Statistica & Applicazioni - Vol. XIX, n. 2, 2021, pp. 161-178
Published online July 2022

THE EVALUATION OF CREDIT RISK USING SURVIVAL MODELS: AN APPLICATION ON ITALIAN SMES

Andrea Marletta*

SUMMARY

The financial literature proposed many contributions to measure the credit risk, in this work a survival approach is proposed to reach this purpose. Having available the survival times for each credit line, the choice was oriented to survival models to evaluate the pathological death of the loan. A survival analysis was conducted on a dataset containing 5322 credits for Italian companies through a Cox model considering some risk factors about both the company and the loan. The selected Cox model led to the identification of risk profiles representing different situations in terms of probability of insolvency.

Keywords: Credit scoring, Survival Analysis, Insolvency risk, Default probability.
DOI: 10.263509/999999_000047
ISSN: 18246672 (print) 2283-6659 (digital)

1. INTRODUCTION AND FRAMEWORK

The banks assumed a central position in the world economy and financial markets. The European Union government promoted support policies in these sectors related to Small and Medium Enterprises (SME) (Cultrera, 2020; Dvoulety, Srhoj and Pantea, 2021). In this context, it is very interesting to observe the behaviour of the Italian credit market regarding SMEs (Angilella and Mazzù, 2015; Nobili, Scalia, Zaccaria and Iannamorelli, 2020).

The analysis of the credit lines here is about in terms of death rate of the credit, intended as insolvency risk of the enterprise and therefore as the probability that the firm will be not able to pay back the loan (Tan and Anchor, 2016; Tan and Floros 2018).

The term "loan death" refers to the beginning of a state of irreversibility, analogously with a natural population. In this context, a loan is born when it is distributed

* Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa - Università di Milano-Bicocca - Piazza dell'Ateneo Nuovo, 1 - Milano (e-mail: andrea.marletta@unimib.it).

and it can follow two different paths: a physiological death, corresponding to the collection of the loan; a pathological death, corresponding to the default of the trustee.

Starting from the analogy between the concept of death in a natural population and a loan, the idea is to use some statistical techniques usually proposed for survival analysis as models able to predict information on the loans. Following this approach, parametric (Weibull) and semi-parametric models (Cox) have been applied for credit data (Jiang, Wang and Zhao, 2019; Pelaez Suarez, Cao Abad and Vilar Fernandez, 2021).

The aim of this paper is to create a useful tool to follow the temporal evolution of credit profiles in order to enact a policy of risk reduction of the loans. This could be done creating different risk scores on the basis of the features of the trustee.

Since the main point of this work is the evaluation of the credit risk, in the next section it seems necessary to introduce some terms typical of the loan market in order to understand more clearly this area of interest.

The paper is organised as follows. Section 2 introduces some definitions related to credit risk, default probability and insolvency in order to have an overview of the topic. Section 3 presents the existing methodology literature in credit scoring and the survival analysis models used for the application. Data and results of the analysis are reported in Section 4. Section 5 is reserved to discussion and final remarks.

2. CREDIT RISK AND DEFAULT

The credit risk is defined as "The risk that an unexpected variation of the credit of an exposed trustee generates a corresponding unexpected variation of the value of the credit position" (Resti, 1999). The credit risk is not only intended as the possibility of insolvency by the trustee, but also the decay of the credit as manifestation of the insolvency risk. The credit risk could be divided into three essential components:

- the expected loss rate, the average value of the distribution of the loss rate;
- the variability of the loss on its average, the risk that the loss is superior to the previous estimated;
- the diversification effect, the decreasing of the unexpected loss rate, suffers when in the same portfolio are included other unexpected loss rates, characterised by imperfect correlation, are included.

Here the attention is focused on the first component, the expected loss rate, usually composed of four key elements: exposure at default (EAD), probability of default (PD), the Loss Given Default (LGD) and the maturity.

The exposure at default is an estimate of the effective value of the loan in case of insolvency. The probability of default is the probability that the obligor is not able to refund the borrowed loan and the matured interests. The LGD is the share of an asset that is lost if a borrower defaults. The maturity is the risk of migration or relegation, namely the danger that a loan could be subjected to decay.

The definition of default is very complex with multiple interpretations. For example, Standard & Poor's (S&P) rating agency defined that there is default when "there is no capability or will of the obligor to keep the financial obligation respecting the original terms" (Ratings, 2016).

S&P defines that "there is insolvency when a payment of interests and/or capital is owed and it is not fulfilled". It is clear that the default appears like an irreversible process, in which it is necessary the enforcement of the guarantees and the probable loss of a part of the borrowed loan (Ratings, 2016).

Credit risk is more complex to measure in comparison with the market risk due to the lower availability of data. Moreover, the distribution of data about credit risk is asymmetric and not so easy to fit to statistical models. Its calculation became one of the issues more examined by economists in recent years. In particular, this is the central topic of the Basel accords.

The Basel Accords refer to the banking supervision Accords (recommendations on banking regulations) – Basel I, Basel II and Basel III – issued by the Basel Committee on Banking Supervision (BCBS). The BCBS is a committee of banking supervisory authorities that was established by the central bank governors of 13 countries in 1974, (Australia, Belgium, Canada, France, Germany, Italy, Japan, Netherlands, Spain, Switzerland, Sweden, United Kingdom and United States). They are called the Basel Accords as the BCBS maintains its secretariat at the Bank for International Settlements in Basel, Switzerland. They are a set of recommendations for regulations in the banking industry and they provide the use of a quantitative and objective method for a more appropriate risk management by the international banking sector (Young, 2011; Penikas, 2015).

Basel I, signed in 1988, is founded on three pillars: the reinforcement of the international banking sector, the homogeneous application among the group of nations, and the circulation of this model outside the countries of the BCBS. Despite some criticisms, Basel I was very successful and more than 100 other countries adopted, even partially, the principles it prescribed. The efficacy with which the principles are enforced varies, even within nations of the Group.

Basel II, signed in 2004, as well as confirming the pillars of Basel I, was founded on three new pillars: minimum capital requirements, supervisory review and market discipline. The first pillar is based on variations applied to the regulatory capital, that is to say the capital in safeguard of the creditors in case of loss due to risk events. For the first time the operational risk and the market risk need to be taken into account, in addition to the credit risk. This pillar leads to the introduction of the definition of rating, defined as a judgement on the actual and future capability of the obligor to refund the borrowed loan. Rating is a quantitative judgement, it is objective, homogeneous and computable using criteria that take into consideration clear evaluation parameters, both quantitative and qualitative.

The second pillar encourages a more strict collaboration among banks and the regulatory authority. They established the general principles that the banks have to follow in order to accomplish a cautious management. In particular, they underline the necessity to endow the banks with a measurement system for the risk management.

The third pillar is about the market discipline, defining contents and communication modalities of the banks on the basis of the risks assumed. This new discipline has to provide the market with all the necessary information to penalise the banks with an insufficient property endowment.

These three pillars are strictly connected themselves, in fact the minimum capital requirements raise the importance of risk management; the Supervisory review raises the collaboration between the bank and regulatory authority; the new market discipline ensures that all could know these new requirements.

Basel III, signed in 2010, was necessary after the world economic crisis, it provides for a substantial reinforcement of the existing patrimonial requirements asking for a higher quality. In particular, this meant the strengthening of the Common Equity and the introduction of two new minimum requirements of liquidity: the LCR (Liquidity Coverage Ratio) and the NSFR (Net Stable Funding Ratio). The first one is an indicator of Liquidity Coverage in the brief period without recourse to the market. The second one is a structural indicator of the financial balance.

3. METHODOLOGY: CREDIT RISK MODELS

The traditional methods to evaluate the credit risk focused their attention on the estimate of the insolvency rate measured in terms of default probability. The set of these statistical models are named credit scoring. This set of models contains all the multivariate models assigning weights to the features of the trustees. These weights are correlated with the importance of the risk factor related to the probability of default.

Credit scoring models are usually divided into four categories: discriminant analysis, probit/logit models, artificial systems and survival analysis.

The economic literature has recently presented many studies facing the problem related to the estimate of the probability of default and the best techniques to use for it. For example, D’Annunzio and Falavigna (2004) examined the frequency distribution of the works related to credit scoring. They showed how 80% of the analysed papers use discriminant analysis or logistic regression. The Artificial Intelligence (AI) systems were introduced in 90’s, but they have been in the new century; they are likely not able to meet the demand of this paper, so the attention will be focused on statistical and econometric models.

3.1 *Survival analysis*

Survival analysis contains all the techniques and statistical models designed to describe and analyse the time events of a statistical unit. It is necessary to identify the unit exposed to risk in relation to these events, the measure of the time duration, and the end of these.

Survival is therefore characterised by a time variable with a start-up and an end-point. In medical research, start-up corresponds to the moment in which an individual is introduced in the experimental study, a clinical treatment or the start of a particular condition for a disease. On the other hand, if the end-point is the death of the patient, data are referred to the time of death. The end-point could not necessarily be the death, but also the end of a pathological state.

For this work, the start-up is the date in which the loan was earmarked and the end-point is represented by the pathological death of the loan (Dirick, Claeskens and

Baesens, 2017; Djeundje and Crook, 2019). The time of interest is the difference between the start-up and the end-point date. All the units at the end of the data collection did not experiment the event of interest have been declared as censored units.

Survival data present some features that require the use of some tailored statistical procedures.

- The first one is distribution – generally survival data are not symmetrically distributed, an histogram based on survival times will tend to be positively asymmetric, this means that all classical models as linear regression are not suitable for these data.
- The second one is the presence of censored data – that is to say, statistical observations that did not experiment the time event. In medical research, these are all patients who are either not dead at the end of the experiment or died for unrelated causes, or retired from the treatment.

Survival analysis can be investigated using non-parametric, parametric or semi-parametric models. The first non-parametric approach considers the estimate of the survival function of a t time variable using the life-tables, obtained dividing the observation period in temporal intervals (Collett, 1994). Suppose that the time period is divided in k intervals, that the $j-th$ interval is from t_j to t_{j+1} and the d_j and c_j are, respectively, the number of deaths and the censored units in the interval. If the censoring process is uniform for the $j-th$ interval, then the number of units at risk is:

$$n_j' = n_j - \frac{c_j}{2} \tag{1}$$

This assumption is called actuarial assumption, it considers the censorship casually appeared in the interval, so the estimate can be approximated using the centre of the interval. The number of units at risk is n_j' and the death probability is $d_j \ / \ n_j$. The probability that a unit survives after the start of the $k-th$ interval is the product of the survival probability from the first to the $k-th$ interval and the estimate of survival function following the life-table approach is:

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j' - d_j}{n_j'} a \tag{2}$$

with $t_k' \leq t < t_{k+1}' = 1,2,\ldots,k$.

One of the limits of this approach is the randomness in the choice of the intervals. To overcome this limit, Kaplan and Meier (1958) proposed a second non-parametric approach. The Kaplan-Meier (K-M) method proposed a univocal rule to build the intervals to divide the time period. The idea is to divide the observation period in intervals identified by the death times. In this way, each interval has a different size and the amount will be dependent by the number of death times.

Suppose $t_1 < t_2 < \ldots < t_r$ with $r < n$ are the death times from a sample of n observations, and the d_j and c_j are the number of deaths and the censored units in the

$j - th$ interval. If the deaths are assumed independent from each other, then the K-M estimate of the survival function will be:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \tag{3}$$

with $t'_k \leq t < t'_{k+1} = 1,2,\ldots,k$.

Non-parametric models are very flexible but they do not guarantee consistent and precise estimates. This is the reason why they are usually used as exploratory tools. For this reason, parametric models have been introduced proposing that the time variable assumes a probability distribution depending on some parameters. This approach allows to determine possible combinations of explanatory variables or risk factors conditioning the risk and the survival function.

Once the probability distribution function $f(t)$ is chosen, then it is possible to obtain the survival function $S(t)$, the hazard risk function $h(t)$ and the cumulative hazard risk function $H(t)$. The survival function $S(t)$ can be obtained as $1 - F(T)$, where $F(T)$ is the cumulative distribution function. The hazard risk function $h(t)$ is the ratio between the probability distribution function $f(t)$ and the survival function $S(t)$. The cumulative hazard risk function $H(t)$ is the total amount of risk for each t cumulated by an individual until t.

The most used probability functions for time variables are the exponential distribution (Epstein, 1958) and the Weibull distribution (Weibull, 1951). For the exponential distribution the formula for the survival function will be equal to: $S(t) = e^{-\lambda t}$. For the Weibull distribution the formula for the survival function will be equal to: $\hat{S}(t) = \exp\left\{-\exp(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi} - \hat{\lambda} t^{\hat{Y}})\right\}$. Using this model, the baseline profile hazard assumes the form of the Weibull distribution $h_0(t) = \lambda\gamma t_i^{\gamma-1}$.

3.2 *Semi-parametric cox model*

Until now, non-parametric and parametric models have been used to model the hazard risk function, with a third alternative being the semi-parametric Cox model. This model was introduced by Cox (1972) and it is defined as semi-parametric because, even if it is based on the hypothesis of proportional hazards, it makes no assumption about the probability distribution for the survival times.

The Cox model assumes the hazard risk function $h_i(t)$ as a product of two components:

$$h_i(t) = h_0(t) * \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) \tag{4}$$

The first component $h_0(t)$ is named baseline hazard risk function, the second one is the exponential of the sum of the combination terms $\beta_p x_{pi}$ extended to all p explanatory variables.

The difference between this approach and the parametric one is that there is no hypothesis about the baseline hazard risk function $h_0(t)$. So the first component involves time variable *t* and not the explanatory variables $X_1, X_2, \ldots, X_p$, on the other hand; the second component involves the explanatory variables without considering time.

The Cox model is named Proportional Hazard (PH) model. The PH assumption entails that HR quantity is constant at a time, or equally, that the hazard for an observation is proportional to the risk of every other observation. From a graphical point of view, this means that there is no change in the vertical distance between two hazard risk curves differing for values of an explanatory variable. In a more flexible approach, it is sufficient that the two hazard risk curves do not intersect each other.

This assumption is rarely verified, for this reason it could be possible to use a stratified or extended Cox model in which interactions between explanatory variables and time variables are included in the model.

3.3 *Survival analysis in credit scoring models*

The application of survival analysis in the economic field could have companies or obligors as statistical units. Even if the analysed units are different, the application of models is similar leading to final considerations in terms of consequences. With regards to the credit market, survival models are based on the definition of the death of the loan, that is to say the entrance in the state of insolvency. So, the random variable time *T*, the loan survival time is measured as the difference between the time in which the obligor obtained the grant for the loan and time in which he becomes insolvent.

There are many contributions in Italy, England and Spain regarding survival analysis on companies. In Italy, Giovannetti estimated the probability of survival of a sample of manufacturing companies using a Cox model. They used as risk factors the firm size, the age of the company, the technological level, the presence on the foreign market and the presence of direct investments (Giovannetti, Ricchiuti and Velucchi, 2007). In North-East region of England, a survival analysis was proposed by Holmes on 931 companies. The probability of survival was estimated using two log-logistic models, the first one for large size enterprises and the second one for the small and medium enterprises. Risk factors included socio-demographic variables (geographic area and firm size), sector-based variables (growth rate and sectorial concentration index), and macro-economic variables (unemployment rate and interest rate) (Holmes, Hunt and Stone, 2010). In Spain, Lopez-Garcia and Puente (2006) analysed a sample of 90,000 companies between 1995 and 2002 using a Cox model. Risk factors involved number of employees, financial ratios and dummy variables related to time and the market sector.

Giambona and Vassallo (2007) built hazard risk profiles for Italian banking credits using a survival model at discrete time on loan data from the Italian banking system from 1985 to 1995. They wanted to observe the hazard risk function in the first decade after the loan assignment. The hazard risk profiles were built using the risk levels compared to the baseline profile through odds ratios. Giambona (2007) also studied the death rate of Italian banking credits using a non-proportional hazard logistic model.

Recent contributions on survival analysis about individuals have been proposed by Xia, He, Li, Fu and Xu (2021), Dirick, Claeskens, Vasnev and Baesens (2022) and

Blumenstock, Lessmann and Seow (2022). Xia *et al.* (2021) proposed the use of gradient boosting decision tree approach. On the other hand, Dirick *et al.* (2022) introduced the use of mixture cure models for competing risks. Blumenstock *et al.* (2022) examined novel machine learning techniques for survival analysis in a credit risk modelling context.

Other contributions applied a mixed model (logistic and Cox) using 27 000 credits with 1-3 years expiration date considering socio-demographic variables about the obligors (age, number of children and house ownership) (Tong, Mues and Thomas, 2012).

Different choices could arise in the nature of variables inserted in the survival model. It is possible to apply a model with only personal variables referred to the obligor like in Andreeva (2006) or Pazdera, Rychnovsky and Zahradnik (2009). Alternatively, Bellotti and Crook (2009) applied a Cox model using only macroeconomic variables.

One of the goals of this work is the use of socio-economic variables related to the company together with variables related to the loan.

4. APPLICATION

The objective of this study is to detect the presence of relationships among the survival time and socio-economic variables related to both the contractor and the loan, using a large dataset in the credit market. After detecting these relationships, it will be possible to create risk profiles of the company based on these variables. As said in the previous section, the time event is the pathological death of the loan, meaning, the entrance of the contractor in a state of insolvency. The survival time is defined as the difference between time in which the loan was obtained and time in which it became insolvent.

Data used for this analysis refer to credit lines of an Italian broker until 2011. This dataset is composed of 5322 credits requested by companies. First of all, it could be interesting to compute the percentage of insolvent loans over to the total. Only 192 credits entered in the insolvency state, so the insolvency rate could be easily computed:

$$\text{insolvency rate} = \frac{\text{insolvency loans}}{\text{total credits}} = \frac{192}{5322} = 3.6\% \tag{5}$$

The variables considered as potential risk factors for the insolvency state are, in brackets the variable type:

- loan amount (numeric): the total amount of funded loan;
- age of the company (numeric): measured in years ;
- year of distribution of the loan (numeric): the year in which the loan was funded;
- quality of life in the city where the company is located: an indicator on 3 levels of the quality of life of the municipality where the company is located;
- legal status of the company (categorical): juridical form of the company;
- form of financing (categorical): type of loan, mortgage or overdraft;
- activity market (categorical): the activity sector of the company according to the Ateco Istat classification.

In Table 1, the frequency distribution for the risk factors are shown. Some operations of aggregation and encoding were applied on the original variables in order to better organise the data. The loan amount was dichotomised in low (< 20000 €) and high amount (≥ 20000 €). The age of the company was categorised in three slots: young (born between 2001 and 2011), medium-age (born between 1991 and 2000) and old companies (born before 1991).

The year of the distribution of the loan was distinguished between crisis age (2009-2011) and before (1994-2009). About the city where the company is located, a tailored indicator was built based on socio-economic data following the guidelines of OECD in constructing composite indicators (Nardo, Saisana, Saltelli, Tarantola, Hoffman and Giovannini, 2005). The result of this indicator gave three slots, identified as high, medium, and low quality of life. The only distinction about legal status of the company regarded the sole trader and the companies. The form of financing was distinguished between defined expiration of the loan (mortgages) and non-defined loan (bank overdraft facilities). Finally, the activity market was divided into three macro-categories: agricultural, industry and services. The majority of the loan has an amount higher than 20000€ (75.2%). Half of the credits regarded the young companies (49.8%). Credits are about equally distributed between crisis years and pre-crisis

TABLE 1. - *Risk factors frequency distributions for involved companies in Italy*

Risk factor	Frequency	Percentage (%)
Loan amount		
0 – 20000 €	4003	75.2%
20000+ €	1319	24.8%
Age of the company		
Young	2650	49.8%
Medium-age	1783	33.5%
Old	889	16.7%
Loan year		
1994-2008	3057	57.4%
2009-2011	2265	42.6%
Life quality of the city		
Low	2978	56.0%
Medium	573	10.7%
High	1771	33.3%
Legal status		
Sole trader	3037	57.1%
Companies	2285	42.9%
Form of financing		
Bank overdraft	2227	41.8%
Mortgage	3095	58.2%
Market activity		
Agriculture	143	2,7%
Services	3706	69.6%
Industry	1473	27.7%

period (42.6% vs 57.4%). The choice to use a composite indicator to aggregate all the municipalities gave good results, with 1 company over 3 active in municipalities with high quality of life. A substantial balance is present between sole trader and companies (57.1% vs 42.9%) and between mortgages and bank overdraft facilities (41.8% vs 58.2%). With regards to the market activity, the prevalent one is the services sector (69.6%).

To measure the credit risk, a semi-parametric Cox model was adopted as described in the previous section. Using the STATA software (StataCorp., 2021), estimated coefficients $\hat{\beta}$, the hazard ratio relative risk $\exp(\hat{\beta})$ and the standard error $err.std(\hat{\beta})$ for all risk factors are presented in Table 2.

Applying this model, it will be possible to create the profile of an enterprise with higher or lower insolvency risk. The frequency distribution allows one to build the baseline profile of, the typical company asking for a loan; a young sole trader, operating in the agricultural market and in a low quality of life municipality, asking for a 0 – 20000 € bank overdraft facility before the crisis. The underlined levels of risk factors represented the baseline profile, the $\hat{\beta}$ coefficients could be interpreted in terms of significance and in terms of insolvency risk considering $\exp(\hat{\beta})$.

TABLE 2. - *Estimates of full semi-parametric Cox model on involved companies in Italy*

Risk factors	$\hat{\beta}$	$\exp(\hat{\beta})$	$err.std(\hat{\beta})$	T test	P-value
Amount					
0 – 20000 €	–				
20000+ €	-0.585	0.557	0.231	-2,53	0.011
Age of the company					
Young	–				
Medium-age	-0.685	0.504	0.166	-4,12	0.000
Old	-1.320	0.267	0.277	-4,78	0.000
Loan year					
1994-2008	–				
2009-2011	-0.573	0.564	0.242	-2,37	0.018
Life quality of the city					
Low	–				
Medium	1.224	3.401	0.213	5,75	0.000
High	0.628	1.874	0.162	3.89	0.000
Legal status					
Sole trader	–				
Companies	-0.373	0.689	0.163	-2,29	0.022
Form of financing					
Bank overdraft	–				
Mortgage	0.075	1.078	0.153	0.49	0.624
Market activity					
Agriculture	–				
Services	-0.755	0.470	0.463	-1.63	0.103
Industry	-0.607	0.545	0.473	-1.28	0.200

TABLE 3. - *Model choice using Likelihood Ratio test on involved companies in Italy*

Model	-2LogL	D	Df	P-Value
Null	3020.713	0.000	-	
Form	3020.583	0.130	1	0.718
Activity	3019.732	0.981	2	0.612
Status	3015.544	5.169	1	0.023
Year	3014.478	6.235	1	0.013
Amount	3007.862	12.851	1	0.000
City	2987.654	33.059	2	0.000
Age	2980.554	40.159	2	0.000
Age+City	2950.086	30.469	2	0.000
Age+City+Amount	2939.313	10.772	1	0.001
Age+City+Amount+Status	2932.943	6.370	1	0.012
Age+City+Amount+Status+Year	**2926.740**	**6.203**	**1**	**0.013**
Complete	2923.753	2.987	3	0.394

It is possible to note that variables Form of financing and Market activity showed non-significant coefficients, for this reason in Table 3, the model selection procedure using the *LR*-test and the test statistic $-2\log\hat{L}$ is shown.

The model selection procedure confirmed the non significance of variables Form of financing and Market activity. So, it is necessary to estimate a new Cox model without these risk factors. Results for this model are displayed in Table 4.

The estimated parameters could be interpreted in terms of sign and value. The positive sign implies higher risk in comparison with the baseline level. The HR exp($\hat{\beta}$) indicates how much the risk increases for the enterprise with that level of the explana-

TABLE 4. - *Estimates of selected semi-parametric Cox model on involved companies in Italy*

Risk factors	$\hat{\beta}$	$\exp(\hat{\beta})$	$err.std(\hat{\beta})$	T test	P-value
Amount					
0 – 20000 €	–				
20000+ €	-0.556	0.573	0.227	-2.43	0.014
Age of the company					
Young	–				
Medium-age	-0.671	0.511	0.165	-4.08	0.000
Old	-1.305	0.271	0.276	-4.74	0.000
Loan year					
1994-2008	–				
2009-2011	-0.562	0.570	0.241	-2.33	0.020
Life quality of the city					
Low	–				
Medium	1.220	3.386	0.213	5.74	0.000
High	0.598	1.818	0.160	3.74	0.000
Legal status					
Sole trader	–				
Companies	-0.384	0.681	0.161	-2.40	0.017

tory variable. For example the value $\hat{\beta} = -0.556$ for 20000 +€ loans means that who asks for a lower credit has a lower risk compared to who asks for a $0 - 20000$€. The value $\exp(\hat{\beta}) = 0.573$ means that there is a -43% $(\exp(\hat{\beta}) - 1)$ of risk for these enterprises.

This hypothesis could be sensed, since who asks for a higher amount is due to provide harder guarantees. The young enterprise is more at risk in comparison to the others. The old companies are less at risk (-73%) compared to the medium-age (-49%).

The negative coefficient related to the year loan was requested, indicates that the risk decreases (-43%) when the credit has been distributed over the course of the previous three years. This fact appears in countertrend with the start of the world economic crisis, although it is possible that Italy experienced a delayed effect in the following years.

The variable about quality of life in the municipality where the loan has been distributed shows that cities with a high quality life are more at a risk about twice than the baseline level ($\exp(\hat{\beta}) = 1.818$).

Finally, in relation to the legal status of the enterprise, the sole trader is more at risk compared to the companies (-32%). This is probably due to the sole trader not having the financial-economic solidity of the companies.

To evaluate the risk trend during time, it is possible to graphically represent some profiles, created on the basis of the values of the risk factors. In Figure1 three profiles

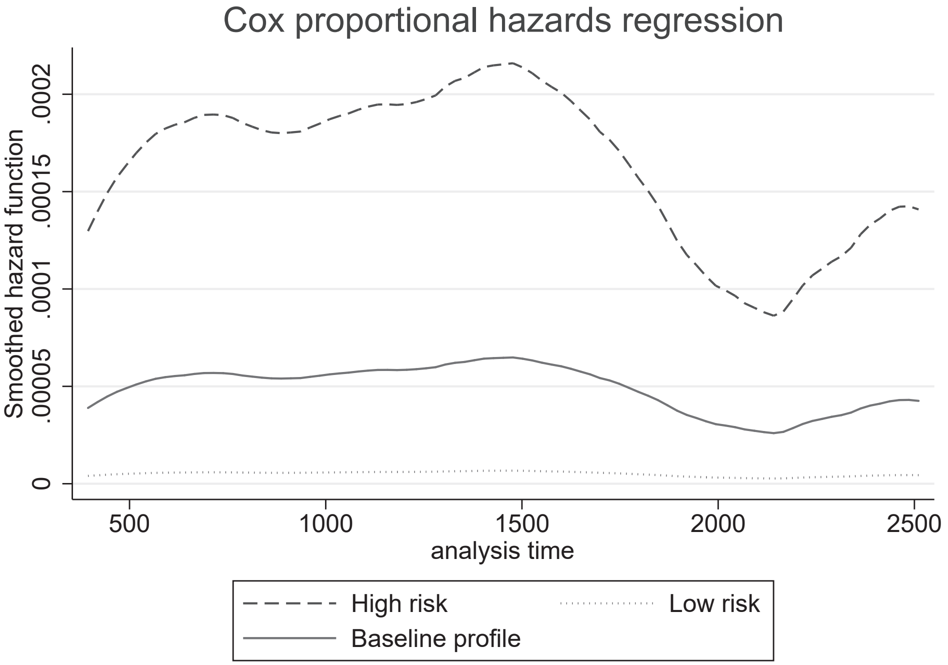

FIGURE 1. - *Hazard function profiles (Model A) on involved companies in Italy*

were displayed, the baseline profile described before, the highest risk profile and the lowest risk one.

The highest risk profile is composed of young sole trader asking for a loan with a $0 - 20000$€ amount in the years before the crisis, located in a medium quality of life municipality.

The lowest risk profile is composed by an old company asking for a loan with a $20000 +$€ in the 2009-2011 period, located in a low quality of life municipality.

The baseline profile curve is in the middle between the highest risk profile and the lowest risk one, the graph shows a similar trend among the three curves. There is an increasing trend until the peak at 1500 survival days (more than 4 years), later a rapid decrease until the minimum at about 2100 days (almost 6 years) and finally a slight comeback until the maximum survival time at 2683 days. Due to the high numbers of censored data, the estimate of the hazard function has very low values, in the order of 10^{-4}.

Another very interesting tool is the risk score. Instead of summarising the effect of a single risk factor, the risk score is able to consider all factors simultaneously. This quantity is very useful when the selected model involves many explanatory variables.

Each risk score compares the level of the risk function with the baseline profile, so it produces a measure of risk in relative and not absolute terms. In Cox model, the risk score is equal to:

$$risk\ score_i = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) \tag{6}$$

As it is possible to note from the formula, the risk score does not depend on the single survival time, so they could be computed for both insolvent and safe companies. Like the hazard ratios, the risk scores are positive values too. The higher the values, the higher is the risk for the company. For the baseline profile the risk score is equal to 1, since all risk factors are equal to 0.

The risk score could also be useful in predictive terms, since when an obligor or a company asks for a loan, their characteristics are known and the risk score can be attributed a-priori.

In Table 5, the risk score is computed for the three profiles represented in graph 1. The risk score for the high risk profile is equal to 3.386, while for the low risk one it is very close to 0.

The only hypothesis of the Cox model is the hazard proportionality, that is, the *HR* has to be constant during time. Table 6 shows the Proportionality Hazard (PH) test for

TABLE 5. - *Risk scores for different profiles on involved companies in Italy*

Profile	High risk	Baseline	Low risk
Amount	$0 - 20000$ €	$0 - 20000$ €	20000+ €
Age of the company	Young	Young	Old
Loan year	1994-2008	1994-2008	2009-2011
Life quality of the city	Medium	Low	Low
Legal status	Sole trader	Sole trader	Companies
Risk score	3.386	1	0.060

TABLE 6. - *Proportionality hazard test for selected Cox model on involved companies in Italy*

Variable	Chi-q	Df	P-value
Amount			
$0 - 20000$ €			
$20000+$ €	0.150	1	0.6990
Age of the company			
Young			
Medium-age	2.950	1	0.0858
Old	0.960	1	0.3272
Loan year			
1994-2008			
2009-2011	1.070	1	0.3018
Life quality of the city			
Low			
Medium	0.050	1	0.8252
High	2.930	1	0.0870
Legal status			
Sole trader			
Companies	3.180	1	0.0746
Global test	11.910	7	0.1035

Cox model. Using this test, if p-value is more than 0.05, the hypothesis of proportional hazards is confirmed refusing the hypothesis to reject the Cox model. Otherwise, the estimate of a stratified Cox model is necessary.

The global test shows a p-value equal to 0.1035 so the risk proportional assumption is confirmed. Therefore, it is possible to refuse the hypothesis to reject the Cox model.

A possible alternative is represented by a parametric approach and the use of a Weibull regression. The model obtained using Weibull distribution is very similar to the Cox one in terms of coefficients for the risk factors. Moreover, since a probability distribution is assumed for the survival time *t*, it is possible to estimate the form parameter $\hat{\gamma}$. For the Weibull model $\hat{\gamma} = 1.492$ with a standard error equal to 0.092. This means that it is significantly different from 1 and it is not possible to hypothesise an exponential distribution for *t*. Since $\hat{\gamma} > 1$, the hazard risk function will be increasing. This represents the only big difference between the two models. To choose the best

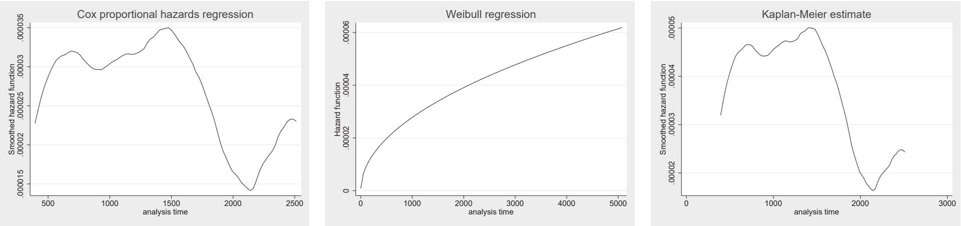

FIGURE 2. - *Hazard risk curves for three approaches: Cox (a), Weibull (b) and Kaplan-Meier (c)*

one, a possible solution could be represented by displaying the hazard risk function of a third approach, the non parametric one. In Figure 2, hazard risk functions are displayed for the 3 approaches, Cox model (a), Weibull (b) and Kaplan-Meier (c). Since the Kaplan-Meier approach shows a hazard curve more similar to the Cox one, it seems plausible to hypothesise a non-increasing trend for the hazard and choose the Cox model.

5. CONCLUSION AND FUTURE RESEARCH

The main aim of this paper was to evaluate the credit risk for loans distributed in Italy. To reach this, a survival analysis was conducted on a dataset containing 5322 credits for Italian companies in 2011.

Having available the survival times for each credit line, the choice was oriented to survival models to evaluate the pathological death of the loan. After the data description and some aggregation and coding about variables, a Cox model was fitted to the final dataset.

The risk factors included in the model are: the requested amount, the age of the company, the year of the loan, the quality of life in which the company is located, and the legal status of the enterprise.

The selected Cox model led to the identification of two risk profiles representing the best and worst situation in terms of probability of insolvency.

The profile with the highest probability of insolvency is composed by a young sole trader asking for a loan with a $0 - 20000$€ amount in the years before the crisis, located in a medium quality of life municipality. This profile corresponds to 57 credit lines in the dataset and 15 over 57 became insolvent, with an insolvency rate equal to the 26.3% against the 3.6% of the total insolvency rate of the full dataset. Moreover, using a risk score indicator, this profile has a risk more than three times bigger than a baseline profile.

On the other hand, the lowest risk profile is composed of an old company asking for a loan with a $20000 +$€ in the 2009-2011 period, located in a low quality of life municipality. The same risk score indicator shows a value very close to 0, therefore a so-composed company can be judged as a safe trustee.

This could represent a very satisfactory result, since the selected model seems to be useful to detect the characteristics of the explanatory variables leading to the insolvency state. This model could be used in predictive terms, in fact through the use of the risk score, a bank could know a-priori if the obligor or the company could be classified as a possible insolvent.

Future researches could generalise this approach using other available risk factors or comparing the Cox model with other techniques different from the survival analysis.

ACKNOWLEDGEMENTS

The author thanks Prof. Francesca Giambona for the inclusion in this research project and the helpful advice.

REFERENCES

Angilella S., Mazzù S. (2015). The financing of innovative SMEs: A multicriteria credit rating model. *European Journal of Operational Research*, **244**(2), 540-554.

Andreeva G. (2006). European generic scoring models using survival analysis. *Journal of the Operational research Society*, **57**(10), 1180-1187.

Bellotti T., Crook J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, **60**(12), 1699-1707.

Blumenstock G., Lessmann S., Seow H.V. (2022). Deep learning for survival and competing risk modelling. *Journal of the Operational Research Society*, **73**(1), 26-38.

Collett D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall: London.

Cox D.R. (1972). Regression models and life-tables (with discussion). *Journal of Royal Statistical Society, Series B* **74**, 187-220.

Cultrera L. (2020). Evaluation of bankruptcy prevention tools: evidences from COSME programme. *Economics Bulletin*, **40**(2), 978-988.

Dvouletý O., Srhoj S., Pantea S. (2021). Public SME grants and firm performance in European Union: A systematic review of empirical evidence. *Small Business Economics*, **57**(1), 243-263.

D'Annunzio N., Falavigna G. (2004). *Modelli di analisi e previsione del rischio di insolvenza: una prospettiva delle metodologie applicate.* Ceris-Cnr.

Dirick L., Claeskens G., Baesens B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, **68**(6), 652-665.

Dirick L., Claeskens G., Vasnev A., Baesens B. (2022). A hierarchical mixture cure model with unobserved heterogeneity for credit risk. *Econometrics and Statistics*, **22**, 39-55.

Djeundje V.B., Crook J. (2019). Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research*, **275**(1), 319-333.

Epstein B. (1958). *The exponential distribution and its role in life testing (No. TR-2).* Wayne State University of Detroit.

Giambona F., (2007). Mortalità dei crediti bancari italiani: Altre evidenze empiriche, *Rivista Minerva Bancaria*, **5**, 1-16.

Giambona F., Vassallo E., (2007). Profili di rischio dei crediti bancari italiani: un'analisi per generazioni di finanziamenti, *Rivista Minerva Bancaria*, **2**, 9-46.

Giovannetti G., Ricchiuti G., Velucchi M., (2007). Dimensione, innovazione e internazionalizzazione: un'analisi di sopravvivenza delle imprese italiane, *Rapporto sul commercio estero*, 386-391.

Holmes P., Hunt A., Stone I. (2010). An analysis of new firm survival using a hazard function. *Applied Economics*, **42**(2), 185-195.

Kaplan E.L., Meier P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**(282), 457-481.

Jiang C., Wang Z., Zhao H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, **277**(1), 20-31.

Lopez-Garcia P., Puente S., (2006). *Business demography in Spain: Determinants of firm survival.* Banco de Espana.

Nobili S., Scalia A., Zaccaria L., Iannamorelli A. (2020). Asymmetric Information and Corporate Lending: Evidence from SME Bond Markets. *Bank of Italy Temi di Discussione (Working Paper)* No, 1292.

Nardo M., Saisana M., Saltelli A., Tarantola S., Hoffman A., Giovannini E. (2005). *Handbook on constructing composite indicators: Metodology and user guide*. OECD Statistics Working Papers n. 2005/03 OECD Publishing, Paris.

Pazdera J., Rychnovský M., Zahradnik P. (2009). *Survival analysis in credit scoring. In Seminar on Modelling in Economics (Vol. 1).*

Pelaez Suarez R., Cao Abad R., Vilar Fernandez J.M. (2021). Probability of default estimation in credit risk using a nonparametric approach. *TEST*, **30**(2), 383-405.

Penikas H. (2015). History of banking regulation as developed by the Basel Committee on Banking Supervision 1974-2014. *Estabilidad financiera*, **28**, 9-47.

Resti A. (1999). *La gestione del rischio di credito con modelli di derivazione attuariale: il caso di CreditRisk.* Fondo interbancario di tutela dei depositi.

Ratings S.P.G. (2016). S&P global ratings definitions. URL: https://www.standardandpoors.com/en_US/web/guest/article/-/view/sourceId/504352

StataCorp. (2021). *Stata Statistical Software: Release 17.* College Station, TX: StataCorp LLC.

Tan A.Y., Anchor J.R. (2016). Stability and profitability in the Chinese banking industry: evidence from an auto-regressive-distributed linear specification. *Investment Management and Financial Innovations*, **13**(4), 120-128.

Tan Y., Floros C. (2018). Risk, competition and efficiency in banking: Evidence from China. *Global Finance Journal*, **35**, 223-236.

Tong E.N., Mues C., Thomas L.C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, **218**(1), 132-139.

Xia Y., He L., Li Y., Fu Y., Xu Y. (2021). A dynamic credit scoring model based on survival gradient boosting decision tree approach. *Technological and Economic Development of Economy*, **27**(1), 96-119.

Young K. (2011). The Basel Committee on Banking Supervision. In T. Hale and T. Held (Eds.) *The Handbook of Transnational Governance: Institutions and Innovations* (pp. 39-45). Willy..

Weibull W. (1951). A Statistical Distribution Function of Wide Applicability. *Journal of Applied Mechanics*, **18**, 293-297.