Statistica & Applicazioni - Vol. XX, n. X, 2023, p. 1-17
First published online July 2023

A COMPARISON OF VARIABLE SELECTION METHODS IN COMPETING RISKS MODELS FOR BUSINESS FAILURES

Francesca Pierri *
Marialuisa Restaino**
Chrys Caroni***

SUMMARY

The statistical techniques that are applied to the analysis and prediction of the failure of business enterprises include regression modelling of hazard functions. Because a large number of financial variables are available as potential predictors, it is necessary to choose an appropriate variable selection procedure in this modelling. Furthermore, alternative models, for cause-specific and subdistribution hazards, are available when different modes of failure are being examined as competing risks. We demonstrate the application of both types of hazards to the study of failures from bankruptcy, liquidation or dissolution among 75479 manufacturing firms in seven European countries in 2000-2018, and we examine the results of variable selection by stepwise and lasso methods for both approaches. All analyses resulted in successful prediction, with areas under the ROC curve of 0.90 or just below, but the lasso approach achieved this with a smaller number of variables selected.

Keywords: Variable Selection, Lasso, Cause-Specific Hazards, Subdistribution Hazards, Competing Risks, Business Failures.
DOI: 1026350/999999_000053
ISSN: 1824-6672 (print) 2283-6659 (digital)

1. INTRODUCTION

The evolution of a firm's activities may eventually lead to its exit from the market, which can happen for several reasons (e.g. bankruptcy, liquidation, merger and acquisition). The various modes of exit may be induced by different factors, with important implications and consequences for the stakeholders and, in general, for the whole economy (Schary, 1991; Harhoff, Stahl and Woywode, 1998). As a consequence, exploring the determinants related to the different types of exit can be particularly relevant and meaningful. Our purpose in this paper is to contribute to discussion on this field, in the context of the illustrative analysis of a large set of data on the survival or failure of European firms. These data will be described in detail in Section 3 below.

Models for analysing firms' survival and for studying the factors associated with

* Dipartimento di Economia, Università di Perugia - Via Pascoli 20 - 06123 PERUGIA (email: francesca.pierri@unipg.it).
** Dipartimento di Economia e Statistica - Università di Salerno - Via Giovanni Paolo II 132 - 84084 FISCIANO (SA) (email: mlrestaino@unisa.it).
*** Department of Mathematics - National Technical University of Athens - Zografou Campus - 15780 ATHENS (✉ e-mail:ccar@math.ntua.gr).

their failure have drawn particular attention from both academics and practitioners over the years. Up-to-date reviews have been presented recently by Severin and Veganzones (2018) and Veganzones and Severin (2021). Early work, starting with the paper by Altman (1968), focused on failure as a binary outcome (i.e. failing versus surviving) and thus fitted binary models, such as logit, probit, discriminant analysis, survival analysis and so on (Ohlson, 1980; Zmijewski, 1984; Lennox, 1999; Shumway, 2001; Brabazon and Keenan, 2004; Amendola, Restaino and Sensini, 2011; Pierri and Caroni, 2017). But since the paper by Rommer (2004), attention has also been paid to different types of financial distress. The data analysed in the present paper record three modes of failure. The models applied to examine and estimate the effects of explanatory variables when more than one failure mode is considered include extended versions of logistic regression (i.e. the mixed logit, multinomial error component logit and nested logit models) (Headd, 2003; Jones and Hensher, 2004; Hensher and Jones, 2007; Jones and Hensher, 2007; Dakovic, Czado and Berg, 2010) and, in survival analysis, competing risks models (Dyrberg, 2004; Rommer, 2005; Chancharat, Tian, Davy, McCrae and Lodh, 2010; Esteve-Pérez, Sanchis-Llopis and Sanchis-Llopis, 2010; Amendola, Restaino and Sensini, 2014; Amendola, Restaino and Sensini, 2015; Caroni and Pierri, 2020). However, little has been said about the similarities or dissimilarities among the factors that are associated with the different modes of failure (Chancharat *et al.*, 2010; Esteve-Pérez *et al.*, 2010), or about the difference in these factors among countries (Altman, Iwanicz-Drozdowska, Laitinen and Suvas, 2017; Tian and Yu, 2017).

All the previous contributions have focused on estimating the probability of a firm's failure and interpreting the effects of covariates on it, without considering how to select the most relevant covariates for inclusion in the model. In fact, since each failure mode could be associated with different factors and the number of potential predictors may be large, it is essential to find efficiently a subset of variables which aids the identification of the determinants of financial risks. Thus a recent review regards the issue of variable selection as 'a crucial step in building a model' (Veganzones and Severin, 2021, p. 213). The optimal choice of those financial indicators that influence each cause of leaving the market should allow better risk assessment and model interpretation. Several variable selection techniques (stepwise regression, lasso, LARS, and others) to predict the failure of firms have been suggested in the literature (Amendola *et al.*, 2011; Amendola *et al.*, 2015; Tian and Yu, 2017).

Starting from this background, in this paper we will consider a Cox proportional hazards competing risks model with three mutually exclusive modes of exit, corresponding to the different causes of a firm's failure – bankruptcy, dissolution and liquidation (winding-up) – defined in Section 3 below. The occurrence of any one of these events ends the firm's participation in the study. The effects of micro-economic indicators and firm-specific variables on the risks of failure by the different routes are examined in the competing risks hazard model. Working within the survival analysis framework, hazard models, unlike discrete outcome models (logit, probit), allow us to account for both whether and when an event occurs. Moreover, the competing risks model provides information regarding possible differences in the effects of each variable

across the multiple states of financial distress. In this paper, we will present and apply the two main approaches to hazard modelling of competing risks data: the cause-specific hazard and subdistribution hazard models. The difference between them in terms of executing the analysis consists in the definition of the risk set, as will be clarified in Section 2 below. The models are fitted to a large set of data on European firms from a publicly available database. This is described in Section 3 below.

Moreover, in order to choose the best set of possible predictors for each event of interest, we will apply both stepwise and lasso variable selection procedures in the regression modelling, these being the two main procedures that have been employed in this context. Thus, we will be able to compare the relevant covariates not only between the two model specifications, but also between the two variable selection procedures.

The rest of the paper is structured as follows. In Section 2, the statistical methods and the variable selection techniques implemented are briefly presented. Section 3 describes the data used in the empirical analysis. Section 4 introduces and discusses the main results of the statistical analysis.

2. METHODOLOGY

2.1 *Competing risks model*

As stated in Section 1, we will examine business failures by applying techniques of survival analysis. In particular, since there are several possible ways of ceasing business activity, we opt for the competing risks model in this application (Crowder, 2001; Fine and Gray, 1999). This extension of the simple mortality model for survival data takes into account multiple events of interest. In this Section, we introduce the notation and recall the basic functions employed in the models considered.

We denote by $\tilde{T}$ and C the actual time until the event of interest occurs and the noninformative censoring time, respectively. $\tilde{T}$ is observable only if failure occurs while the firm is under study. Let $T = \min(\tilde{T}, C)$ be the recorded time until failure or censoring, whichever occurs first, and define the indicator function $\delta = I(T \leq C)$, which is equal to 1 if failure is observed, and zero otherwise. We denote by D the modes of failure, numbered from 1 to K. (In our application, $K = 3$.) Thus, the observed data are given by a set of pairs (T, δ) and also, if $\delta = 1$, the value of D.

Interest in the competing risks model lies in analysing the joint distribution of T and D. Two different approaches are proposed in the literature, corresponding to the construction of alternative forms of the hazard function: *cause-specific hazard* and *subdistribution hazard* (Putter, Fiocco and Geskus, 2007; Austin, Lee, D'Agostino and Fine, 2016). The *cause-specific hazard function* at time t is the instantaneous rate of failure from cause k at that point, among the firms that are still surviving:

$$h_k(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t \cap D = k | T > t]}{\delta t}, \qquad k = 1, \ldots, K. \tag{1}$$

A major difficulty related to the use of the cause-specific hazard function is that it does not give correctly the cumulative incidence of failure from a particular cause,

$F_k(t) = P(T \leq t \cap D = k)$, because surviving up to time t means that the firm must so far have avoided not only this cause of failure but also all the other causes as well. Therefore, one cause cannot be considered in isolation from the others; it is not possible for the firm to fail from cause k at time t unless it has survived the threat of other causes up until t.

On the other hand, the cumulative incidence of failure cause k can be estimated correctly from the *subdistribution hazard function* introduced by Fine and Gray (1999):

$$h_k^s(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t \cap D = k | T > t \cup (T \leq t \cap D \neq k)]}{\delta t}, \qquad (2)$$

$k = 1, \ldots, K$ by means of the relation

$$h_k^s(t) = -\frac{d}{dt} \ln (1 - F_k(t)). \qquad (3)$$

The subdistribution hazard is the instantaneous rate of failure from cause k at time t among firms that up to then have not experienced an event of type k, although they may have experienced a different event. In contrast, the cause-specific hazard in (1) refers to the failure of firms that have not yet experienced an event of any type. Thus, the main difference between the two hazard functions lies in the risk set, i.e. the set of firms that are counted as being at risk at time t. For the cause-specific hazard in (1), the risk set consists of all firms that still survive at time t. For the subdistribution hazard in (2), the risk set also includes the firms that have already failed before time t from causes other than cause k (despite the fact that, obviously, they cannot in fact fail again). This departure from reality is a major obstacle to the interpretability of the model. However, as remarked above, it is what allows the correct calculation of the cumulative incidence $F_k(t)$, which is a quantity of central interest.

Both versions of the hazard function under competing risks can be extended to include the multiplicative effects of covariates. This is usually done exactly as in Cox's semi-parametric proportional hazards model (Cox, 1972)

$$h_k(t|Z_k) = h_{k,0}(t) \exp\{\beta_k^T Z_k(t)\} \qquad (4)$$

(and similarly for $h_k^s(t)$), where $h_{k,0}(t)$, the baseline hazard of cause k, does not need to be specified explicitly, $Z_k(t)$ is a vector of covariates potentially affecting the hazard for the cause k at time t, and the vector of unknown regression coefficients β_k represents the covariate effects on cause k which are to be estimated. Since the same variables could have different effects on the various risks, it is reasonable to assume separate values of β_k for each k. In order to obtain estimates of the coefficient vectors, we maximise the partial likelihood function for each k as in the Cox proportional hazards model with a single cause of failure (Cox, 1972, 1975). Because time-dependence of the hazard function appears only in the term $h_{k,0}(t)$ which does not depend on the covariates, the proportional hazards assumption implies that the ratio of hazards for two units with different values of the covariates is constant for all time.

The two models lead to different estimates of the coefficients $\hat{\beta}_{jk}$ where j denotes

a particular covariate and $k = 1, 2, 3$ denotes the competing events. In both models, the hazard ratios are obtained by taking $\exp(\hat{\beta}_{jk})$. A value of $\hat{\beta}_{jk}$ greater than zero, and equivalently a hazard ratio greater than one, indicates that as the value of the j-th covariate for the k-th event increases, the associated event hazard increases, and thus the length of survival tends to decrease. In other words, a hazard ratio above one indicates a covariate that is positively associated with the event's probability of occurrence, and thus negatively associated with the length of survival. A hazard ratio below one indicates a covariate that has a protective effect in that higher values of this covariate are associated with lower risk and therefore longer survival.

2.1.1 *Variable selection techniques*

Several studies have been devoted to looking for the optimal set of predictors of failure in order to construct a successful model in terms of accuracy, interpretability and predictive ability. Various procedures for identifying the best set of predictors have been proposed in the literature. The simplest method is stepwise selection, which is among the most widely used despite the fact that it suffers from certain drawbacks (Tian and Yu, 2017). One popular way of overcoming these weaknesses is to apply the lasso or another penalised shrinkage approach.

In this paper, we opt for applying the lasso technique, introduced by Tibshirani in linear regression (Tibshirani, 1996), and subsequently extended by him to the Cox model (Tibshirani, 1997) and then to the competing risks model by Fu, Parikh and Zhou (2017). In contrast to subset regression that either sets a coefficient to zero or inflates it, this shrinkage method seeks to set some coefficients to zero while shrinking others and thus producing more stable results. In our data analysis we compare the performance of the lasso with stepwise regression, for both cause-specific and subdistribution hazards. Some details of these variable selection procedures follow.

The stepwise variable selection procedure is an automated procedure for obtaining the best candidate final regression model. In forward selection, a model without covariates is estimated and a univariate model for each covariate is fitted. Then, the variable that maximally improves model fit enters the model. We employed minimisation of Akaike's Information Criterion (AIC) in our application. Each remaining variable is then examined for inclusion, and the process is repeated. The procedure stops when AIC cannot be reduced further. Backward elimination proceeds in the opposite direction. First, the full model which includes all the covariates is estimated, and the variables that maximally reduce AIC are dropped one by one, until AIC cannot be reduced further. In the stepwise procedure, both forward selection and backward elimination are performed at each step.

However, all automatic procedures suffer from drawbacks. In particular, biased and unstable coefficient estimates and prediction could be obtained, since small changes in the data may be reflected in changes in the set of variables selected and consequently in the coefficients and predictions (Hesterberg, Choi, Meier and Fraley, 2008). Furthermore, the presence of multicollinearity among variables can produce

an increase in the variance of estimated coefficients, leading to the erroneous identification of some predictors as irrelevant (Harrell, 2015). Finally, they face significant challenges in high dimensional data (Fan and Lv, 2010).

These drawbacks justify the need to consider other approaches such as the lasso and its variants, which are based on the maximisation of different forms of penalised likelihood. These methods are able to produce interpretable models, accurate predictions and approximately unbiased inferences. The lasso estimate of the vector of coefficients for the subdistribution hazard model is defined (Fu *et al.*, 2017) as

$$\hat{\beta}_{lasso} = \underset{\beta \in \mathrm{R^p}}{\operatorname{argmax}} \{l(\beta) - n \sum_{j=1}^{p} p_\lambda(|\beta_j|)\} \tag{5}$$

where R^p is a p-dimensional space of covariates, $l(\beta)$ is the log-partial likelihood for cause $k = 1, \ldots, K$, $p_\lambda(|\beta_j|) = \lambda|\beta_j|$ is the penalty function, and λ is a tuning parameter that controls the complexity of selected models. Due to the form of the penalty function, the coefficients of less important variables are set equal to zero. Thus the lasso acts as a variable selection method as well as an estimation procedure. A larger value of λ tends to choose a simpler model containing fewer selected variables.

3. DATA DESCRIPTION

The data used in our study were drawn from the Organisation for Economic Cooperation and Development's (OECD) database of firm-level microdata based on the commercial ORBIS database of the Bureau van Dijk (Pinto Ribeiro, Menghinello and De Backer, 2010). It contains administrative information on over 40 million companies and entities worldwide. ORBIS is not an exhaustive database of all companies around the world; it is a collection of business records rather than a comprehensive business register. Our sample consists of 75479 firms operating in the manufacturing sector in several European countries during the period 2000-2018. The countries we considered were Belgium, France, Germany, Italy, Portugal, Spain, and the United Kingdom. For each firm, the financial data for the last available year, its legal form, current legal status and geographical location were extracted.

Following the classification of company status available in the database, we identified three categories of inactive firms: dissolved, liquidated (wound up) and bankrupt. The first category covers companies that no longer exist as legal entities, but the reason for this is not specified. Thus, these companies are defunct, no longer active or are no longer included in the register of companies. Firms in the second category (liquidated) are in process of liquidation, have been dissolved after liquidation of their assets, or are in default (i.e. they are not able to pay their debts). The last category (bankrupt) comprises the firms that are in the process of bankruptcy, have been dissolved at the end of the bankruptcy process, or have been declared insolvent (i.e. they are still active, but are under legal protection). Table 1 shows the distribution of firms with respect to these possible states, by country.

TABLE 1. - *Distribution of firms by status and country*

	Active	Dissolution	Liquidation	Bankruptcy	Total
Belgium	1193	0	12	163	1368
France	4823	262	3	1803	6891
Germany	1120	32	319	14	1485
Italy	34046	43	474	3312	37875
Portugal	5575	10	4	60	5649
Spain	18747	206	0	421	19374
United Kingdom	2642	7	133	55	2837
Total	68146	560	945	5828	75479

TABLE 2. - *Financial indices examined as potential covariates in the analyses*

ID	Formula	ID	Formula
ind001	EBITDA* (logarithm)	ind080	Quick Assets / Total Assets
ind011	Cash flow / Shareholders Funds	ind081	Net Income / Total Assets
ind013	Cash flow / Total Liabilities	ind083	Return on Equity
ind020	Total Assets (logarithm)	ind087	Sales / Cash Flow
ind021	(Creditors / Operating Revenue)*360	ind088	Sales / Current Assets
ind024	Current Assets / Total Assets	ind089	Sales / EBIT**
ind025	Current Liabilities / Current Assets	ind090	Sales / Equity Ratio
ind030	Working Capital / Net Worth	ind092	Sales / Total Assets
ind031	Current Assets / Current Liabilities	ind094	Shareholders Funds / Capital
ind033	Debtors / Sales	ind098	Total Assets / Sales
ind044	Equity / Fixed Assets	ind104	Sales / Shareholders Funds
ind051	Inventory / Total Assets	ind105	Working Capital
ind053	Loans / Total Assets	ind108	Working Capital / Sales
ind055	Long Term Debts / Sales	ind116	EBIT / Interest Paid
ind056	Long Term Debts / Net Capital	ind117	Long Term Debts / Equity
ind057	Long Term Debts / Total Assets	ind118	Net Worth / Total Liabilities
ind063	Net Income / Cash flow	ind119	Net Worth / Total Assets
ind065	Net Income / Fixed Assets	ind124	Receivables / Current Assets
ind071	Non-Current Liabilities / Current Liabilities	ind132	Equity / Sales
ind079	Quick Assets / Sales		

*Earnings before interest, taxes, depreciation and amortisation; **Earnings before interest and tax.

On the basis of the financial statements and information collected for all firms in the sample, we built a dataset containing the predictors that potentially could influence the probability of entering one of the failure states. These predictors are financial ratios that were chosen according to three criteria: i) they have a relevant financial meaning in the context of failure; ii) they have been widely used in the failure prediction literature; and iii) the information needed for their calculation was available. Furthermore, a choice was made between ratios that were highly correlated (≥ 0.7) with each other. After this procedure, the final number of financial variables was 39, as shown in Table 2. They reflect the main aspects of the firm's structure, such as profitability, efficiency,

solvency and liquidity. In addition, some non-financial information – the age of the company, its legal form, and the country – was also included in the list of possible predictors. For analysis of the categorical variables, we chose as reference group Italy (for country) and private limited companies (for legal form). Where a type of event was not recorded, the corresponding category was excluded (i.e. no dissolved firms in Belgium, no liquidation in Spain). For the restricted purposes of this illustrative application, we carried out a 'complete case' analysis - that is, cases with missing values for any of the selected variables were omitted. For the same reason, we did not carry out extensive examination of the data to verify the proportional hazards assumption or to investigate possible transformations of the variables.

4. RESULTS

In this Section we present the main results of fitting the competing risks models. We evaluate the sign and magnitude of the covariates' effects on the competing events and, for easier interpretation of the results, we exhibit the values of the hazard ratios and their 95% confidence intervals. The estimates for the cause-specific hazard functions of all three failure states are presented in Table 3 for both variable selection methods and are discussed in Section 4.1. Corresponding results for the subdistributional hazard functions are shown in Table 4 and discussed in Section 4.2. A comparison between the estimates for the two hazard functions follows in Section 5.

4.1 *The cause-specific hazard function*

Data analysis applying the cause-specific hazards approach was carried out using the R library 'survival', employing the libraries 'MASS' and 'glmnet' for the stepwise and lasso variable selection methods, respectively. Looking at the variables selected by the stepwise and lasso techniques (Table 3), we notice that some significant variables that affect the three competing events are in common between the results of the two variable selection methods, with estimated coefficients that are consistent in sign and magnitude between the two techniques.

Bankruptcy models were the most complex, with a larger number of financial ratios affecting the event. The lasso selected more parsimonious models than the stepwise method for Liquidation and Bankruptcy as causes of failure. In the latter case in particular there was an appreciable difference in the number of variables selected (10 against 19). For the liquidation event, 80% of variables were the same between the two methods.

For both methods and for all the causes of exit, the ratios ind001, ind119 and ind033 were selected. For the first two (ind001 and ind119), higher values led to a decrease in the risk of failure for every exit type. The opposite (increased failure risk for all exits at higher values of the variable) held for ind033. In particular, a higher value of ind119 (Net Worth/Total assets) - a financial ratio that provides an indication of the organisation's financial health - had a very strong protective effect against the bankruptcy event.

TABLE 3. - *Results for Subdistributional hazards regression analyses of competing risks. Stepwise and Lasso selection models are shown for comparison*

	Dissolution-Stepwise HR (95% CI)	Dissolution-Lasso HR (95% CI)	Liquidation-Stepwise HR (95% CI)	Liquidation-Lasso HR (95% CI)	Bankruptcy-Stepwise HR (95% CI)	Bankruptcy-Lasso HR (95% CI)
Country*						
Belgium			3.56 (1.97 - 6.41)		1.70 (1.43 - 2.01)	3.22 (2.73 - 3.80)
France	29.94 (22.06 - 40.63)	29.89 (22.30 - 40.05)	0.08 (0.02 - 0.24)	0.07 (0.02 - 0.22)	3.72 (3.49 - 3.97)	5.67 (5.30 - 6.07)
Germany	30.08 (19.24 - 47.01)	40.79 (25.62 - 64.95)	226.69 (183.33 - 280.32)	181.41 (147.82 - 222.63)	0.29 (0.17 - 0.50)	
Portugal			0.07 (0.03 - 0.20)	0.07 (0.03 - 0.20)	0.17 (0.13 - 0.22)	0.21 (0.16 - 0.27)
Spain	10.45 (7.67 - 14.23)	7.75 (5.74 - 10.45)			0.32 (0.28 - 0.36)	0.47 (0.42 - 0.52)
United Kingdom			18.67 (14.61 - 23.84)	17.52 (13.76 - 22.31)	0.45 (0.34 - 0.59)	
Legal Form**						
Partnerships					0.50 (0.39 - 0.63)	
Public limited companies	1.41 (1.18 - 1.68)	1.70 (1.40 - 2.06)			1.39 (1.29 - 1.50)	
ind001	0.95 (0.94 - 0.96)	0.96 (0.95 - 0.98)	0.93 (0.92 - 0.94)	0.92 (0.91 - 0.93)	0.94 (0.94 - 0.95)	0.94 (0.94 - 0.95)
ind020		0.85 (0.81 - 0.91)	0.60 (0.58 - 0.63)	0.61 (0.58 - 0.64)	0.83 (0.81 - 0.84)	0.84 (0.82 - 0.85)
ind024			2.35 (1.69 - 3.27)		0.27 (0.22 - 0.32)	
ind025						1.06 (1.04 - 1.07)
ind031					1.12 (1.11 - 1.14)	
ind033	1.32 (1.15 - 1.51)	1.38 (1.22 - 1.57)	1.13 (1.02 - 1.25)	1.18 (1.07 - 1.29)	1.32 (1.23 - 1.40)	1.06 (1.01 - 1.11)
ind051	0.54 (0.34 - 0.87)		0.30 (0.21 - 0.45)		6.09 (4.96 - 7.48)	1.69 (1.47 - 1.94)
ind053			1.79 (1.16 - 2.77)	1.63 (1.05 - 2.54)	2.62 (2.34 - 2.94)	2.96 (2.69 - 3.26)
ind055					1.28 (1.20 - 1.35)	
ind057					0.56 (0.42 - 0.76)	0.39 (0.31 - 0.50)
ind071					0.77 (0.73 - 0.81)	
ind079					1.48 (1.38 - 1.59)	
ind081		1.00 (1.00 - 1.00)				
ind092			0.77 (0.70 - 0.85)	0.84 (0.77 - 0.92)		0.67 (0.64 - 0.70)
ind098					0.82 (0.78 - 0.86)	
ind108					0.76 (0.72 - 0.80)	
ind118					0.88 (0.83 - 0.93)	
ind119	0.37 (0.24 - 0.56)	0.39 (0.25 - 0.59)	0.10 (0.08 - 0.14)	0.11 (0.08 - 0.16)	0.03 (0.02 - 0.03)	0.02 (0.02 - 0.02)
ind124					1.28 (1.11 - 1.48)	
ind132			1.14 (1.09 - 1.19)	1.13 (1.08 - 1.19)	1.27 (1.22 - 1.32)	

Reference categories: *Italy; **Private Limited Companies

In addition to covariates derived from the balance sheets, the categorical variables *Country* and *Legal Form* were found to be statistically significant using both selection methods. The analysis revealed differences among countries. The lasso model, compared to stepwise, had for Germany a hazard ratio (HR) not statistically significantly different from 1 for the bankruptcy event, a larger HR in the case of dissolution and a smaller HR for liquidation. The risk of dissolution was much larger in all other countries than in Italy. Varying results were registered for the other events. The risk of entering liquidation was higher in Belgium, Germany and United Kingdom compared to Italy, but lower in France and Portugal. Firms in Belgium and France were the most likely to experience bankruptcy.

With regard to the firm's legal form, both variable selection methods showed a significant differentiation between public limited companies and private ones for the dissolution event; the HR had the same sign but its value was a little higher in the lasso method. For the bankruptcy cause, the legal form was significant only in the stepwise model.

4.2 *The subdistribution hazard*

Data analysis under the subdistribution hazards approach was carried out using the R library 'cmprsk', with 'crrstep' for the stepwise variable selection technique and 'fastcmprsk' for the lasso. Table 4 shows the results of applying the stepwise and lasso selection methods to fitting the subdistribution hazards model for the three competing events. The magnitude and sign of estimates were consistent between the two techniques, except for bankruptcy models. In fact, examining the HR values for that event, two ratios (ind031 and ind132) both acted in opposite directions in the two models. Moreover, ind051 (Inventory/Total Assets) and ind053 (Loans/Total Assets) had almost double the effect under lasso compared to stepwise selection. The lasso bankruptcy model was more parsimonious than the stepwise one, with 21% fewer variables. The country variable and two ratios (ind001 and ind020) were repeatedly relevant for each cause of exit using both selection techniques. EBITDA (ind001) played a protective role in every model, with similar magnitude of the HR for all three events. Total assets (ind020), on the other hand, showed differences: as its value increased, the risk of failure fell for every failure mode, but by a greater factor for liquidation than for the others. Furthermore, we observed close agreement of the estimates of the coefficients of the categorical variables country and legal form between the two selection methods.

5. DISCUSSION AND CONCLUSIONS

The analyses of competing risks resulted in quite similar models for the two different hazard functions. For each cause of failure, almost the same sets of variables were included in both models, with the same signs on their coefficients although sometimes with differing intensities.

TABLE 4. - *Results for Subdistributional hazards regression analyses of competing risks. Stepwise and Lasso selection models are shown for comparison*

	Dissolution-Stepwise HR (95% CI)	Dissolution-Lasso HR (95% CI)	Liquidation-Stepwise HR (95% CI)	Liquidation-Lasso HR (95% CI)	Bankruptcy-Stepwise HR (95% CI)	Bankruptcy-Lasso HR (95% CI)
Country*						
Belgium			3.19 (1.72 - 5.90)	3.22 (1.74 - 5.96)	2.30 (1.96 - 2.71)	1.69 (1.44 - 1.98)
France	26.59 (18.85 - 37.50)	26.19 (18.65 - 36.77)	0.05 (0.02 - 0.16)	0.05 (0.02 - 0.16)	5.02 (4.63 - 5.44)	3.14 (2.91 - 3.39)
Germany	30.75 (19.00 - 49.76)	29.08 (17.98 - 47.05)	221.87 (172.47 - 285.42)	216.63 (168.68 - 278.21)	0.32 (0.19 - 0.55)	0.23 (0.14 - 0.39)
Portugal			0.08 (0.03 - 0.21)	0.08 (0.03 - 0.22)	0.17 (0.13 - 0.22)	0.10 (0.07 - 0.14)
Spain	11.16 (7.88 - 15.81)	10.90 (7.79 - 15.26)			0.34 (0.30 - 0.39)	0.22 (0.20 - 0.25)
United Kingdom	2.65 (1.19 - 5.87)	2.51 (1.14 - 5.56)	22.60 (16.84 - 30.33)	23.52 (17.60 - 31.43)	0.59 (0.45 - 0.77)	0.34 (0.26 - 0.45)
Legal Form**						
Partnerships					0.45 (0.35 - 0.59)	0.54 (0.42 - 0.70)
Public limited companies	1.49 (1.22 - 1.81)	1.47 (1.21 - 1.80)			1.42 (1.30 - 1.54)	1.27 (1.17 - 1.38)
ind001	0.96 (0.94 - 0.97)	0.96 (0.94 - 0.97)	0.93 (0.92 - 0.94)	0.92 (0.91 - 0.93)	0.95 (0.94 - 0.95)	0.92 (0.91 - 0.92)
ind020	0.93 (0.87 - 0.98)	0.92 (0.87 - 0.98)	0.59 (0.56 - 0.63)	0.60 (0.56 - 0.63)	0.81 (0.79 - 0.82)	0.78 (0.77 - 0.80)
ind024					0.40 (0.31 - 0.51)	0.65 (0.50 - 0.84)
ind031					1.09 (1.07 - 1.10)	0.93 (0.89 - 0.96)
ind033	1.40 (1.20 - 1.64)	1.45 (1.27 - 1.67)	1.18 (1.08 - 1.29)		1.18 (1.03 - 1.34)	
ind051		0.53 (0.30 - 0.93)	0.43 (0.29 - 0.64)	0.43 (0.29 - 0.64)	5.28 (3.59 - 7.77)	10.93 (8.13 - 14.71)
ind053					2.52 (1.75 - 3.64)	4.10 (2.67 - 6.30)
ind055					1.43 (1.25 - 1.63)	2.06 (1.69 - 2.50)
ind057					0.43 (0.30 - 0.62)	
ind071					0.80 (0.74 - 0.86)	
ind079					1.63 (1.41 - 1.89)	2.66 (2.14 - 3.30)
ind092			0.83 (0.75 - 0.92)	0.81 (0.73 - 0.90)	0.63 (0.55 - 0.71)	0.66 (0.60 - 0.73)
ind098					0.76 (0.69 - 0.84)	0.57 (0.49 - 0.66)
ind108					0.71 (0.63 - 0.79)	0.43 (0.36 - 0.52)
ind119	0.64 (0.41 - 0.99)		0.11 (0.07 - 0.16)	0.10 (0.07 - 0.14)	0.02 (0.01 - 0.03)	
ind124					1.40 (1.05 - 1.86)	1.58 (1.25 - 2.00)
ind132			1.11 (1.08 - 1.15)	1.12 (1.08 - 1.17)	1.24 (1.09 - 1.40)	0.85 (0.74 - 0.98)

Reference categories: *Italy; **Private Limited Companies

TABLE 5. - *Number of covariates selected by stepwise and lasso in cause-specific and subdistribution hazard models*

	Cause-Specific		Subdistribution	
	Stepwise	Lasso	Stepwise	Lasso
Dissolution	6	7	6	6
Liquidation	10	8	8	7
Bankruptcy	19	10	19	15

Country and legal form variables were significant in both hazard functions, while the set of financial ratios included was a little wider for the cause-specific hazards (Table 5). The subdistribution hazard function did not include for any cause of failure the ratios Current Liabilities on Current Assets (ind025), Net Income on Net Assets (ind081) and Net Worth on Total Liabilities (ind118), all of which appeared in cause-specific hazards.

The cause-specific lasso model turned out to be the most parsimonious one. All models for each cause had in common some indexes, increases in which had a protective (ind001, ind020, ind019) or threatening (ind033) effect on each event. Moreover, some predictors, such as ind051, had different intensities and signs among the causes. The simplest model was for the Dissolution cause, then in Liquidation models we observe one more ratio that may decrease the probability of the event (ind093) taking into account the incidence of Sales on Total Assets; moreover, we notice two new ratios (ind053 and ind132), with an accelerating effect on the event. The first was significant only in the cause-specific model and focuses attention on the incidence of Loans on Total Assets, while the second underlines the relation between Equity and Sales. Bankruptcy models were the most complex ones. Total Assets composition and in particular, the rising incidence of Long Term debts and Current Assets, may restrain the bankruptcy event. Acting in the opposite direction, we observe the ratios Quick assets on Total Sales (ind0079) and Inventory on Total Assets (ind051), the increase in which is particularly risky. A wide discussion within the framework of competing risk models (Caroni and Pierri, 2020) has highlighted the main differences. In this context we observe that the models are very similar both in terms of variables selected, and also in predictive ability which we examined by means of the Area Under the ROC Curve (AUC). This was calculated using the R package 'ROCR' (Table 6). In particular Liquidation and Bankruptcy have AUC values that do not vary over time. The absence of closure events before 16 years precludes AUC estimation before this point.

In this application the use of either the lasso or stepwise methods led to very similar models, both for the cause-specific and subdistribution hazard functions. There is a rather extensive range of other variable selection procedures, such as SCAD; see Beretta and Heuchenne (2019) for an application of this in a study that also examined business failures. There are also different versions of the lasso, including the group lasso which may be useful when there are categorical covariates: see Lin, Wang, Liu and Holtkamp (2013) and Zhao, Zhang and Liu (2014) for examples of

its use. An alternative analysis approach, the use of decision trees and random forests, is also in effect a variable selection procedure since not all variables will appear in the trees that are finally constructed. An example of random forests applied to a competing risks problem can be found in Cafri, Li, Paxton and Fan (2018).

TABLE 6. - *Values of AUC, the area under the ROC curve, for each event at different time horizons. Dissolution events did not occur before 16 years*

	Forward		Lasso	
	CSH	CIF	CSH	CIF
1 year				
Dissolution	-	-	-	-
Liquidation	0.9073	0.9042	0.9026	0.9038
Bankruptcy	0.8854	0.8893	0.8834	0.8492
3 years				
Dissolution	-	-	-	-
Liquidation	0.9072	0.9042	0.9026	0.9038
Bankruptcy	0.8854	0.8893	0.8834	0.8492
5 years				
Dissolution	-	-	-	-
Liquidation	0.9072	0.9042	0.9026	0.9038
Bankruptcy	0.8854	0.8893	0.8834	0.8492
10 years				
Dissolution	-	-	-	-
Liquidation	0.9068	0.9042	0.9022	0.9038
Bankruptcy	0.8854	0.8893	0.8834	0.8492
16 years				
Dissolution	0.8716	0.8786	0.8776	0.8713
Liquidation	0.8854	0.8893	0.8834	0.8492
Bankruptcy	0.8855	0.8893	0.8834	0.8492
18 years				
Dissolution	0.8706	0.8786	0.8768	0.8713
Liquidation	0.9058	0.9042	0.9012	0.9038
Bankruptcy	0.8855	0.8893	0.8834	0.8492

Our objective in this study was to investigate variable selection methods within a context of competing risks. We did not aim to build a predictive model suitable for real-world application. Among other things, that would have required careful construction of the dataset on which the model would be calibrated (Severin and Veganzones, 2018). We used a publicly available database which does not claim to be comprehensive and consequently can be expected to have biases, such as under-representation of smaller enterprises (Pinto Ribeiro *et al.*, 2010). Within this data, we carried out only a complete case analysis, deleting firms that had missing data on any of the variables that we were considering. A full analysis intending to deliver a working predictive model would not be limited to complete cases but would seek to use all the available information. Furthermore, more detailed verification of the assumptions

behind the proportional hazards model should be carried out. For further discussion of the construction of predictive models based on hazards, see Austin *et al.* (2016).

There are two main 'consumers' of survival analysis methods. One is the field of sciences associated with engineering and technology, where (under the name of 'reliability modelling') it is more usual to apply parametric modelling instead of the semi-parametric approach that we have adopted here. The other is the field of biomedical sciences, where the Cox model predominates. Most of the important methodological developments in survival analysis have been made by researchers associated with biostatistics and have subsequently been adopted as required by practitioners in other fields, such as finance. It is relevant to our study to consider whether there are important differences between financial and medical applications. In fact, we do not see any difference in principle between the financial and medical fields in terms of theory or the application of the methodologies that we have employed in the present study. However, there may be some practical differences. The most significant, in our opinion, lies in the size of the data sets, in terms of the numbers of variables in typical applications of survival analysis. In applications such as the present one, the variables are obtained from routinely collected administrative data and consequently there will be a large number to select from. In medical applications, there tend to be fewer variables, in our experience, because many of them are special laboratory tests that are not carried out routinely because of the cost. This means that the choice of variable selection method is less often important in medical applications than financial ones, because the researcher may be able to compare alternate sets of variables in detail 'by hand' without resorting to an automatic method. Another possible difference is that the class imbalance problem, which is undoubtedly of major importance in classification (Severin and Veganzones, 2018; Veganzones and Severin, 2021), is almost always present in financial applications, often to a severe degree, whereas it is not an issue in many medical applications. An investigation of possible interaction between variable selection methods and class imbalance could form a topic for further research.

ACKNOWLEDGEMENTS

Francesca Pierri received the support of ''Fondo Ricerca di Base, 2017-2019'' from the University of Perugia for the project ''Multistate models for competing risk analysis: an analysis for the European manufacturing sector''.

REFERENCES

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

Altman E.I, Iwanicz-Drozdowska M., Laitinen E.K., Suvas A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-Score model. *Journal of International Financial Management & Accounting*, **28**, 131-171.

Amendola A., Restaino M., Sensini L. (2011). Variable selection in default risk models. *The Journal of Risk Model Validation*, **5**, 3-19.

Amendola A., Restaino M., Sensini L. (2014). An empirical comparison of variable selection methods in competing risks model. *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, Springer 13-25.

Amendola A., Restaino M., Sensini L. (2015). An analysis of the determinants of financial distress in Italy: A competing risks approach. *International Review of Economics & Finance*, **37**, 33-41.

Austin P.C., Lee D.S., D'Agostino R.B., Fine J.P. (2016). Developing points-based risk-scoring systems in the presence of competing risks. *Statistics in Medicine*, **35**, 4056-4072.

Beretta A., Heuchenne C. (2019). Variable selection in proportional hazards cure model with time-varying covariates, application to US bank failures. *Journal of Applied Statistics*, **46**, 1529-1549.

Brabazon A., Keenan P.B. (2004). A hybrid genetic model for the prediction of corporate failure. *Computational Management Science*, **1**, 293-310.

Cafri G., Li L., Paxton E.W., Fan J. (2018). Predicting risk for adverse health events using random forest. *Journal of Applied Statistics*, **45**, 2279-2294.

Caroni C., Pierri F. (2020). Different causes of closure of small business enterprises: alternative models for competing risks survival analysis. *Electronic Journal of Applied Statistical Analysis*, **13**, 211-228.

Chancharat N., Tian G., Davy P., McCrae M., Lodh S. (2010). Multiple states of financially distressed companies: Tests using a competing-risks model. *Australasian Accounting, Business and Finance Journal*, **4**, 27-44.

Cox D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187-202.

Cox D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Crowder M.J. (2001). *Classical competing risks*. Chapman and Hall/CRC, New York.

Headd B. (2003). Redefining business success: Distinguishing between closure and failure. *Small Business Economics*, **21**, 51-61.

Dakovic R., Czado C., Berg D. (2010). Bankruptcy prediction in Norway: a comparison study. *Applied Economics Letters*, **17**, 1739-1746.

Dyrberg A. (2004). Firms in financial distress: An exploratory analysis. *Tech. Rep., Danmarks Nationalbank Working Papers*.

Esteve-Pérez S., Sanchis-Llopis A., Sanchis-Llopis J.A. (2010). A competing risks analysis of firms' exit. *Empirical Economics*, **38**, 281-304.

Fan J., Lv J. (2010). A selective overview of variable selection in high dimensional feature space, *Statistica Sinica*, **20**, 101.

Fine J.P., Gray R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**, 496-509.

Fu Z., Parikh C.R., Zhou B. (2017). Penalized variable selection in competing risks regression. *Lifetime Data Analysis*, **23**, 353-376.

Harhoff D., Stahl K., Woywode M. (1998). Legal form, growth and exit of West German Firms - Empirical results for manufacturing, construction, trade and service industries. *The Journal of Industrial Economics*, **46**, 453-488.

Harrell E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer Cham.

Hensher D.A., Jones S. (2007). Forecasting corporate bankruptcy: Optimizing the performance of the mixed logit model. *Abacus*, **43**, 241-264.

Hesterberg T., Choi N.H., Meier L., Fraley C. (2008). Least angle and $\ell 1$ penalized regression: a review. *Statistics Surveys*, **2**, 61-93.

Jones S., Hensher D.A. (2004). Predicting firm financial distress: A mixed logit model. *The Accounting Review*, **79**, 1011-1038.

Jones S., Hensher D.A. (2007). Modelling corporate failure: A multinomial nested logit analysis for unordered outcomes. *The British Accounting Review*, **39**, 89-107.

Lee M.-C. (2014). Business bankruptcy prediction based on survival analysis approach. *International Journal of Computer Science & Information Technology*, **6**, 103-119.

Lennox C. (1999). Identifying failing companies: a re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business* , **51**, 347-364.

Lin H., Wang C., Liu P., Holtkamp D.J. (2013). Construction of disease risk scoring systems using logistic group lasso: application to porcine reproductive and respiratory syndrome survey data. *Journal of Applied Statistics*, **40**, 736-746.

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

Pierri F., Caroni C. (2017). Bankruptcy prediction by survival models based on current and lagged values of time-varying financial data. *Communications in Statistics: Case Studies, Data Analysis and Applications*, **3**, 62-70.

Pinto Ribeiro S., Menghinello S., De Backer K. (2010). The OECD ORBIS Database: Responding to the need for firm-level micro-data in the OECD. *OECD Statistics Working Papers*, 2010/01.

Putter H., Fiocco M., Geskus R.B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389-2430.

Rommer A.D. (2004). Firms in financial distress: An exploratory analysis. *Danmarks Nationalbank Working Papers*.

Rommer A.D. (2005). A comparative analysis of the determinants of financial distress in French, Italian and Spanish firms. *Danmarks Nationalbank Working Papers*.

Schary M. A. (1991). The probability of exit. *The RAND Journal of Economics*, 339-353.

Severin E., Veganzones D. (2018). Sixty years of bankruptcy models: issues, limits, and progress. *Bankers, Markets & Investors*, **154**, 01.

Shumway T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, **74**, 101-124.

Tian S. and Yu Y. (2017). Financial ratios and bankruptcy predictions: An international evidence. *International Review of Economics & Finance*, **51**, 510-526.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267-288.

Tibshirani R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Veganzones D., Severin E. (2021). Corporate failure prediction models in the twenty-first century: a review. *European Business Review*, **33**, 204-226.

Zhao W., Zhang R., Liu J. (2014). Sparse group variable selection based on quantile hierarchical lasso. *Journal of Applied Statistics*, **41**, 1658-1677.

Zmijewski M.E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, **22**, 59-82.