \end{aligned} \tag{4}$$

DOI: 10.1002/asmb

Note that the covariates vector $\mathbf{x}_i$ is a function of t_k, not of the integration variable u, because we assume that the independent variables are measured for an entire interval and not for every instant in the interval $[t_k, t_{k+1})$. We can thus take $e^{\mathbf{x}_i'(t_k)\boldsymbol{\beta}}$ out of the sign of integral. Defining $\alpha_{t_k} = \log \int_{t_k}^{t_{k+1}} \lambda_0(u)\,du$, Equation (4) can be rewritten as

$$\lambda_i(t_k; \mathbf{x}_i) = 1 - \exp[-\exp(\mathbf{x}_i'(t_k)\boldsymbol{\beta} + \alpha_{t_k})] \tag{5}$$

Hence, when considering the time variable as grouped in discrete-time intervals, the Cox's proportional hazards model degenerates to a binary regression with a complementary-log–log link function. Let t_{C_i} be the censoring time for the ith firm (i.e. the time when the ith firm ceases being observed) and t_{B_i} the time when ith firm begins being observed. Following Meyer [22], we can then write the likelihood function associated with Equation (5) as

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N} \left\{ [1 - \exp[-\exp(\mathbf{x}_i'(t_{C_i - 1})\boldsymbol{\beta} + \alpha_{t_{C_i - 1}})]]^{\delta_i} \prod_{k=t_{B_i}}^{t_{C_i}-1-\delta_i} \exp[-\exp(\mathbf{x}_i'(t_k)\boldsymbol{\beta} + \alpha_{t_k})] \right\} \tag{6}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_T]$, with T denoting the time when the survey comes to its end and $\delta_i = 1$ if $T_i \leqslant T$ and 0 otherwise.

The first term of (6) is equal to 1 except when a company defaults in t_{C_i}. The second term is the probability that the ith company survives at least until t_{C_i} or, in case it defaults in t_{C_i}, that it survives at least until $t_{C_i - 1}$. The corresponding log-likelihood function can be expressed as

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \left[\delta_i \log[1 - \exp\{-\exp[\mathbf{x}_i'(t_{C_i - 1})\boldsymbol{\beta} + \alpha_{t_{C_i - 1}}]\}] - \sum_{k=t_{B_i}}^{t_{C_i}-1-\delta_i} \exp[\mathbf{x}_i'(t_k)\boldsymbol{\beta} + \alpha_{t_k}] \right] \tag{7}$$

We move now to the frailty version of model (5). As known, the use of a limited set of covariates could be not completely appropriate in explaining the variability in observed times to default. The excess in unexplained heterogeneity, known as *overdispersion*, could result in an inadequacy of the model in accounting for why firms with shorter times to failure are more frail than others. To address this problem, Vaupel *et al*. [23] firstly introduced a random effect into a survival model framework and applied the model in a demographic setting to account for population heterogeneity. To denote such random effect, they coined the term *frailty*. Thus, a frailty model essentially entails a proportional hazards model conditioned on the random effect. If the unobserved heterogeneity is assumed to take a multiplicative form, then the frailty version of the continuous hazard function can be expressed as

$$\lambda_i(t; \mathbf{x}_i \,|\, V_i = v_i) = v_i \lambda_0(t) \exp[\mathbf{x}_i'(t)\boldsymbol{\beta}] \tag{8}$$

where V_i is a random variable that is assumed to be independent of $\mathbf{x}_i(t)$. It varies over the population of firms and represents the omitted covariates. It is easy to demonstrate that the discrete version of Equation (8) becomes

$$\lambda_i(t_k; \mathbf{x}_i \,|\, V_i = v_i) = 1 - \exp[-\exp(\mathbf{x}_i'(t_k)\boldsymbol{\beta} + \alpha_{t_k})\, v_i] \tag{9}$$

DOI: 10.1002/asmb

If we assume that v_i is Gamma distributed with mean 1 and variance σ^2, the log-likelihood associated with Equation (9) can be expressed as follows:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{N} \log \left[\left[1+\sigma^2 \sum_{k=t_{B_i}}^{t_{C_i}-1-\delta_i} \exp[\mathbf{x}_i'(t)\boldsymbol{\beta}+\alpha_t] \right]^{-\sigma^{-2}} - \delta_i \left[1+\sigma^2 \sum_{k=t_{B_i}}^{t_{C_i}-\delta_i} \exp[\mathbf{x}_i'(t)\boldsymbol{\beta}+\alpha_t] \right]^{-\sigma^{-2}} \right] \tag{10}$$

where the variance σ^2 is to be estimated together with the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

In the following application, we are interested in comparing the classification performances of models (5) and (9). For this purpose, we consider the expected hazard rate, expressed as a function of its variance σ^2. Following Scheike and Jensen [24], we derive such a relationship starting from the conditional probability of default in the interval $[t_k, t_{k+1})$

$$\begin{aligned} P\{T_i \in [t_k, t_{k+1}) | v_i\} &= \lambda_i(t_k|v_i) \prod_{j=1}^{k-1} [1-\lambda_i(t_j|v_i)] \\ &= \lambda_i(t_k|v_i) \prod_{j=1}^{k-1} \exp\{-\exp[\mathbf{x}_i'(t_j)\boldsymbol{\beta}+\alpha_{t_j}]v_i\} \\ &= \lambda_i(t_k|v_i) \exp\left\{ -\sum_{j=1}^{k-1} \exp[\mathbf{x}_i'(t_j)\boldsymbol{\beta}+\alpha_{t_j}]v_i \right\} \\ &= \exp[-F_i(t_{k-1})v_i] - \exp[-F_i(t_k)v_i] \end{aligned} \tag{11}$$

where

$$F_i(t_k) = \sum_{j=1}^{k} \exp[\mathbf{x}_i'(t_j)\boldsymbol{\beta}+\alpha_{t_j}]$$

with the definition $F_i(t_0)=0$. The marginal distribution of default time for the ith company is derived integrating out Equation (11) from the random effect v_i

$$\begin{aligned} P\{T_i \in [t_k, t_{k+1})\} &= \int \exp[-F_i(t_{k-1})v_i]h(v_i)\,\mathrm{d}v - \int \exp[-F_i(t_k)v_i]h(v_i)\,\mathrm{d}v \\ &= \left(\frac{1}{\sigma^2 F(t_{k-1})+1}\right)^{\sigma^{-2}} - \left(\frac{1}{\sigma^2 F(t_k)+1}\right)^{\sigma^{-2}} \end{aligned} \tag{12}$$

where $h(\cdot)$ is the gamma density function.

Dividing Equation (12) by $\mathrm{E}(\exp[-F_i(t_{k-1})v_i])$, we obtain the hazard rate

$$\lambda_i(t_k; \mathbf{x}_i) = 1 - \left(\frac{\sigma^2 F_i(t_{k-1})+1}{\sigma^2 F_i(t_k)+1}\right)^{\sigma^{-2}} \tag{13}$$

DOI: 10.1002/asmb

In the following application, we use SAS statistical software [25] (LOGISTIC procedure with the c-log–log link function) to perform the estimation of model (5). The estimation of the Gamma-frailty model (9) is performed using the PGMHAZ procedure implemented by Jenkins [26] for STATA statistical software.

3. THE DATA

Both the Prentice–Gloeckler's model and its gamma-frailty extension are estimated using data collected over the period 1995–1998 from a sample of small and middle-sized Italian firms provided by Intesa SanPaolo. The data set consists of 25 600 firm-year observations, representing 7711 individual companies belonging to different industries. The proportion of firms experiencing a default from 1996 to 1999 is about 2.5%. In order to verify if the use of a more homogeneous data

Table I. Sample industry breakdown.

Industry	Obs.	Firms	Defaults	% Defaults
Agriculture	114	38	2	0.026
Agricultural intermediation	98	29	2	0.260
Automobiles and motorbikes*	206	63	2	0.026
Clothing*	924	280	10	0.130
Commodities*	647	186	1	0.013
Construction	1364	419	29	0.376
Drugs*	215	64	0	0
Electronics*	543	166	4	0.052
Electrotechnics*	559	167	4	0.052
Financial intermediation	394	126	3	0.039
Food*	1107	339	7	0.091
Households: building materials*	293	90	3	0.039
Households: equipment*	325	97	7	0.091
Households: miscellaneous*	1575	468	13	0.169
Industrial transportation*	139	41	2	0.026
Intermediate goods for chemicals*	364	103	1	0.013
Intermediate goods for clothing*	1660	486	12	0.156
Intermediate goods for construction*	445	131	6	0.780
Intermediate goods for metallurgy*	1806	530	10	0.130
Intermediate goods: miscellaneous	765	223	5	0.065
Leisure products*	132	42	1	0.013
Machinery*	2336	687	17	0.221
Media	157	47	0	0
Packing*	335	97	4	0.052
Personal services	657	209	3	0.039
R&D	61	19	1	0.013
Transportation	784	240	5	0.065
Utilities	125	37	1	0.013
Wholesalers	7259	2223	40	0.519
Widely used goods*	211	64	1	0.013
Total	25 600	7711	196	2.542

*Indicates units belonging to the manufacturing sub-sample.

DOI: 10.1002/asmb

sample could improve the classification performance of the model, we also select a sub-sample of 4101 firms belonging to the manufacturing industries, consisting of 13 826 observations. Table I provides the sample industry breakdown.

Table I shows that there are significant differences in default rates between industries. For example, 'Commodities' and 'Utilities' are industries with a relative low default frequency while 'Wholesalers', 'Intermediate goods for construction,' and 'Construction' are at the opposite end of the scale. These differences reflect well enough the specific risk profiles of the different industries in the years to which the study is referred. Patterns of defaults are shown in Table II.

The geographical distribution of the sample shows a strong concentration of companies in the northwest of Italy as a result of the higher degree of industrial development in this region of the country (Table III).

The set of explanatory variables are based on the official balance sheets and income statements of the sampled firms and on data from the Central Credit Register, which is a department of the Bank of Italy in charge of collecting data on individual loans over €75 000 granted by Italian banks to companies and individuals. These data are compulsorily filed by banks and are made available upon request to individual banks to monitor the total exposure of their customers. In our study, we use the variable CRMARG that is the difference between the total amount of granted loans and the overall exposure recorded by Central Credit Register at the end of each year. Table IV provides the description of the covariates.

Note that the set of coefficients of the temporal dummies is the vector $\boldsymbol{\alpha}=[\alpha_1, \alpha_2, \ldots, \alpha_T]$ whose components $\alpha_{t_k}=\log \int_{t_{k-1}}^{t_k} \lambda_0(u)\,\mathrm{d}u$ are the constants in Equations (5) and (9). Geographical

Table II. Year-by-year default patterns.

	Aggregated sample			Manufacturing sub-sample		
Year	Obs.	Defaults	% Defaults	Obs.	Defaults	% Defaults
1995	6076	—	0	3312	—	0
1996	6553	62	0.804	3541	32	0.780
1997	6925	60	0.778	3541	29	0.707
1998	6046	36	0.467	3712	23	0.561
1999	—	38	0.493	—	21	0.512
Total	25 600	196	2.542	13 825	105	2.560

Table III. Sample geographical distribution.

	Aggregated sample			Manufacturing sub-sample		
Region	Units	% Units	Defaults	Units	% Units	Defaults
Northwest	6228	80.77	146	3331	81.22	76
Northeast	615	7.98	15	378	9.22	13
Center	573	7.43	20	279	6.80	10
South and Islands	295	3.82	15	113	2.76	6
Total	7711		196	4101		105

DOI: 10.1002/asmb

Table IV. Description of the covariates.

Variable		Definition
Financial structure		
EQ		Equity—Intangibles
INT/OPEPR		Interest Expense/(Operating Profit+Depreciation)
LEV		Leverage
DBANK/SALES		Short-term Borrowing from Banks/Sales
FIN/SALES		(Short-term+Long-term Borrowing)/Sales
DEBT/EQ		Debt/Equity ratio
GRICOV		Gross Interest Cover
CRMARG		Margin in Central Credit Register
Profitability		
PROF		Profit (Loss) after Taxation
OPEPR/SALES		(Operating Profit (Loss)+Depreciation)/Sales
PRF/SALES		(Net Income+Depreciation)/Sales
ROI		Operating Profit (Loss)/Total assets
ROA		(Net Income+Interest)/Total Assets
Efficiency		
SAL		Sales
TURN		Sales/Total Assets
SAVAL/CLAB		(Sales−Cost of Sales)/Labor Cost
Liquidity		
ΔCURR/PRF		Current Assets Variation/(Net Income+Depreciation)
STK/CURR		Stocks/Current Assets
LIQ		Current Assets/Current Liabilities
Dummy	Value	If the observation is referred to…
YEARXX	1	year 19xx;
	0	otherwise
NEAST	1	a firm located in northeastern Italy
	0	otherwise
CENTER	1	a firm located in Central Italy
	0	otherwise
SOUTH	1	a firm located in Southern or Insular Italy
	0	otherwise

dummies are included to capture the specific effects associated with the correspondent areas. Some summary statistics of the covariates are reported in Table V.

4. EMPIRICAL ESTIMATES

In this section, we present the estimates resulting from the application of model (5) and its frailty extension (9). A preliminary backward analysis is performed in order to select a limited set of significant covariates and to check the significance of the interaction terms between the temporal dummies reported in Table IV and the other variables. None of these interactions are statistically

DOI: 10.1002/asmb

Table V. Descriptive statistics.

Variable	Mean	Std dev.	Median	1st P.le	99th P.le
EQ*	1521.730	4353.680	850.863	−914.387	9571.500
INT/OPEPR	36.210	67.159	30.366	−290.000	300.000
LEV	9.560	118.627	5.865	−16.217	117.192
DBANK/SALES	18.243	17.784	14.530	0.000	77.295
FIN/SALES	28.650	27.467	22.310	0.000	140.886
DEBT/EQ	247.034	336.863	149.158	−250.000	1000.000
GRICOV	5.431	12.608	2.029	−5.796	80.000
CRMARG*	1241.050	2524.360	723.040	−3100.290	9946.960
PROF*	94.635	558.634	39.767	−1422.320	1549.630
OPEPR/SALES	9.349	8.181	8.350	−9.150	34.732
PRF/SALES	3.272	6.066	2.410	−16.247	20.453
ROI	8.919	9.125	7.934	−14.429	35.281
ROA	16.128	20.326	12.922	−45.000	60.000
SAL*	9485.870	5243.890	8037.880	2833.280	23 734.550
TURN	1.519	0.871	1.337	0.331	4.783
SAVAL/CLAB	197.088	128.487	166.846	19.064	801.469
ΔCURR/PRF	126.393	1037.890	42.212	−4344.540	5000.000
STK/CURR	28.409	19.552	26.049	0.000	81.646
LIQ	90.184	54.208	82.650	17.015	258.075

*Data expressed in € thousands.

Table VI. C-log–log maximum likelihood estimates for the aggregated sample and the manufacturing sub-sample.

	Aggregated sample			Manufacturing sub-sample		
Variables	Coefficient	Std err.	χ^2	Coefficient	Std err.	χ^2
YEAR95	−5.59889	0.163534	1172.1599	−6.098518	0.268783	514.8088
YEAR96	−5.67967	0.153885	1362.2376	−6.125003	0.249869	600.8784
YEAR97	−5.89254	0.188184	980.4867	−5.859651	0.256456	522.0587
YEAR98	−5.66202	0.180462	984.3954	−5.777982	0.266584	469.7692
SOUTH*	0.540197	0.283196	3.6386			
EQ	−0.00012	0.000039	8.8367			
LEV				0.001806	0.000589	9.3961
CRMARG	−0.00028	0.000025	125.8781	−0.000355	0.000051	48.2270
PRF/SALES	−0.05429	0.006729	65.1087	−0.070917	0.006209	130.4566
ROA	−0.01271	0.004120	9.5138			
TURN	−0.54028	0.129564	17.3890	−0.754023	0.236734	10.1449
LIQ	−0.01288	0.001610	63.9779	−0.022626	0.004459	15.1389

Note: The model is estimated by maximizing the log-likelihood function (7). All the parameters are significant at less than 0.5% with the exception of * whose P-value is 0.0565.

significant. Table VI provides the results of maximizing equation (7) for the overall sample and the manufacturing sub-sample with the selected set of explicative variables.

A few brief comments about Table VI seem to be in order. First, we should observe that the model offers a very parsimonious analytical framework that is quite similar for both samples. The

DOI: 10.1002/asmb

negative and significant CRMARG coefficient is as expected and suggests that firms with unused borrowing capacity have a relative lower risk profile. In other words, the more flexibility there is in seeking borrowed funds, the larger the buffer that a firm can use in times of financial difficulties. In the same way, we can interpret the negative sign of EQ for the aggregated sample and the positive sign of LEV for the sub-sample: the higher the proportion of debt financing in a firm's capital structure, the higher its solvency risk.

Second, as might be expected, the probability of default decreases as the liquidity of a firm's assets increases. Firms with relatively large proportions of current assets will either have or will generate soon the necessary cash to meet creditors' claims. This aspect is tightly connected with the role of assets turnover in mitigating the risk of insolvency. The faster assets turn over, the more quickly funds work their way towards cash on the balance sheet.

Third, we should observe an inverse relationship between the risk of default and the level of profitability. The model confirms this: the coefficient of PRF/SALES is negative and highly significant in both samples. Profitable firms ultimately turn their profits into cash and are also usually able to access borrowing more easily and at a relatively lower cost than unprofitable companies. Firms with low or negative profitability must often rely on available cash or additional borrowing to meet financial commitments as they become due. As known, these difficulties can degenerate to a vicious circle as the operating cycle must not only create sufficient cash to supply operating working capital needs but also throw off sufficient cash to service debt. The variable ROA that is significant only for the aggregated sample has a similar rationale to the preceding one.

The last point to note about the estimates is that of the three geographical dummies we considered in our preliminary analysis, only the SOUTH variable in the aggregated sample has a certain statistical significance. This result seems to reflect the existence of marked disparities in the economic development of the two halves of Italy. Indeed, it is common knowledge that Italy shows one of the widest geographical dualisms among other European countries. In some southern areas, the unemployment rate is even four times higher than in the center-north despite recent signs of dynamism. The infrastructure endowment remains far below the national average and organized crime still constitutes a heavy deterrent both for investment and endogenous development. Furthermore, access to external financial resources is generally more difficult and more expensive than in the other areas of the country. As a result, firms located in the south of Italy *ceteris paribus* tend to be more vulnerable than those operating in more developed regions.

Let us now compare the previous results with those provided by the Gamma-frailty model (9) that are reported in Table VII. The size of the estimated log-variance of the gamma distribution relative to its standard error suggests that unobserved heterogeneity is significant in both data sets. Comparing the c-log–log model with its frailty version, we see that the duration dependence parameters that are the coefficients of the temporal dummies are larger in absolute value. This is not a surprising result because not accounting for unobserved heterogeneity induces an under-estimate (over-estimate) of the extent to which the hazard rate declines (increases) with duration. Furthermore, as might be expected, most of the coefficients of other variables are slightly larger in absolute value. To catch this point, it should be noted that the presence of the heterogeneity mitigates the response of the hazards to an infinitesimal change in a single regressor. Thus, the c-log–log model that ignores the unobserved heterogeneity tends to replicate such attenuated effect in the estimation of the regressors' coefficients (for a detailed discussion, see, for example, [27]).

We note that the inclusion of frailty discloses the significance of the SOUTH dummy in both samples. As for the rest, the frailty model shows the same picture as the c-log–log model: a

DOI: 10.1002/asmb

Table VII. Gamma-frailty model maximum likelihood estimates for the aggregated sample and the manufacturing sub-sample.

	Aggregated sample			Manufacturing sub-sample		
Variables	Coefficient	Std err.	χ^2	Coefficient	Std err.	χ^2
YEAR95	−6.247919	0.237032	694.7973	−6.635103	0.363652	332.9070
YEAR96	−6.336123	0.228997	765.5737	−6.847412	0.366765	348.5588
YEAR97	−6.049961	0.220235	754.6265	−6.094553	0.311255	383.3998
YEAR98	−5.717513	0.219105	680.9443	−5.967095	0.329212	328.5301
SOUTH*	0.984529	0.413078	5.6806	1.646178	0.791555	4.3251
EQ*	−0.000112	0.000057	3.8371			
LEV*				0.003427	0.001565	4.7944
CRMARG	−0.000512	0.000064	63.5775	−0.000536	0.000089	35.9073
PRF/SALES	−0.108501	0.019677	30.4039	−0.141748	0.023633	35.9757
ROA	−0.010001	0.005357	3.4851			
TURN	−0.738959	0.162097	20.7823	−1.310090	0.377702	12.0311
LIQ	−0.018363	0.003578	26.3434	−0.023957	0.005916	16.3988
$\log \sigma_\Gamma^2$	1.978602	0.354367	31.1752	2.228773	0.394480	31.9214

Note: The model is estimated by maximizing the log-likelihood function (10). All the parameters are significant at less than 0.5% with the exception of * which are significant at less than 5%.

combination of weak profitability, low liquidity and turnover, and high debt ratios usually spells financial distress.

5. SYSTEM-LEVEL COVARIATES

In this section, we wish to verify the statistical significance of covariates that reflect some of the underlying economic conditions under which the observed companies operate. The main feature of these system-level variables is that, varying over time but not across firms, they can be viewed as the output of a stochastic process that is external to the units under study, provided that the same process is not really affected by the individual contributions of the sampled units (for a detailed discussion, see for example, [28]). It should be noted that the inclusion of one or more system-level variables can really improve the potentialities of the models as predictive tools. A sensible approach could be, for example, to use their expected values to make bankruptcy predictions over more than one year or under different future possible scenarios.

The inclusion of the external variables requires a slight adaptation of the c-log–log model and its frailty extension. Indeed, as observed by Beck *et al.* [29], the coexistence of the time dummies and system-level covariates causes a collinearity problem, with the result that the coefficients of the external variables are always statistically not significant. The presence of system-level covariates thus implies the removal of the dummy time variables and their replacement with the intercept. Assuming that the values of the external variables change over time by the different industries, model (5) becomes

$$\lambda_i(t_k; \mathbf{x}_{is}; \mathbf{z}_s) = 1 - \exp[-\exp(\alpha + \mathbf{x}'_{is}(t_k)\boldsymbol{\beta} + \mathbf{z}'_s(t_k)\boldsymbol{\gamma})] \tag{14}$$

DOI: 10.1002/asmb

and model (9) becomes

$$\lambda_i(t_k; \mathbf{x}_{is}; \mathbf{z}_s | V_i = v_i) = 1 - \exp[-\exp(\alpha + \mathbf{x}'_{is}(t_k)\boldsymbol{\beta} + \mathbf{z}'_s(t_k)\boldsymbol{\gamma})v_i] \quad (15)$$

where α is the intercept, $\mathbf{x}_{is}(t_k)$ is the vector of time-dependent variables for the ith company belonging to the sth industry, $\boldsymbol{\beta}$ is its parameter vector, $\mathbf{z}_s(t_k)$ is the vector of time-dependent system-level covariates referred to the sth industry, and $\boldsymbol{\gamma}$ is its parameters vector. Note that $\lambda_i(t_k; \mathbf{x}_i; \mathbf{z}_s)$ now takes the meaning of instantaneous default risk at time t_k for the ith company belonging to sth industry.

Among possible candidates, we tested the significance of the following system-level variables:

- Economy-wide default rate by industries;
- Total volume of production by industries at 1995 constant prices;
- Total volume of demand by industries at 1995 constant prices;
- Operating profitability (ROI) by industries;
- Labor cost as percentage of the volume of production by industries.

The first variable that has the same rationale as the ANNUALRATE indicator used in [12] to proxy the baseline hazard rate is the mean of the quarterly default rates for Italian loan facilities, distributed by macrobranch of economic activity over the previous year. This indicator that is published by Bank of Italy on a quarterly basis [30] is available in two versions. The first one is defined as the ratio whose denominator is the *amount* of credit used by all the borrowers covered by the Central Credit Register not classified as 'adjusted bad debtors'[‡] at the end of the previous quarter and whose numerator is the *amount* of credit used by such borrowers who become 'adjusted bad debtors' during the same quarter. The second version is constructed using at the denominator the *number* of borrowers not classified as 'adjusted bad debtors' and at the numerator the *number* of such borrowers who become insolvent. Trying with both versions, we matched the yearly means of the sectorial default rates to the respective industry, obtaining a new variable whose values change over time depending on the industry to which the observed unit belongs (i.e. for each industry, we have as many values as the number of the years of observation: four, in our study).

The other four variables that are distributed by industrial macrobranches are surveyed on a yearly basis by Prometeia, an independent Italian research institute.

Including the aforementioned system-level variables, models (14) and (15) are estimated and our findings are that none of these covariates are statistically significant. This result is not surprising, since the length of the period covered by our data is too short to embody any relevant change in the general conditions of the economy or any meaningful variation in the industrial environment. It is certainly possible that samples collected over a longer period of time would lead to different results.

[‡]Bank of Italy defines adjusted bad debts as 'the total loans outstanding when a borrower is reported to the Central Credit Register: (a) as a bad debt by the only bank that disbursed credit; (b) as a bad debt by one bank and as having an overshoot by the only other bank exposed; (c) as a bad debt by one bank and the amount of the bad debt is at least 70% of its exposure towards the banking system or as having overshoot equal to or more than 10% of its total loans outstanding; (d) as a bad debt by at least two banks for amounts equal to or more than 10% of its total loans outstanding.'

DOI: 10.1002/asmb

6. VALIDATION AND PREDICTIVE POWER

In order to get unbiased estimates of the classification errors, we check the predictive ability of the models by means of a holdout sample test. Following Beaver *et al*. [14], both the aggregated sample and the manufacturing sub-sample are randomly divided into two parts, each accounting for 50% of the observations. The first (training) set is used to compute the parameter estimates on the basis of the variable selection performed on the entire sample (see Section 4) and the second (validation) set to compute the individual hazards for each of the observed years (i.e. in case of an uncensored unit, from 1995 to 1998). The yearly hazards are then employed to measure the ability of the c-log–log and gamma-frailty models to correctly predict the possible outcome in each of the following years (i.e. for an uncensored unit, from 1996 to 1999).

To make a comparative analysis with the results provided by a single-period model, we estimate for both samples four logistic regressions based on data collected in each of the observed years. We then use the training and validation sets to assess the classification performance of the logistic regressions. Both for the duration models and for the logistic model, the overall classification performance is the mean of the yearly results.

As usual, we define two types of prediction error. A type I error occurs when a company fails but is predicted to survive; a type II error occurs when a company survives but is predicted to fail. The occurrences of type II misclassifications are counted without considering the cases where a firm defaults two or more years later than predicted by the model. Indeed, in these cases, the models forecast the distress two or more years earlier.

As the models are to be used to make a binary prediction, a cut-off point for the estimated probabilities of default must be chosen depending on the relative costs of types I and II errors. Nevertheless, as we need a concise indicator that is independent of any cut-off, for each model we gauge the area under the receiver operating characteristic (ROC) curve that shows the trade-off between the correct classifications of defaults (sensitivity) and the incorrect classifications of non-defaulted cases (1-specificity). This indicator ranges from a minimum of 0.5, denoting a random model with no predictive power, to a maximum of 1.0 in the case of perfect discrimination. Our results are reported in Table VIII.

Our findings thus show that the duration models perform better than the single-period logistic regression. Furthermore, the classification accuracy is always better in the sub-sample than in the aggregated sample and with the exception of few cut-off points, the gamma-frailty model is almost always more accurate than the c-log–log model.

These results are shown in Figure 1 that depicts the ROC curves for each model.

As noted in the previous section, the SOUTH effect becomes significant in the manufacturing sub-sample only when the model allows for unobserved heterogeneity. Then in comparing the predictive power of the two models, it is important to note that for the sub-sample a different number of variables is used. In order to address whether the improvement in the classification

Table VIII. Area under the ROC curve.

	Logit	C-log–log	Gamma-frailty
Aggregated sample	0.8138	0.8368	0.8420
Manufacturing sub-sample	0.8627	0.8846	0.8931

DOI: 10.1002/asmb

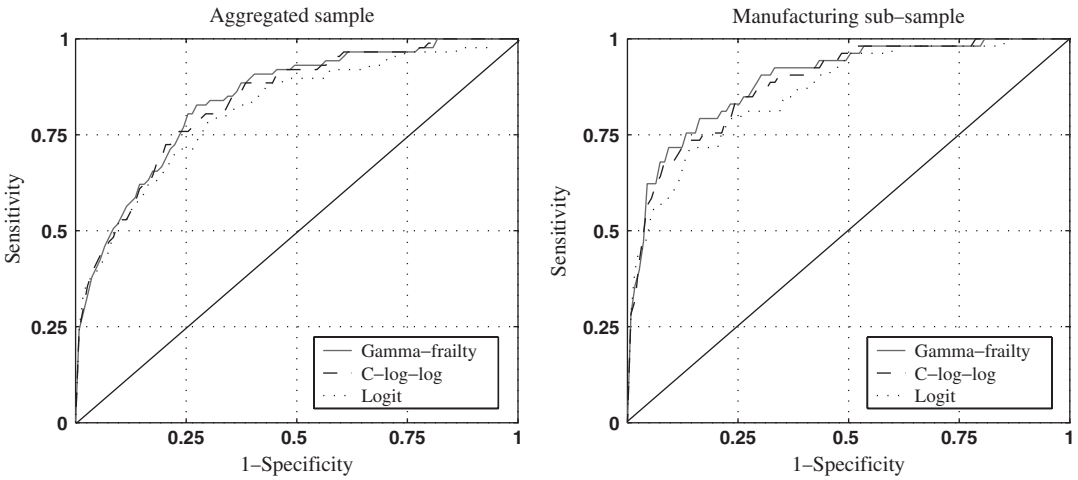

Figure 1. ROC curves.

performance was just an artifact of the use of one more regressor, we estimate the c-log–log model adding the SOUTH variable. The classification results are nearly the same as those obtained when the SOUTH variable is omitted suggesting that the better accuracy can be ascribed to the inclusion of the frailty.

7. SUMMARY AND DISCUSSION

In this paper, the problem of default prediction was addressed by means of a discrete-time survival analysis approach. We proposed the application of the Prentice–Gloeckler's model that is the discrete-time counterpart of the Cox's proportional hazards model and its frailty extension allowing for unobserved heterogeneity. The empirical results, based on a sample of small and middle-sized Italian firms, offer a clear and coherent framework in the analysis of the driving forces behind corporate distress and confirm that survival analysis is well suited to depict the dynamic process that leads to the default. We also discussed the possibility of including system-level covariates and found that none of the variables we have considered are statistically significant. Since this result is almost certainly due to the shortness of the period over which our data were collected, a worthwhile analysis would be to test the significance of these external indicators over a longer horizon and to study the use of their values forecast to predict the default over more than one period after the point at which the regressors are taken or under different future scenarios. This analysis could be the object of a future research.

Compared with the results of a single-period logit regression, we found that the output of our model is richer and more accurate. Not only does it incorporate the synthesis of the evolution of the risk profile of a firm (the sequence of its hazard rates) but it also provides a better performance in terms of prediction accuracy. Taking into account the unobserved heterogeneity can really improve the predictive accuracy of the analysis. We also noted that when passing from the c-log–log model

DOI: 10.1002/asmb

to the frailty model, the parameter estimates grow in magnitude. This result is coherent with the evidence of statistical literature.

Conforming to most of the existing literature, we implemented the frailty model assuming that the unobserved heterogeneity was gamma distributed. As known, this choice is made mainly for mathematical convenience because it leads to a closed-form expression for the likelihood function, avoiding complex numerical integration. It is important to point out, however, that a disadvantage of any approach based on a specified functional form for the distribution of heterogeneity is the possible sensitivity of the conclusions to this specification. In the analysis of our aggregated sample, one would have expected that the frailty model would have led to almost the same accuracy as that obtained in the more homogeneous manufacturing sub-sample. Since the reverse was true, a possible explanation is that the gamma distribution could be not the most appropriate choice for modelling the frailty effect in the population from which our sample was drawn. The use of different distributions for modelling the unmeasured heterogeneity could be the object of further analysis.

ACKNOWLEDGEMENTS

The authors thank an anonymous referee for his/her precious comments and suggestions that really improved the original structure of the paper. The authors also thank the Risk Management Department of Intesa SanPaolo for providing the sample and Prometeia for the data discussed in Section 5. The opinion expressed herein are those of the authors and do not represent the policies or positions of Intesa SanPaolo SpA.

REFERENCES

1. Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 1968; **4**:589–608.
2. Altman EI, Hadelman RG, Narayanan P. Zeta analysis: a new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance* 1977; **1**:29–51.
3. Santomero A, Vinso JD. Estimating the probability of failure for commercial banks and the banking system. *Journal of Banking and Finance* 1977; **10**:185–205.
4. Ohlson J. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 1980; **18**(1):109–131.
5. Zmijewski M. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research* 1984; **22**(Suppl.):59–86.
6. Merton RC. On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* 1974; **29**:449–470.
7. Jarrow R, Turnbull S. Pricing options on financial securities subject to default risk. *Journal of Finance* 1985; **50**:53–86.
8. Duffie D, Singleton KJ. An econometric model of the term structure of interest-rate swap yields. *Journal of Finance* 1997; **52**:1287–1322.
9. Lane WR, Looney SW, Wansley JW. An application of the Cox's proportional hazards model to bank failure. *Journal of Banking and Finance* 1986; **10**:511–531.
10. Luoma M, Laitinen E. Survival analysis as a tool for company failure prediction. *OMEGA—International Journal of Management Science* 1991; **19**(6):673–678.
11. Shumway T. Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business* 2001; **74**(1): 101–124.
12. Hillegeist SA, Cram DP, Keating EK, Lundstedt KG. Assessing the probability of bankruptcy. *Review of Accounting Studies* 2004; **9**:5–34.
13. Chava S, Jarrow RA. Bankruptcy prediction with industry effects. *Review of Finance* 2004; **8**:537–569.

DOI: 10.1002/asmb

14. Beaver WH, McNichols MF, Rhie J. Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies* 2005; **10**:93–122.
15. Deng Y. Mortgage termination: an empirical hazard model with stochastic term structure. *Journal of Real Estate Finance and Economics* 1997; **14**(3):309–331.
16. Ambrose BW, Capone CA. The hazard rate of first and second defaults. *Journal of Real Estate Finance and Economics* 2000; **20**(3):275–293.
17. Stepanova M, Thomas L. Survival analysis methods for personal loan data. *Operations Research* 2002; **50**(2): 277–289.
18. Ciochetti BA, Deng Y, Lee G, Shilling JD, Yao R. A proportional hazards model of commercial mortgage default with originator bias. *Journal of Real Estate Finance and Economics* 2003; **27**(1):5–23.
19. Pennington-Cross A. Credit history and the performance of prime and nonprime mortgages. *Journal of Real Estate Finance and Economics* 2003; **27**(3):279–301.
20. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 1978; **34**:57–67.
21. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman & Hall: London, 1984; 91–111.
22. Meyer BD. Unemployment insurance and unemployment spells. *Econometrica* 1990; **58**(4):757–782.
23. Vaupel JV, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**:439–454.
24. Scheike TH, Jensen TK. A discrete survival model with random effects: an application to time to pregnancy. *Biometrics* 1997; **53**:318–329.
25. SAS Institute Inc. *SAS/STAT User's Guide*, *Version 8*. SAS Institute Inc.: Cary, NC, 1999; 1901–2041.
26. Jenkins SP. Discrete time proportional hazards regression. *Stata Technical Bulletin* 1997; **39**:17–32.
27. Lancaster T. *The Econometric Analysis of Transition Data*. Cambridge University Press: Cambridge, 1990; 65–70.
28. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980; 122–127.
29. Beck N, Katz J, Tucker R. Taking time seriously: time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* 1998; **42**:1260–1288.
30. Bank of Italy. *Statistical Bulletin* (*Table Ref. TDB 30518*). Bank of Italy Website. http://www.bancaditalia.it/ (1 September 2006).

DOI: 10.1002/asmb