



# Time varying effects in survival analysis: a novel data-driven method for drift identification and variable selection

Zakaria Babutsidze<sup>1,2</sup> · Marco Guerzoni<sup>3,4</sup>  · Luigi Riso<sup>5</sup>

Received: 21 December 2022 / Revised: 26 October 2023 / Accepted: 8 January 2024 /

Published online: 26 February 2024

© The Author(s) 2024, corrected publication 2024

## Abstract

In this paper we address the problem of survival models when high-dimensional panel data are available. We discuss two related issues: The first one concerns the issue of variable selection and the second one deals with the stability over time of such a selection, since presence of time dimension in survival data requires explicit treatment of evolving socio-economic context. We show how graphical models can serve two purposes. First they serve as the input for a first algorithm to assess the temporal stability of the data: Secondly, allow the deployment of a second algorithm which partially automates the process of variable selection, while retaining the option to incorporate domain expertise in the process of empirical model-building. To put our proposed methodology to the test, we utilize a dataset comprising Italian firms funded in 2009 and we study the survival of these entities over the period of 10 years. In addition to revealing significant volatility in the set of variables explaining firm exit over the years, our novel methodology enables us to offer a more nuanced perspective than the conventional one regarding the critical roles played by traditional variables such as industrial sector, geographical location, and innovativeness in firm survival.

**Keywords** Firm survival · Model drift · Variable selection

---

✉ Marco Guerzoni  
marco.guerzoni@unimib.it

<sup>1</sup> SKEMA Business School, Université Côte d’Azur (GREDEG), Valbonne, France

<sup>2</sup> OFCE, Sciences Po, Paris, France

<sup>3</sup> DEMS-University of Milan-Bicocca, Piazza dell’Ateneo Nuovo 1, 20126 Milan, Italy

<sup>4</sup> BETA, University of Strasbourg, Strasbourg, France

<sup>5</sup> Università Cattolica del Sacro Cuore, Milan, Italy

## 1 Introduction

Understanding the reasons behind firms' exit from markets can help design appropriate industrial policy and managerial strategies. The theoretical problem of exit finds its empirical counterpart in survival models. Classical approaches to modeling firm exit have three main characteristics: the dependent variable is the time span before the realization of an event (exit/death); Data are right censored, since some observations never experience the event; Covariates explain the average waiting time for the exit to occur and they can be considered either risk factors to be analyzed or control variables, depending on the problem at hand. On this basis, survival analysis tests which mechanisms explain a firm's exit.

This paper discusses a crucial assumption in survival models, that is that the firm exit mechanisms are constant over the considered time-span. When the modeling period is relatively short, the argument for a stable mechanism is fair. However, we surmise that when observations span over many years that are characterized by different economic conditions and evolving institutional landscape, the stability of exit mechanisms should be carefully tested rather than straightforwardly assumed.

The potential damage from assuming stable exit mechanisms is exacerbated by the current digital revolution that provides researchers with longer data-sets for survival exercises. Large data-sets can greatly improve both prediction capability and the causal analysis of specific risk factors of interest. However, the vast availability of data might create the 'embarrassment of riches' (Altman & Krzywinski, 2018; Coad et al., 2013; Eklund et al., 2007) in the choice of the variables (and thus exit mechanisms) since the requirements of model conciseness, exogeneity of the covariates and absence of collinearity force the researcher to make an educated selection among many variables. Ample data, in combination with the potential of changing causes for firm's death, present an important challenge for scientists studying firm survival. We like to clearly state that the problem at hand is much deeper than a model with time-varying coefficients (Cefis & Marsili, 2019), but that in different time we might require different models.

In this paper we present a data-driven alternative to conventional data selection. We rely on the advances in graphical modeling, in particular the *High-Dimensional Graphical Model* (Edwards et al., 2010), in order to develop a strategy for variable selection that can be used for econometric modeling of firm survival. Most notably, the proposed approach does not impose a unique mechanism for firm survival over the period of analysis. Instead, it is flexible in allowing for variations in the selection mechanism over time. Additionally, we maintain that any reasonable data-driven methodology should also allow for the researcher-in-the-loop feature. It is important that the final word in the variables selection process stays with the researcher who can enrich the estimated survival model by the theory-driven knowledge accumulated in the literature. We do this by splitting the proposed algorithm in two phases. At the first stage, we use a

data-driven graphical model, for variable selection, and for the computation of the statistical drift over the time (Riso & Guerzoni, 2022). The assessment of drift does not provide a resolution to the issue of time stability; rather, it should be regarded almost as a diagnostic test. There are established procedures for detecting structural breaks in time series, such as the sudden shift in a parameter value, as outlined by Chow (1960), or, more generally, change detection methodologies (Picard, 1985). More recently, the literature in machine learning has coalesced under the encompassing term “drift detection.” This incorporates the estimation of time-sensitive parameters, the potential alteration in the functional form characterizing the relationship among variables, and, ultimately, the variables to be included in the data set. Riso and Guerzoni (2022) reviewed the pertinent literature and introduced the current general framework applied herein. A distinctive feature of this framework is that the analysis is not contingent upon the model being utilized but calculates the magnitude of a drift. Consequently, in the absence of a relevant drift, the same variable selection and model can be applied consistently over time. On the contrary, in presence of a drift, the variable selection is re-evaluated at every period (length of which needs to be specified by the researcher).

In the second stage, we tackle variable selection by using a secondary algorithm, which also receives its primary input from the graphical model. The algorithm (Riso et al., 2023) automatically selects the set of variables which maximize the relevance of the the information in the data, while minimizes its redundancy, as explained later in the paper. In this version of the algorithm we allow the research to modify the initial graphical model. This modification empowers researchers to guide the algorithm’s consideration of variables, ensuring it takes into account those of theoretical significance and excludes others that may be dismissed due to known measurement errors or a lack of meaningful relevance. Since the graphical is a derived automatically, any deviation from the original graph demands a clear and comprehensive motivation, redenring tra variable selection procedure very transparent. Eventually, we can proceed with the econometric analysis, with the certainty that we accounted for any drift and the variable selection remains consistent with the underlying data structure.

The key characteristic of the proposed framework is the use of data science algorithms to empower and complement, rather than completely substitute out the researcher since we allow for the possibility of the researcher to act in a codified and transparent way at a certain stage of the algorithm. With this aim in mind, Sect. 2 overviews the literature related to conventional and data-driven firm survival analysis. Section 3 shortly overviews graphical models, which constitute the backbone of the proposed methodology, and presents the proposed algorithm. Section 4 presents an empirical exercise where we apply both conventional and data-driven survival analysis methods to Italian startup data. We present two main results. Firstly, we show a significant data drift overtime, which needs to be taken account by survival models. Secondly, we expose the differences between conventional firm survival approach and the methodology proposed in this paper. Section 5 concludes.

## 2 Conventional and data-driven survival analysis

Firm survival and its determinants have long been acknowledged as key issues in business studies, as survival is a necessary condition for success (Barnard, 1938; Coeurderoy et al., 2012; Delmar et al., 2022; Suárez & Utterback, 1995). However, a surge in the analysis took place over past three decades as the focus of innovation studies shifted toward entry, exit and growth as key elements of industrial dynamics (Geroski, 1992; Klepper, 1996).

Traditional approach to survival analysis in such a context relies on an extensive literature that highlights (both empirically and theoretically) the main determinants of survival (Giot and Schwienbacher 2007; Pérez et al. 2004, among others). One can distinguish six fundamental elements in traditional survival analysis which have emerged over the past decades:

- Age and size (Audretsch, 1995; Geroski, 1995; Pérez et al., 2004; Savin et al., 2023);
- Sector or Industry of belonging; (Geroski, 1995; Klepper, 1996; Malerba & Orsenigo, 1997)
- Geographical localization; (Acs et al., 2007; Sternberg & Litzengerger, 2004; Sternberg et al., 2009)
- Profitability; (Delmar et al., 2013; Mogos et al., 2021)
- Liquidity constraints; (Holtz-Eakin et al., 1994; Musso & Schiavo, 2008)
- Innovativeness and Entrepreneurship. (Cefis & Marsili, 2005; Guerzoni et al., 2020; Santarelli & Vivarelli, 2007)

While the empirical literature tends to concur on the influence of certain factors, there remains a lack of consensus regarding others. It is widely acknowledged that older and larger firms generally exhibit a higher likelihood of survival. This phenomenon is often tied to their enhanced profitability and reduced liquidity constraints. Additionally, geographical location serves as a crucial control variable due to the presence of agglomeration economies, which increase survival rate.

Furthermore, the life cycle of an industry and its knowledge base undoubtedly represent pertinent control factors. In some cases, they might even account for the majority of the variance previously attributed to location. Lastly, innovation and entrepreneurship, particularly in the context of young startups, introduce a dynamic dimension to the survival equation. On one hand, they can be a source of vitality and adaptability, enabling firms to seize new opportunities and pivot when needed. However, this dynamism also comes with inherent risks, such as increased uncertainty and resource constraints, which may impact a startup's survival prospects.

More recently, the increased availability of register data for longer time-spans has enabled the production of a vast array of survival exercises. This literature can be divided in two streams of research.

The first research stream collects theory driven, econometric works aiming at testing the direction, significance, and magnitude of a specific causal impact of

a variable on the probability to survive. A sound econometric exercise is able to elicit a causal effect, but the choice of models is limited by the capacity to derive estimators with adequate properties for the inference process: Endogeneity, multicollinearity, reverse causality, degrees of freedom, and heteroskedasticity are the main points of concern which accompany the usually theory driven process of the variable choice. The second research stream unites survival studies that exploit new tools in data science and focus on the prediction of survival using any possible variable at disposal. However, despite the high flexibility in the choice of model and in variable selection, these studies are silent on the impact of a specific variable on the survival probability.

A comprehensive literature review of the hundreds of articles employing econometric survival analysis is beyond the scope of this work, but can be found in recent reviews (Cefis et al., 2021; Hyytinen et al., 2015). Here we highlight a few recent works that are exemplary for the choice of variables and the methodology. Zhang et al. (2018) employ a dataset of Chinese firms and focus on six variables including size, proxy for innovative performance, productivity, capital intensity, and industry dummies. Ortiz-Villajos and Sotoca (2018) use variables on innovative performance, size, corporate social responsibility, entrepreneurs psychological traits, but do not include productivity, location, nor capital intensity. Jung et al. (2018) analyze a sample of South-Korean firms, and study the impact of innovation and R &D on firm survival. Using the dataset AIDA-BvD (Analisi informatizzata delle aziende italiane), by Bureau van Dijk Electronic Publishing Ltd, on Italian firms, Grazzi et al. (2021) explain different types of exit with the innovative performance, firms size, productivity, financial stability, age, and both industry and geographical controls. Using the same source of data, Agostino et al. (2021) discuss the impact of R &D and innovative activities on the risk of bankruptcy, Basile et al. (2017) examine the effect of agglomeration economies on firms survival, and Guerzoni et al. (2020) explore the survival of innovative start-ups. Useche and Pomet (2021) analyze the exit routes of high-tech firms considering information on the venture capitalist. Using data on Dutch firms, Zhou and van der Zwan (2019) examine the impact on survival of firms' growth controlling for age, size, sector and firm urban location, while Cefis and Marsili (2019) explore the impact of economic downturn on firms exits controlling for their innovative activities.

All of these works and, to the best of our knowledge, any other firm survival exercise exploit hazard models to test the impact of the variable on the probability to survive. The choice of the independent variables and controls depends on the data collected and is usually theory-driven.

The second research stream builds on the idea of employing data-driven algorithms to predict firm bankruptcy. This dates back to Altman (1968)'s use of discriminant analysis. The first generation of works falling under this stream are reviewed by Bellovary et al. (2007). Later years have seen an explosion of research on survival prediction, including in the field of industrial dynamics. Bargagli-Stoffi et al. (2021) review 26 studies and summarize the accuracy of exit prediction results. The number of variables employed in these analyses could go as high as 190 [as in Liang et al. (2016)]. Other recent studies employ variables from unconventional sources such as company website features and content (e.g.,

Crosato et al., 2021). In these prediction exercises, the variable choice is not an issue of primary importance since machine learning models do not rely on inference for assessing model uncertainty and, therefore, do not impose the usual constraints employed in econometric models.

In this paper, we maintain as the main objective to derive an econometric model and not a prediction exercise. However, in the presence of a rich set of variables and a relative long time-span, we exploit an unsupervised machine learning tool to improve the variable selection process. Importantly, this process is dynamic and takes into consideration potential changes in market forces and environment. Once data are collected and the variable selected, in this realm authors consider the average effect over time of covariates upon the dependent variable, assuming that the underlying process is rather stable. There are few studies which account for a potential dynamic over time. For instance Cefis and Marsili (2005) discusses options when the proportional hazard assumption is rejected and opt for an accelerated time model in which the effect of a covariate does not represent an impact on the hazard, but rather accelerate or decelerate the process leading to the event, object of study. Instead Cefis and Marsili (2019)'s solution is the adoption of a piece-wise exponential model, which assume a constant baseline risk in different periods. However, also this rare attempt of estimating time-varying coefficients do not evaluate the stability of the overall process as represented by the available data, nor operate a variable selection procedure, which, accordingly, can be flexible over time. In order to allow for this, we consider the potential statistical drift in data and allow for time-varying set of explanatory variables. We focus on the issue of variables selection, which, in the presence of high number of variables, creates the trade-off between the need of minimizing the information loss and satisfying standard constraints of econometric models (i.e., exogeneity, coherence with the theory, etc.).

Traditionally, variable choice in an econometric exercise is theory driven and operated by the researcher based on both its educated guess and, in some cases, a process of trials-and-errors. While this process is theoretically sound, in practice it could presents significant drawbacks. If the variable set is extremely large, this task can go beyond the cognitive ability of the researcher, who will opt for cognitive shortcuts (Gigerenzer & Selten, 2002). Thus, this process could be influenced by cognitive biases and be subject to scientific malpractices, such as the p-hacking (Carota et al., 2015; Head et al., 2015). Moreover, the process of selection can be opaque and ex-post justified leading to reverse p-hacking problem and selective reporting (Chuard et al., 2019). On the contrary, an automated purely data-driven process is fast, transparent and free from biases, but it does not allow to leverage the information coming from theory, expertise, and scientific literature.

The proposed method hinges on combination of the two approaches in order to capitalize on their respective strengths. We surmise that advances in data science could be productively used to cut through wide data sets by taking the first step of exposing the structure of data, which would then guide the researcher to identify smaller set of potentially important variables to be considered for inclusion in the final econometric model. Here we propose the use of Graphical Models which are particularly powerful in uncovering hidden structures in high-dimensional data.

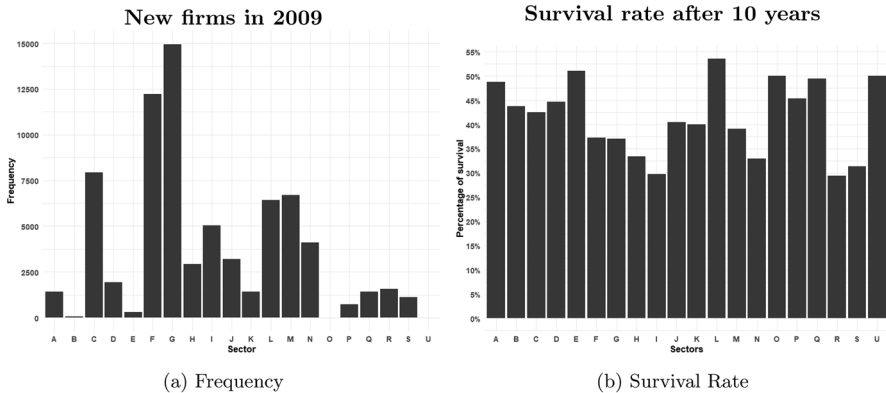
### 3 Graphical models

In this section, we present Graphical Models (GM) as a data-driven approach to structural learning. Structural learning aims at inferring structural relations among a high number of variables in the context of big data (Koller et al., 2007). Graphical models are flexible enough to allow for performing the drift analyses, as well as for the implementation of the variable selection algorithm. Carota et al. (2015) present a simple introduction to graphical models with an application in innovation studies.

#### 3.1 Basic elements of graphical models

GM are a method to display the conditional (in)dependence relationships between variables through a network representation. A network is a graph, that is a mathematical object  $G(V, E)$ , where  $V$  is a finite set of nodes with direct correspondence with the variables present in the dataset, and  $E \subset V \times V$ , is a subset of ordered couples of  $V$  representing the edges of the network and the dependence relationship between variables (Lauritzen, 1996). GM employed in this paper belong to *classes of multivariate distributions* (de Abreu et al., 2009), whose conditional independence properties are encoded by a graph in the following way: the variables have a direct representation as the nodes of the graph and the absence of the edges between nodes represents conditional independence between the corresponding variables. In this paper we make use of undirected graphical models,  $G = (V, E)$ , where  $V = \{v_1, \dots, v_p\}$  is the set of vertices and  $E$  is the set of edges. An edge  $e = (u, v) \in E$  indicates that the variables associated to  $u$  and  $v$  are conditionally dependent (Jordan, 2004).

The empirical problem of model selection consists in learning the structure of the probability function or, in learning the relations among variables in a complex system encoded in the graphical structure itself (Carota et al., 2015). In order to make the estimation of the graph feasible, we restrict the analysis to undirected strongly decomposable graphs (Lauritzen, 1996). In such graphs, two non-adjacent nodes are connected (if at all) by a unique path. Such graphs are referred to as *trees*. As a consequence, these graphs do not include cycling paths between pairs of variables that would significantly complicate the problem. If a dataset can be represented as a collection of trees, it is referred to as a *forest*. The statistical problem is to estimate the maximum spanning tree, that is the tree which maximizes the mutual information among variables. Although the Maximum Likelihood Estimator for this problem exists in explicit form, its calculation is extremely demanding. Instead we take a computational shortcut and carry out the estimation relying on the *Chow–Liu Algorithm* (Chow & Liu, 1968). In particular, we adopt the extension of the *Chow–Liu Algorithm* proposed by Edwards et al. (2010), which allows the use of discrete and continuous random variables in the same graphical model. In this context, the connections between nodes in a tree, or tree dependencies, represent their (unknown) joint probability distributions, thereby providing information regarding their mutual dependence. In summary, the connection between two generic nodes (variables) is determined by calculating their mutual information, which serves as a measure of



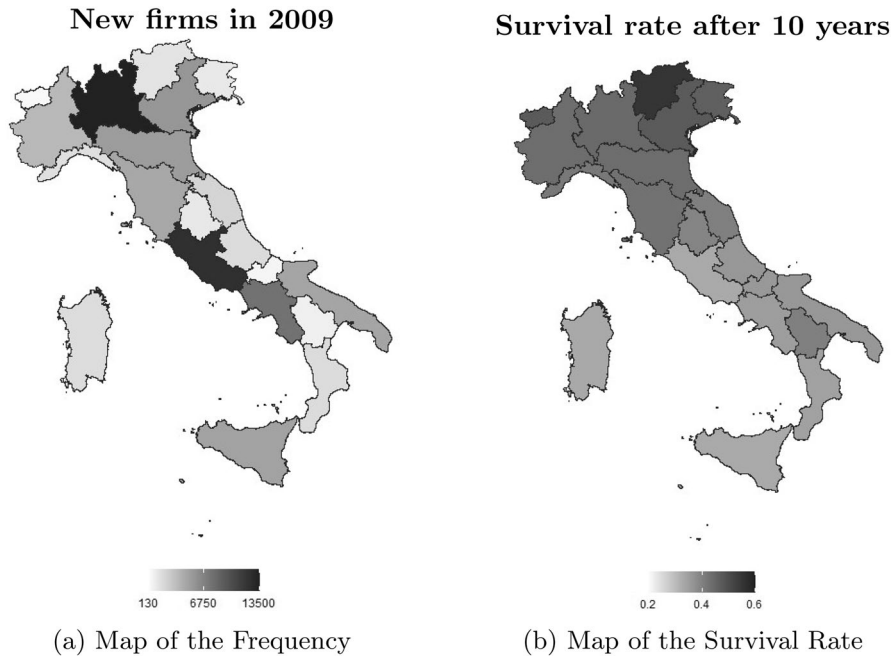
**Fig. 1** Histograms of the sectors in the *start-ups*

their proximity (Lewis, 1959). The algorithm proposed by Edwards et al. (2010) is designed to identify the maximum weight-spanning tree within an arbitrary undirected connected graph with positive edge weights and has been extensively investigated. In this context, Kruskal's algorithm (Kruskal, 1956) offers a straightforward and efficient solution to this problem. It starts with an empty graph and proceeds by adding, at each step, the edge with the greatest weight that does not create a cycle with the previously selected edges. All in all, the graph represents the conditional independence structure of a dataset and, as suggested by (Jordan et al. 2004), the use of graphs to represent both causal relations and sets of conditional independence relations is relevant to econometric non only because the graphical approach to causal inference has led to a more explicit formulation of the assumptions implicit in some social science methodology. Moreover, in this paper GM serves as the main input of both the algorithm for detecting drift and the algorithm for variable selection.

### 3.2 Graphical models for survival analysis: detecting drift and variable selection

As GM compute conditional dependencies in a given data-set, they allow the researcher to scrutinize direct associations between any variable and the designated variable of interest. In survival analysis such a variable interest could be an indicator of firm exit. In modeling such variable (or its complement), conventional survival models assume that the data generation process is stable over time. However, this assumption is likely to be violated when we are considering extended periods of time. It is more reasonable to allow for the possibility that there exist a non-observable hidden context responsible for the data generation that could change overtime (Gama et al., 2014). Presence of such dynamics could introduce the statistical drift in the data. Under such circumstances, GM could be used to estimate both the presence and the magnitude of the drift by comparing the data structure encoded





**Fig. 2** Descriptive look at Italian startups

in different time periods. In this paper we borrow from Riso and Guerzoni (2022) who develop a Bayesian model which estimates the probability that connections (or their absence) in an estimated graphical model are stable over time. In the presence of significant drift, the appropriate approach is to estimate a sequence of graphical models as a single graphical model spanning the whole study period cannot deliver correct estimates (Riso & Guerzoni, 2022).

Beyond estimating the drift, in this paper we use GM to select model variables in the context of high dimensionality. We employ an algorithm using the Minimum Redundancy Maximum Relevance (mRMRe) approach (De Jay et al., 2013) that ranks variables according to their relevance of the information for the target variable (Kratzer & Furrer, 2018) (survival, in our case) by considering the conditional dependence as defined by GM. On this basis, the algorithm can autonomously select variables to include in the regression analysis, requiring no additional effort from the researcher. However, variable selection is not driven solely by informational content but is also influenced by theoretical underpinnings. Therefore, we introduce the concept of 'the-man-in-the-loop' into the otherwise automated process. This step allows for the integration of the researcher's knowledge and scientific judgment into the procedure. Exclusions can be grounded in economic theory or based on statistical attributes of the considered variable that might introduce undesirable assumption violations into the econometric analysis. Consequently, 'the-man-in-the-loop' adjusts the GM by adding and removing links, after which the automatic selection mechanism operates. It is crucial to emphasize that this process is fully transparent, as any modification to the original GM must be explicitly justified.

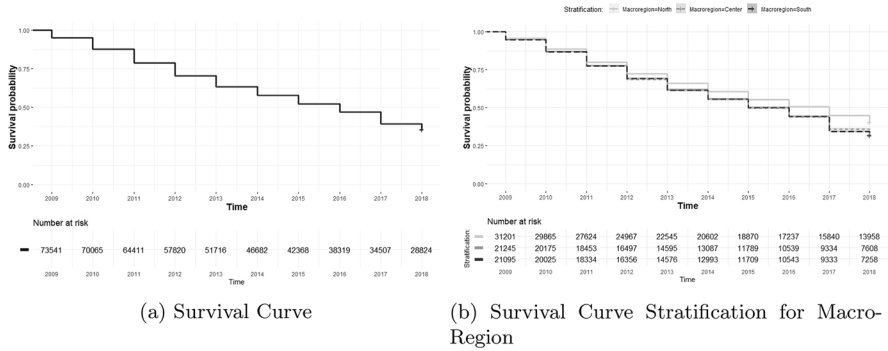


Fig. 3 Survival curves for the firms born in 2009 and number of firms at risk

Table 1 Log-rank test for macro-region

Macro-region	<i>N</i>	Observed	Expected	$\frac{(O-E)^2}{E}$
North	31,201	18,634	20,876	241
Center	21,245	14,366	13315	83
South	21,095	14,409	13,218	107

$$\chi^2 = 480, \text{ on } 2 \text{ degrees of freedom, } p\text{-value} \leq 2e - 16$$

Further technical details and the comparison with other feature selection methods can be found in Riso et al. (2023). In short, the process is described by the following steps:

- (unsupervised) production of a graphical model for each year;
- evaluation of the drift;
- researcher’s analysis of the dependency structure of the GM and theory-driven transparent intervention removing or adding variables and subsequent transformation of the GM;
- automatic variables selection on the transformed GM;
- econometric analysis.

This procedure allows for a transparent selection of the variables, by blending automatic algorithms and theory.

## 4 Empirical application

### 4.1 Data

The analysis is based on AIDA-BvD data, which contains comprehensive information on all Italian firms required to file accounts. Each firm is described by a large

**Table 2** Test for the proportional-hazards assumptions

Covariates	$\chi^2$	df	p-value
Region	55.25	19	2.1e-05
Sector	107.76	17	3.2e-15
Innovative startup	9.78	1	0.0018
Production costs	3.89	1	0.0485
Total from sales	18.54	1	1.7e-05
Index liquidity	27.44	1	1.6e-07
Employees	80.22	1	< 2e-16
Global test	299.78	41	< 2e-16

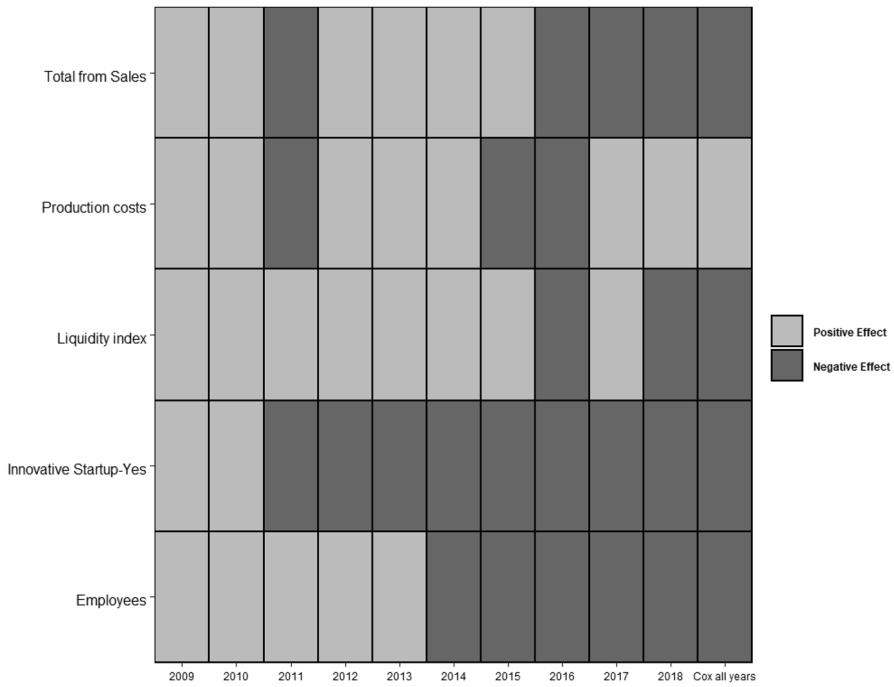
number of variables in the following categories: identification codes and vital statistics; activities and commodities sector; legal and commercial information; share accounting and financial data; shareholders, managers, company participation. From this database we consider variables with the lowest percentage of missing data and that describe all macro categories. Specifically, we observe all firms funded in 2009 and we observe them along a time span of 10 years. Details of the variables are presented in Table 4 in Appendix 1. Since there is still a percentage of missing variables, we use *Random Forest Missing Algorithm* (Tang & Ishwaran, 2017) as the missing data imputation strategy.<sup>1</sup>

## 4.2 Traditional survival analysis

The traditional approach of variable selection (including for survival analysis) consists in deriving from both theory and literature the most promising hypotheses to be tested. Among the variables in the dataset, following the literature reviewed in Sect. 2, we selected *Region*, *Sector*, *Total from sales* and *Production cost* as a proxy for profitability, *Liquidity index* to capture liquidity constraints, *Employees* and *Sales* for the size, and *Innovative Startups* for the innovativeness. The latter represents a variable encoding whether a given firm is registered in the register of Italian innovative start-ups (after 2012). The variable *Sector* is at *Nace Rev.2* level. Figure 1 (panel (a)) shows the distribution of all firms born in 2009 in Italy across sectors. The same figure (panel (b)) reports 10-year survival rates in each of the sectors. As observed in Fig. 1a, the most prevalent sectors are G (wholesale and retail trade, with 14,955 new firms) and F (construction, with 12,214 new firms). They are followed by sectors C (manufacturing), L (real estate activities), and M (professional, scientific, and technical activities), with 7933, 6422, and 6703 new firms, respectively. In contrast,

<sup>1</sup> This method has some desirable properties, since it is able to handle mixed types of missing data. Furthermore, it is adaptive to interactions and non-linearity and it has the potential to scale to big data settings (Tang & Ishwaran, 2017) We implemented the *Random Forest Missing Algorithm* with the support of Open Computing Cluster for Advanced data Manipulation (OCCAM) at the University of Turin (Aldinucci et al., 2017, 2018).

### Comparison of the odds ratio



**Fig. 4** Impact on probability of survival (regions and sectors control are not shown)

the sectors at the lower end of the spectrum include B (mining and quarrying, with only 64 new firms), as well as O (public administration and defence) and U (activities of extraterritorial organizations and bodies), each with just 2 firms. Regarding the survival rates depicted in Fig. 1b, the sectors exhibiting the highest rates are L (real estate activities, at 53%) and E (water supply; sewerage, waste management, and remediation activities, at 51%). Conversely, the sectors with the lowest 10-year survival rates are R (arts, entertainment and recreation), I (accommodation and food service activities), and S (other service activities) with rates of 29.4%, 29.7%, and 31.3%, respectively. The maps in Fig. 2 describe the geographical distribution and corresponding 10-year survival rates in the population of Italian firms created in 2009. It is worth noting, as depicted in Fig. 2a, that Lombardy is the region that had the highest number of new firms in 2009, totaling 13,430. Other regions that also exhibited a significant presence of new firms in 2009 include Lazio (12,217), Campania (7645), and Veneto (5453). Slightly trailing behind is Emilia Romagna with 5079 new firms established in 2009. At the lower end of the spectrum, we find Basilicata with 590 new firms, Molise with 317, and Valle d’Aosta with 132 new

firms, all in the same year. Analyzing Fig. 2b, it becomes evident that Trentino Alto Adige (54%), Valle d'Aosta, and Veneto (47%) are the regions boasting the highest survival rates after a decade. Conversely, Calabria (34%), Sardinia, and Sicily (both 32%) occupy the lower end of the survival rate ranking. More details are reported in Appendices in the Tables 5 and 6.

We can define  $X_i = \{X_{i,1}, \dots, X_{i,p}\}$  as the realized values of the  $p$  covariates for firm  $i$ , and  $Y_i$  as the corresponding survival status. For this exercise, we adopt semi-parametric hazard models, that are specifically designed to examine the duration phenomena to ascertain survival determinants by explaining the time period between a firm's birth and its cessation of economic activity. The most commonly used models for survival data describe the transition rate from one state to another, where in this case the transition is represented by the death of the firm (Kyle et al., 1997). These models belong to a class of Poisson regressions, in particular the *Cox* proportional hazard models:

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp \sum_{j=1}^p \beta_j \mathbf{X}_{i,j}. \quad (1)$$

It is worth noting that some variables are time variant. Following the standard approach to survival analysis, we consider the time dimension according to:

$$\lambda(t|\mathbf{X}(t)_i) = \lambda_0(t) \exp \sum_{j=1}^p \beta_j \mathbf{X}(t)_{i,j}, \quad (2)$$

where the covariate  $X(t)$  is the value of time-varying covariate for the  $i_{th}$  subject at time  $t$ , with  $t = 1, \dots, T$ . The partial likelihood, in general, can be written out as

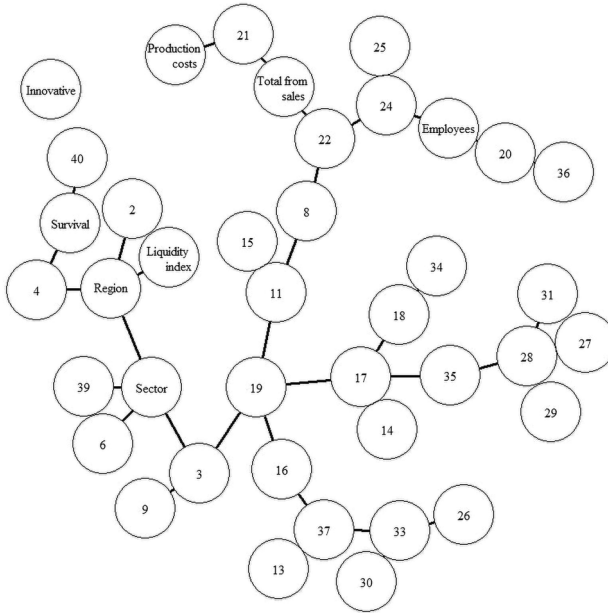
$$L(\beta) = \prod_{t=1}^T \left[ \frac{\lambda(\mathbf{Y}_i|\mathbf{X}_i(t))}{\sum_{i \in R_i(t)} \lambda(\mathbf{Y}_i|\mathbf{X}_i(t))} \right], \quad (3)$$

where the expression  $i \in R_i$  indicates that the sum is taken over all subject in the risk set  $R_i$  at time  $t$ . Figure 3a shows the survival curve for the firms born in 2009, while Fig. 3b shows the survival curves with the stratification for *Macro-Region*.

Figure 3b emphasizes the role played by the macro-regions in Italy, revealing a clear socio-economic divide between the North and the remaining parts of Italy. Table 1 presents the results of the *log-rank* (Somes, 1986) test applied to macro-regions, which is based on the  $G - \rho$  family method introduced by Harrington and Fleming (1982). With 2 degrees of freedom, the  $\chi^2$  test statistic yields a value of 480, and the corresponding  $p$ -value is  $\leq 2e - 16$ . This result allows us to reject the null hypothesis in favor of the alternative hypothesis, supporting our initial observation from Fig. 3b<sup>2</sup>.

<sup>2</sup> The *log-rank* is the most widely used method for comparing different survival curves. It is approximately distributed as a  $\chi^2$  test statistic and is a non-parametric test, which makes no assumptions about the survival distributions.

### Spanning Tree 2009



### Spanning Tree 2010

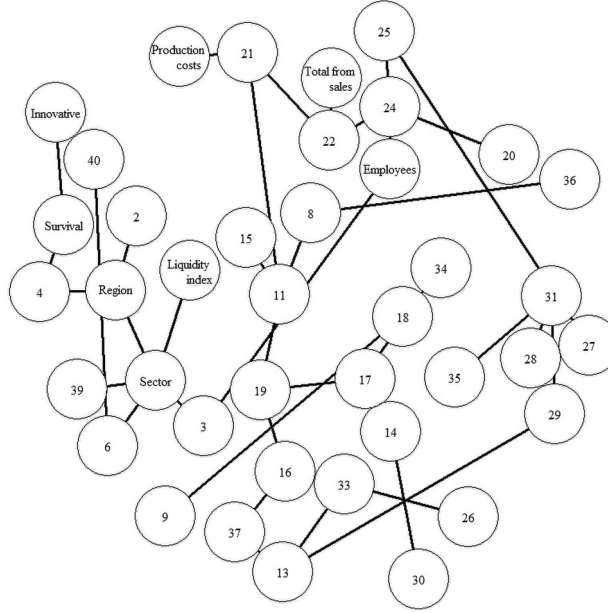
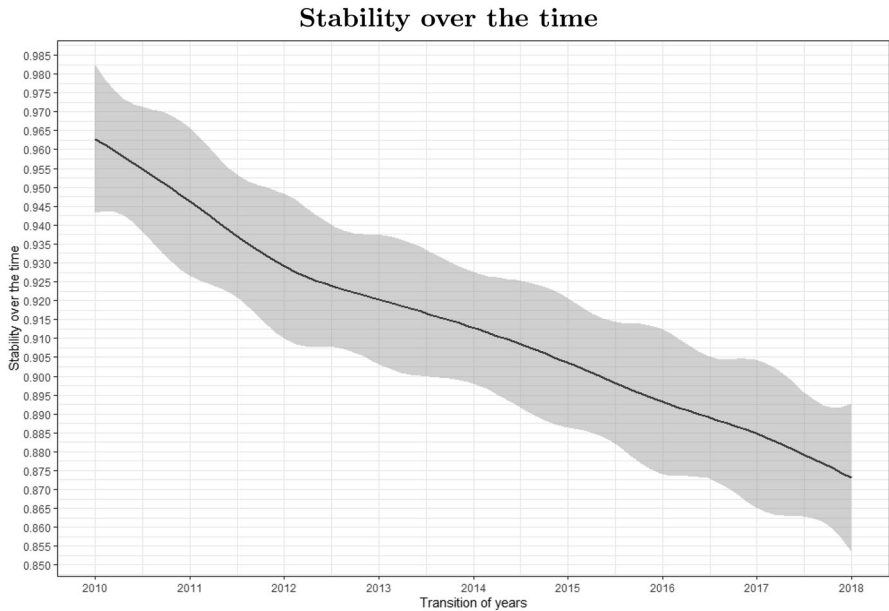


Fig. 5 Graphical models for Italian startup over 2009–2010



**Fig. 6** Evolution of *stability* in the dataset (black line) and the grey area indicates the credible interval

**Table 3** Posterior summaries for the regression parameters of the Stability

Coefficients	Mean	SD	Lower C.I..5%	Upper C.I..95%
$\beta_0$	1812.46	294.5	1376.05	2329.22
$\beta_t$	-0.90	0.15	-1.15	-0.68
$\beta_{2^T-1}$	5.12	5.91	-1.51	18.26

Table 7 in Appendix 1 reports the result of the multivariate *Cox* regression, in which sector *G* (wholesale and retail trade; repair of motor vehicles and motorcycles) is the reference level for the variable *Sector*, while for the variable *Region* the reference is *Lombardia*. The results are consistent with the literature and all of the selected variables are significant. Risky ventures such as innovative firms have a lower chance to survive as well as firms with liquidity constraints. Firms in most of the regions and sectors have a lower chance to survive vis á vis firms in *Lombardia* and in the automotive, respectively. These results assume both a stable relation among variables over a time span of 10 years, as well as stable coefficients. The result obtained from the traditional survival analysis is based on the assumptions of the *Cox* proportional hazard model which can be summarized in the following way: at any time *t* all observations are assumed to have the same baseline, the effects of the regression variables remain constant over time and eventually the regression coefficients do not vary with time (Cox, 1975). These assumptions easily can be checked using the *Schoenfeld Residuals* (Grambsch & Therneau, 1994). In fact, Grambsch and Therneau (1994) have shown that many of the popular tests for proportional hazards are essentially tests for nonzero slopes in a generalized linear regression of the rescaled residuals on chosen function(s) of time.

Table 2 reports the result of the Schoenfeld (1982) test (Grambsch & Therneau, 1994), in which for each of the covariates and for the global test we reject the null hypothesis of proportional hazards. This result reinforces the idea of the presence of a statistical drift in data. A possible solution to this significant difference across the periods is a piecewise exponential model (Cefis & Marsili, 2019). In detail, this model leads to a flexible or semi-parametric approach to fitting a survival model when the cox assumptions do not hold. This is a semi-parametric approach, that allows for penalized estimation of very flexible survival models (Gasparrini, 2014). Given the formulation of the proportional hazard model in the Eq. (1), we can rewrite the piecewise exponential model as:

$$\lambda_{i,k} = \lambda_k \exp \sum_{j=1}^p \beta_j \mathbf{X}(t)_{i,j} \quad (4)$$

where  $\lambda_{i,k}$  is the hazard of subject  $i$  in the period  $k$ ,  $\lambda_k$  is the baseline hazard of such a time period (Cefis & Marsili, 2019). Following Bender and Scheipl (2018), the log-baseline in Eq. (4) can be splitted in two terms:

$$\log \lambda_{i,k} = a_k + \sum_{j=1}^p \beta_j \mathbf{X}(t)_{i,j}$$

Thus, it is possible to select time-constant covariates and time-dependent covariates as in the following specification:

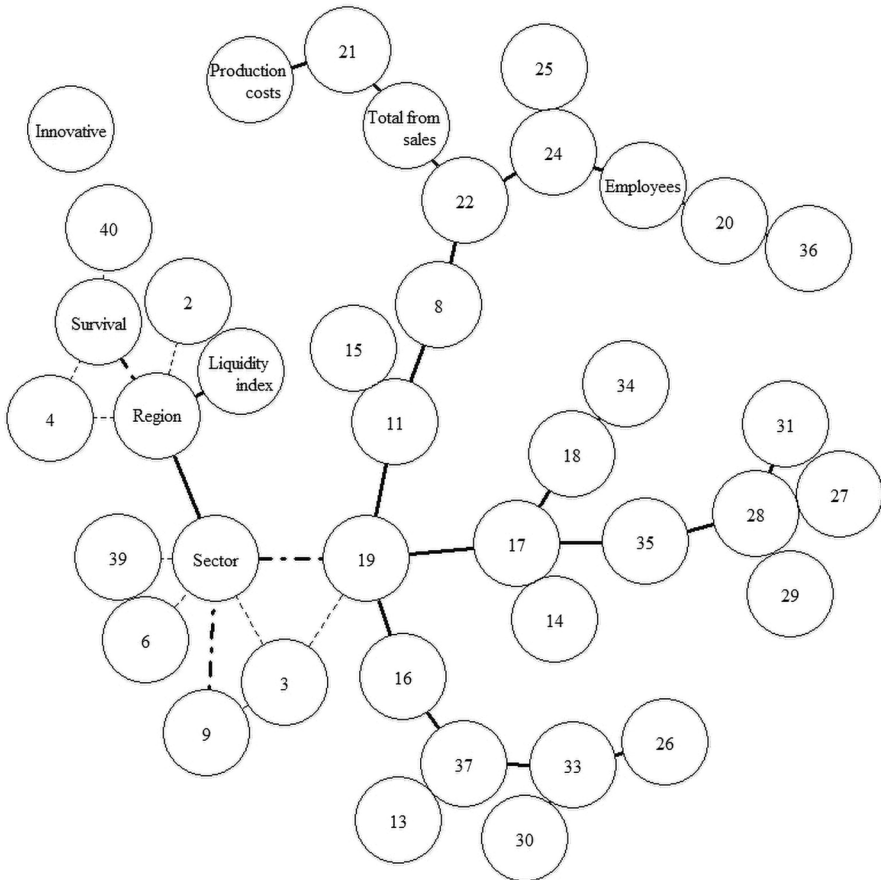
$$\log \lambda_i(t) = f_0(t) + \sum_{p=1}^P f_p(\mathbf{x}_{i,p}, t) + \sum_{m=1}^M g_m(\mathbf{z}_{i,m}, t) \quad (5)$$

where the term  $f_p(\mathbf{x}_{i,p}, t)$  denotes time-varying effects of time-constant covariates  $\mathbf{x}_{i,p}$  and the term  $g_m(\mathbf{z}_{i,m}, t)$  represents time-dependent covariates, while  $f_0(t)$  the baseline of the model. In our case, the use of this approach is to test the presence of time-varying effects, where the effects of continuous covariates are assumed to vary non-linearly in time, but linearly in the covariates. In this way, we can reinforce the idea that the process of firm survival is a dynamic one and it includes potential changes in market forces and in the surrounding environment. Figure 4 reports the time-varying effects of selected variables and compare with the result of a traditional Cox-regression in the last column,<sup>3</sup>. The most interesting results is that for the selected variables the time-varying effect exhibit a change in the direction of the effect over time, suggesting that the Cox-model's time average effect are not very informative. In the next section we apply the theoretical framework presented in

<sup>3</sup> Table 8 in the appendix reports all the effects. The effects of variables *Region Sector* and *Innovative Startup* were estimated as linearly time-varying effects (Time varying effects=1) and they must be interpreted as relative changes compared to the baseline hazard (*Year of survival*), which itself is non-linear and statically significant. For time-dependent covariates, the variable *Employees* is not linear and non significant while *Production Cost*, *Total From Sales* and *Index Liquidity* are non-linear and statically significant.



### Spanning Tree 2009 Pruning



**Fig. 7** Pruning GM 2009, solid black lines indicate the original connection, while dot–dash lines indicate the connection between the variables after the pruning

Sect. 2 to the same data and present an alternative view in which the variable selection is computer-aided and stability is not taken for granted. We thus make a time varying impact, but we test whether, when many variables are available, the variable selection should be stable overtime or not.

### 4.3 Application of graphical models

Section 3 introduced GM as a method to map the conditional dependence structure of a dataset, to evaluate its stability overtime, and to select variables for including in (eventual) econometric analysis. In this section we apply the method to the data at hand and examine qualitative differences with the results derived from canonical survival analyses presented in the Sect. 4.2.

### 4.3.1 A graphical model of new Italian firms

The starting point of the proposed method is the construction of GMs for each year using the algorithm proposed by Edwards et al. (2010). Figure 5 provides a close-up view of the inferred GMs obtained using the algorithm proposed by Edwards et al. (2010) for the years 2009 and 2010. Each node in the graph represents a variable in the dataset, and the absence of a link indicates a relation of independence. If two nodes are connected through a third variable, it means they have a dependent relationship, conditioned on the third variable.

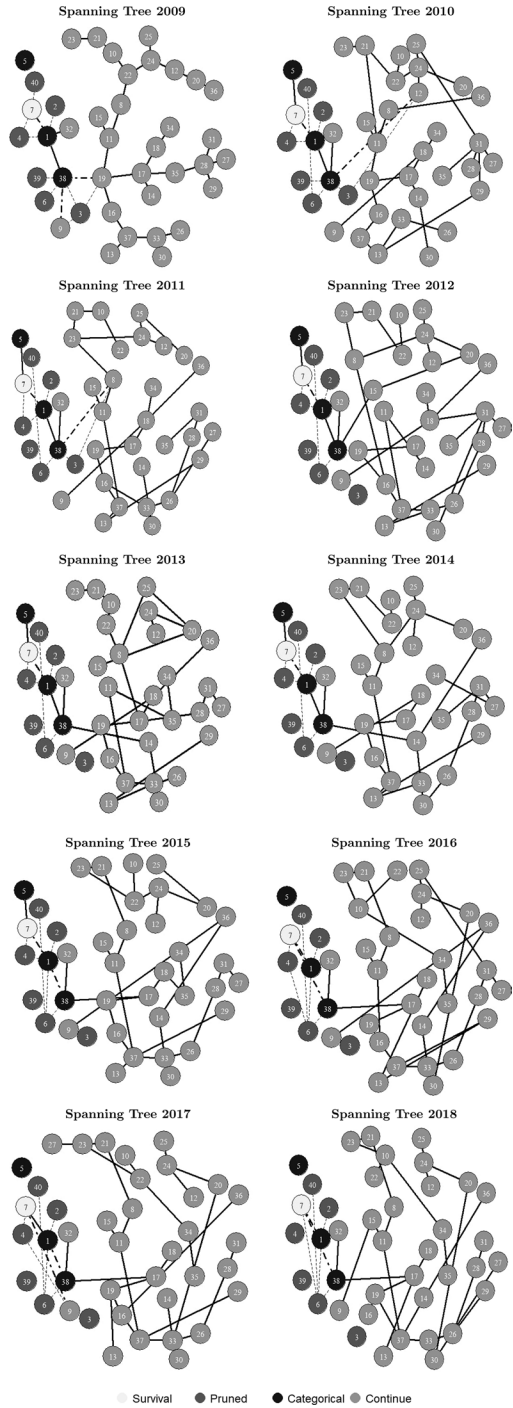
In this figure, for the sake of clarity, we have labeled the variables used in traditional survival analysis with their names, while labels for the remaining variables can be found in Table 4. The focal point of our study is the dummy variable representing firm survival, denoted as node 7 and labeled as ‘*Survival*.’ Notably, the graphs differ across time periods. For example, the variable ‘*Innovative*,’ a dummy variable identifying innovative start-ups, exhibits conditional independence from survival in 2009 but not in 2010. To assess how the relationships between variables evolve over time, we calculate the drift of the inferred graphs and visualize it in Fig. 6. This metric compares the GM inferred for a given year with the one inferred in the preceding year starting from 2009 (Riso & Guerzoni, 2022).

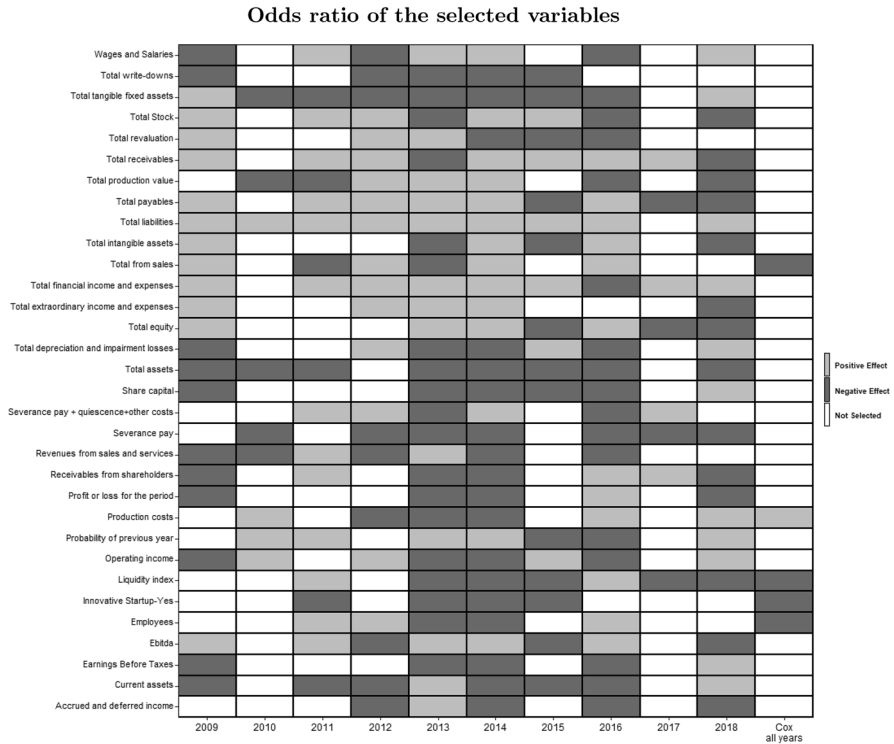
Figure 6 and Table 3 show the results of the drift analysis. The line in Fig. 6, that decreases monotonically, indicates that with each subsequent period, the dependency structure of the variables, as depicted by the graphical model, diverges more from that of the first period. The trend line’s shape conveys two pieces of information: First, it signals the presence of a drift. Secondly, its consistent decrease suggests that this isn’t a result of random events in specific years, but rather an intrinsic characteristic of the process. This observation aligns with our expectations, considering we are monitoring the same firms over time. As these firms grow, fail, or prosper, they gradually alter the relationships of their characteristics. Table 3 shows the posterior summaries for the regression parameters in this data-set: the mean intercept value  $\beta_0 = 1812.46$  indicates that despite the presence of drift, the *stability* index is high in the periods under observation, while the mean of  $\beta_1 = -0.9$  portrays a risk of drift over the long term (Riso & Guerzoni, 2022). Eventually, the mean  $\beta_{2T-1} = 5.12$  captures the component of *stability* which originates in the persistence of existing connections with respect to the component of *stability* which originates in absence of connections. Given the evidence of drift, we can conclude that the Cox regression model is not appropriate and that the procedure for variable selection and regression should be executed for each individual period.

### 4.3.2 Variable selection

The automated process generating the Fig. 5 can help the researcher in identifying variables to include in the econometric analysis. However, as explained above, we allow for the researcher to modify the basic GM in order to include or exclude variables. In this way, the process is transparent since any deviation from the basis GM needs to be motivated.

**Fig. 8** Graphical models for Italian startup survival over 10 years. Solid black lines indicate the original connection, while dot–dash lines indicate the connection between the variables after the pruning





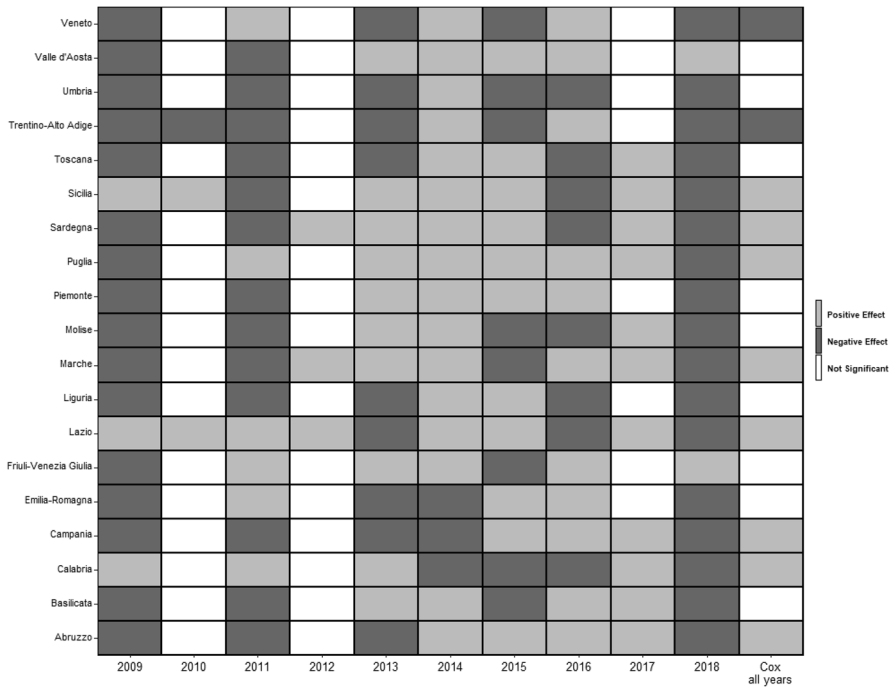
**Fig. 9** Impact on probability of survival

In the case at hand we decided to exclude six variables. In what follows, we give reasons for the exclusion of each of them:

- *Province* (2) direct dependence with the variable *Region*, presence of many levels
- *Legal form* (3) presence of many levels, among which many are not informative.
- *Legal status* (4) Presence of many non informative levels, not mentioned in the literature, data non reliable since they do derive from administrative process.
- *Artisan Companies* (6): zero-inflated, non informative, not mentioned in the literature.
- *Constitution quarter* (40): not informative since in Italy, the timing of constitution relates more with administrative deadlines.
- *Sales description* (39): many non-informative levels, not mentioned in the literature.

This process in which we remove variables from the model and generate a new link to account for the omitted variable is called pruning. This process generates a new

### Odds ratio of the Region levels



**Fig. 10** Impact on probability of survival of *Regions* Reference level: region *Lombardia*

GM for each year, as represented in Fig. 7 for the year 2009. In Fig. 7 solid black lines indicate the original connection, while dot-dash lines indicate the connection between the variables after the pruning.

For instance, node 4 (*Legal Status*) mediates the impact of *Region* on *Survival*. This means that the survival is conditionally independent of the region once we remove the legal status variable. Thus, we need to add a link between *Survival* and *Region* following the elimination of the legal status variable from the model.

Figure 8 depicts the GMs generated using the algorithm proposed by Edwards et al. (2010) over a 10-year period following the pruning process. Various sets of variables are color-coded for clarity. Specifically, the legend in Fig. 8 categorizes the variables into four groups: ‘Survival’, which represents the variable of interest (node 7, *Survival*), ‘Pruned,’ encompassing all variables involved in the pruning process, ‘Categorical,’ which includes the variables *Region* (node 1), *Innovative* (node 5), and *Sector* (node 38), and finally, the ‘Continuous’ group, which comprises all continuous variables. The correspondence between node labels and variables is detailed in Table 4.

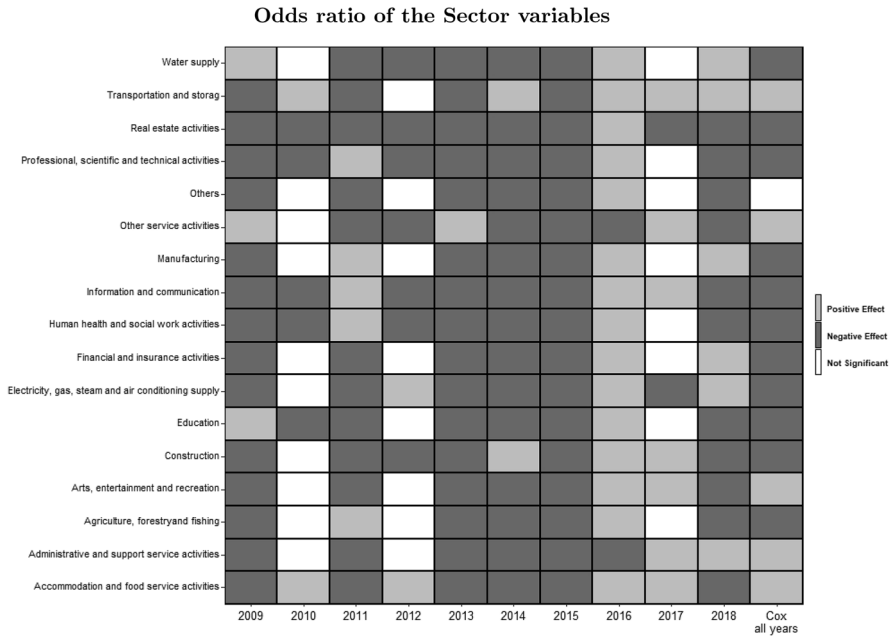


Fig. 11 Impact on probability of survival of Sector Reference level: sector G

The variable selection process applied to the GMs after the pruning process (Fig. 8) is conducted using the Best Path Algorithm (BPA), as proposed by Riso et al. (2023). The BPA leverages the probabilistic structure of the GM to identify the optimal set of variables for explaining or predicting a target variable (in our case, the variable Survival). Assuming the existence of at least one relationship between the target variable and the variables in the dataset implies that there is at least one variable, directly or indirectly linked to the target variable through one or more intermediary variables. This assumption is equivalent to supposing the presence of distinct subsets of variables. The primary objective of the BPA (Riso et al., 2023) is to identify the most suitable subset of variables that exhibits the highest Mutual Information (MI) (Lewis, 1959) with the target variable. The MI between a specific subset of variables and the target variable is quantified using the Kullback–Leibler information index, which measures the divergence between the probability density function of the target variable and its projection onto the subspace defined by the variables in the set. This measure aligns with the concept of entropy coefficient of determination (Eshima & Tabata, 2007, 2010), which simplifies the standard coefficient of determination in the case of a linear relationship between the variables. In order to further increase models reliability, we decide to add an additional variable in the forthcoming econometric analysis beyond the variables selected by the automatic

process. We add the estimated probability of survival in the previous period as a way to deal with the auto-correlation in the survival process. We do this since each graph is a snapshot in time and does not include information from previous years which could be useful for analyzing survival. We estimate the probability of survival in the previous year, as a propensity score employing all variables selected by the algorithm in the previous year. This algorithm identifies (potentially) a different number of regressors each year. These regressors are then used to estimate the probability of survival (for a given year) with a logistic regression:

$$\log \left( \frac{\rho_{i,t}}{1 - \rho_{i,t}} \right) = \beta_0 + \sum_j^N \beta_{j,t} x_{j,t} \quad \text{for } t = 2009, \dots, 2018, \quad (6)$$

where on the left-hand side, we have the odds ratio of surviving explained by the set of selected variables  $X_t$ . Since in the process of variable selection we now include the probability of survival in the previous year, we can observe if it is uncorrelated with the present probability of surviving or if it has an effect on the present probability of surviving. In this way, we control not only if the likelihood of surviving depends on the past history of the firms, but also study the direction of this effect.

### 4.3.3 Results

In this section we present the results of the variables selection carried out through the algorithm presented in the Sect. 3.2. The Fig. 9 shows the selected variables and the sign of the coefficients, while Figs. 10 and 11 focus on the results for all the levels of the variables *Sector* and *Region* respectively. In the three figures, the last column summarizes the result of the Cox regression for comparison. In Fig. 9, gray color denotes coefficients with positive coefficients, black signifies coefficients with negative coefficients, while white corresponds to variables not selected for the analysis. It is very important to notice that the variable selection approach allow for the inclusion of many more variables than the traditional analysis, the variables selected change every year and the significance and sign of some variables change across periods. In Figs. 10 and 11, gray and black shading has the same interpretation as in Fig. 9, while white denotes variables that are included in the analysis but are estimated to be statistically insignificant.

In line with the conjecture articulated by Cefis et al. (2021), our results show significant dynamics over the years. We see multiple variables changing their significance levels and even flipping coefficient signs across the 10-year period of study. Traditional survival analysis, highlight the role of firm innovativeness, size, cost of production, revenues, as well as that of the region and sector. This alternative approach does not outright dismiss any of these variables, but rather provides a richer, finer-grained view exposing marked changes across time. At an aggregate level, survival in 2010 and 2017 does not seem to be captured by the current

variables as if in these years other events (external to our, and in general to survival models) were responsible for Italian firms' exit.

A closer scrutiny reveals that firm innovativeness is an important predictor of survival, however this is not true in every single year. This finer-grained view helps to explain mild significance of sector and region variables in the traditional survival analysis. In both cases we see strong heterogeneity of effects across years within virtually every category. For instance the non significant coefficient of the region *Piemonte* in the Cox regression could be the result of the varying impact of the coefficient over time, as elicited by our method. Theoretical implications of these results are important. Results support the argument that the underlying mechanisms of firm dynamics are much more complex than assumed in a traditional survival analysis. There seem to be strong unobservable factors influencing firm survival even in the short run. In the conclusion, we summarize the results and derive the policy implications from the empirical evidence presented.

## 5 Summary and conclusion

This paper contributes to the ongoing attempt in the literature of combining data science tools with a traditional econometric approach. We propose an alternative method of survival analysis which is particularly suited for high-dimensional data. It allows for automatic feature selection for the estimation of an econometric model. It also allows for including the researcher in the loop of the otherwise automated process. The researcher carries the burden of modifying and approving the final set of model variables. To conclude, the researcher is responsible for making sure the model is built on sound economic (as well as econometric) theory.

The proposed methodology addresses two important current issues in survival models. First, the increasing availability of data with a large number of variables makes the process of variable selection both cumbersome and, in some cases, opaque. Secondly, the hidden socio-economic context not captured by the observable data might change over time which makes the assumption of the stable data generation process implausible. Although, we cannot capture the hidden context, we can deal with its impact. We employ graphical models as a tool for both empowering the research in the process of variable selection and for testing the stability of the explanatory variables of firm survival over time. When applying this method and comparing it with a traditional survival exercise, results are striking. While the traditional methodology designates some variables as stable determinants of firm survival, a fine-grained analysis using the innovative approach suggests that above-mentioned variables explain firm survival only in a handful of years.

An important advantage of the proposed new algorithm of variable selection is transparent and significantly reduces the risk of selective reporting and p-hacking from the side of a researcher. Variable shortlisting is performed by unsupervised learning method of graphical modeling. Under the assumption of prior choice of variable remoteness by the researcher, this algorithm automatically generates the list of candidate variables to be included in the statistical model. This least could further



be adjusted by the researcher allowing her to introduce theoretical considerations in otherwise data-driven analysis.

However, the proposed methodology has a significant shortcoming. Due to the specificity of graphical models, categorical variables cluster in the generated variable relation graphs. In other words, the method used in this paper does not allow for the possibility of the relationship between two categorical variables to be mediated by a numeric variable. This is significant as, under the condition that our variable of interest (firm survival) is a categorical variable, it pushes categorical covariates to more pronounced positions in the estimated graphical model. In order to overcome this problem, the researcher is advised to keep the categorical variables in survival analysis to the minimum, or be open to estimating more complex models with higher number of variables. Both of these approaches would guarantee that important numerical variables make through the pre-set selection threshold into the ultimate econometric model. This shortcoming could also be moderated by further developments in machine learning techniques, and more precisely in graphical modeling. New methods could allow for the relaxation of the variable-type clustering in resulting graphical models.

As for the economic interpretation of the results, we have emphasized that the relationship between different survival drivers is more complex than typically discussed in previous literature. If we take this consideration seriously, it means that especially for businesses in their early years, the mechanisms determining their success or failure depend on the interaction of many time-volatile factors. Public policies aimed at supporting business survival, especially in the case of startups, should develop a very specific understanding of the realities rather than relying on generic literature. Furthermore, national-level policy programs must necessarily be complemented by regional-level policies to ensure flexibility in intervention tools. This conclusion comes with some caveats. Our sample focused solely on startups, where it's normal to expect more volatility and complexity. Probably, considering a sample of businesses that also includes existing enterprises, the issue of drift may be less significant, and survival-related factors may be less pressing. We hope to continue this research along this line of analysis and include more firms and an extended time span.

## Appendix

### Data

In this section are reported the details of missing data (Table 4), the frequencies distributions for *Regions* and *Sector* (Tables 5 and 6). Table 7 shows coefficients of the Cox Regression.

**Table 4** Name of the variables and percent of missing for year

Name of the variable	Node label	Variable type	Percentage of missing data													
			2009	2010	2011	2012	2013	2014	2015	2016	2017	2018				
Region	1	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Province	2	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Legal form	3	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Legal state	4	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Innovative startup	5	Dichotomous	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Artisan companies	6	Dichotomous	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Death (survival)	7	Dichotomous	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total receivables	8	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Receivables from shareholders	9	Continue	17%	4%	4%	4%	4%	4%	4%	3%	3%	2%	2%	1%	1%	0%
Total from sales	10	Continue	20%	5%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Current assets	11	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Employees	12	Continue	28%	26%	12%	8%	7%	4%	4%	4%	4%	4%	3%	3%	3%	3%
Total tangible fixed assets	13	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Total intangible assets	14	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Total stock	15	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Total assets	16	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Total equity	17	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Share capital	18	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Total liabilities	19	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Severance pay	20	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Total production value	21	Continue	20%	5%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Revenues from sales and services	22	Continue	20%	5%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Production costs	23	Continue	17%	4%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%
Wages and salaries	24	Continue	20%	5%	4%	4%	4%	3%	3%	3%	2%	2%	2%	1%	1%	0%

Table 4 (continued)

Name of the variable	Node label	Variable type	Percentage of missing data											
			2009	2010	2011	2012	2013	2014	2015	2016	2017	2018		
Severance pay + quiescence + other costs	25	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%		
Operating income	26	Continue	17%	4%	4%	4%	3%	3%	2%	1%	0%			
Total extraordinary income and expenses	27	Continue	20%	5%	4%	4%	3%	3%	2%	1%	0%			
Earnings before taxes	28	Continue	17%	4%	4%	4%	3%	3%	2%	1%	0%			
Total financial income and expenses	29	Continue	17%	4%	4%	4%	3%	3%	2%	1%	0%			
Total depreciation and impairment losses	30	Continue	17%	4%	4%	4%	3%	3%	2%	1%	0%			
Profit or loss for the period	31	Continue	20%	5%	4%	4%	3%	3%	2%	1%	0%			
Liquidity index	32	Continue	27%	10%	9%	9%	8%	8%	7%	7%	6%			
Ebitda	33	Continue	17%	4%	4%	4%	3%	3%	2%	1%	0%			
Total revaluation	34	Continue	45%	30%	24%	17%	9%	3%	2%	1%	0%			
Total write-downs	35	Continue	20%	5%	4%	4%	3%	3%	2%	1%	0%			
Accrued and deferred income	36	Continue	17%	4%	4%	4%	3%	3%	2%	1%	0%			
Total payables	37	Continue	17%	4%	4%	4%	3%	3%	2%	1%	0%			
Sector	38	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%			
Sae description	39	Discrete	1%	1%	1%	0%	0%	0%	0%	0%	0%			
Constitution quarter	40	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%			

**Table 5** Distribution of the firms for region

Region	Number of firms born in 2009
Lombardia	13,430
Veneto	5453
Emilia-Romagna	5079
Piemonte	3552
Liguria	1408
Trentino-Alto Adige	1204
Friuli-Venezia Giulia	943
Valle D'Aosta	132
<b>North</b>	<b>31,201</b>
Lazio	12,217
Toscana	4435
Marche	1975
Abruzzo	1631
Umbria	987
<b>Center</b>	<b>21,245</b>
Campania	7645
Sicilia	4766
Puglia	4580
Calabria	1614
Sardegna	1583
Basilicata	590
Molise	317
<b>South</b>	<b>21,095</b>

**Table 6** Frequency of the sector

Label	Frequency	Description
A	1426	Agriculture, forestry and fishing
B	64	Mining and quarrying
C	7933	Manufacturing
D	1936	Electricity, gas, steam and air conditioning supply
E	317	Water supply; sewerage, waste management and remediation activities
F	12,214	Construction
G	14,955	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	2923	Transportation and storage
I	5053	Accommodation and food service activities
J	3192	Information and communication
K	1433	Financial and insurance activities
L	6422	Real estate activities
M	6703	Professional, scientific and technical activities
N	4117	Administrative and support service activities
O	2	Public administration and defence; compulsory social security
P	716	Education
Q	1428	Human health and social work activities
R	1588	Arts, entertainment and recreation
S	1117	Other service activities
U	2	Activities of extraterritorial organizations and bodies

**Table 7** Cox regression summary

Variables	Coefficient	exp(Coef)	SE(Coef)	Robust SE	$z$	Pr(> z )	Signif.
Region Abruzzo	1.18e−01	1.12e+00	3.23e−02	3.26e−02	3.609	3.07e−04	***
Region Basilicata	−8.21e−03	9.92e−01	5.29e−02	5.15e−02	−0.159	0.873	
Region Calabria	1.26e−01	1.13e+00	3.23e−02	3.30e−02	3.809	1.40e−04	***
Region Campania	1.37e−01	1.15e+00	1.79e−02	1.86e−02	7.357	1.89e−13	***
Region Emilia-Romagna	4.09e−03	1.00e+00	2.11e−02	2.13e−02	0.192	0.847	
Region Friuli-Venezia Giulia	−6.11e−02	9.41e−01	4.39e−02	4.39e−02	−1.393	0.163	
Region Lazio	2.13e−01	1.24e+00	1.56e−02	1.61e−02	13.209	2e−16	***
Region Liguria	5.57e−02	1.06e+00	3.53e−02	3.61e−02	1.543	0.123	
Region Marche	6.59e−02	1.07e+00	3.02e−02	3.07e−02	2.144	0.032	*
Region Molise	8.57e−02	1.09e+00	6.96e−02	6.76e−02	1.268	0.205	
Region Piemonte	1.46e−02	1.01e+00	2.42e−02	2.45e−02	0.597	0.551	
Region Puglia	1.28e−01	1.14e+00	2.13e−02	2.22e−02	5.780	7.46e−09	***
Region Sardegna	1.95e−01	1.22e+00	3.21e−02	3.29e−02	5.936	2.92e−09	***
Region Sicilia	1.92e−01	1.21e+00	2.07e−02	2.18e−02	8.800	2e−16	***
Region Toscana	−6.24e−03	9.94e−01	2.21e−02	2.23e−02	−0.280	0.780	
Region Trentino-Alto Adige	−3.55e−01	7.01e−01	4.31e−02	4.28e−02	−8.292	2e−16	***

**Table 7** (continued)

Variables	Coefficient	exp(Coef)	SE(Coef)	Robust SE	z	Pr(> z )	Signif.
Region Umbria	5.44e-02	1.06e+00	4.13e-02	4.15e-02	1.312	0.190	
Region Valle d'Aosta	-9.14e-02	9.13e-01	1.14e-01	1.17e-01	-0.780	0.436	
Region Veneto	-8.76e-02	9.16e-01	2.10e-02	2.12e-02	-4.130	3.63e-05	***
Sector A	-4.14e-01	6.61e-01	3.77e-02	4.01e-02	-10.311	2e-16	***
Sector Others	-1.63e-01	8.50e-01	1.57e-01	1.55e-01	-1.053	0.292	
Sector C	-6.59e-02	9.36e-01	1.77e-02	1.93e-02	-3.42	6.26e-04	***
Sector D	-1.72e-01	8.42e-01	3.18e-02	3.50e-02	-4.897	9.75e-07	***
Sector E	-3.66e-01	6.93e-01	7.90e-02	7.93e-02	-4.616	3.91e-06	***
Sector F	-6.60e-02	9.36e-01	1.51e-02	1.91e-02	-3.448	5.64e-04	***
Sector H	1.23e-01	1.13e+00	2.46e-02	2.75e-02	4.474	7.66e-06	***
Sector I	1.54e-01	1.17e+00	1.95e-02	2.70e-02	5.704	1.17e-08	***
Sector J	-1.28e-01	8.80e-01	2.46e-02	2.79e-02	-4.584	4.56e-06	***
Sector K	-8.20e-02	9.21e-01	3.46e-02	3.80e-02	-2.157	0.031	*
Sector L	-5.30e-01	5.89e-01	2.06e-02	2.60e-02	-20.369	2e-16	***
Sector M	-6.11e-02	9.41e-01	1.85e-02	2.26e-02	-2.701	0.007	**
Sector N	8.03e-02	1.08e+00	2.14e-02	2.68e-02	3.000	0.003	**
Sector P	-2.78e-01	7.57e-01	5.03e-02	5.38e-02	-5.158	2.49e-07	***
Sector Q	-3.96e-01	6.73e-01	3.77e-02	4.18e-02	-9.467	2e-16	***
Sector R	1.65e-01	1.18e+00	3.09e-02	3.44e-02	4.788	1.69e-06	***
Sector S	1.14e-01	1.12e+00	3.67e-02	4.16e-02	2.731	0.006	**
Innovative startup-yes	-1.99e+00	1.36e-01	3.33e-01	3.19e-01	-6.252	4.05e-10	***
Production costs	2.96e-05	1.00e+00	4.43e-06	8.48e-06	3.495	0.001	***
Total from sales	-1.13e-04	1.00e+00	6.52e-06	3.51e-05	-3.208	0.002	**
Index liquidity	-2.58e-02	9.74e-01	3.03e-03	3.35e-03	-7.717	1.19e-14	***
Employees	-7.58e-03	9.92e-01	7.35e-04	3.18e-03	-2.384	0.017	*

Significant code:  $p^{***} < 0.0001$ ,  $p^{**} < 0.001$ ,  $p^* < 0.05$ ,  $p < 0.1$

## Statistical results

See Table 8.

**Table 8** Approximate significance of smooth terms

Variables	Time varying effects	$\chi^2$	p-value
Year of survival	3.80	52.78	< 2e - 16***
Region	1.00	0.39	0.534
Sector	1.00	0.30	0.583
Innovative startup	1.00	0.01	0.936
Production costs	5.21	133.29	< 2e - 16***
Revenus	5.28	121.99	< 2e - 16***
Index liquidity	8.72	54.94	< 2e - 16***
Employees	2.00	1.78	0.411

**Funding** Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement.

**Data availability** The data that support the findings of this study are available from Bureau van Dijk Electronic Publishing, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Bureau van Dijk Electronic Publishing.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Acs, Z. J., Armington, C., & Zhang, T. (2007). The determinants of new-firm survival across regional economies: The role of human capital stock and knowledge spillover. *Papers in Regional Science*, 86(3), 367–391.
- Agostino, M., Scalera, D., Succurro, M., & Trivieri, F. (2021). Research, innovation, and bankruptcy: Evidence from European manufacturing firms. *Industrial and Corporate Change*, 31(1), 137–160.
- Aldinucci, M., Bagnasco, S., Lusso, S., Pasteris, P., Rabellino, S., & Vallero, S. (2017). Occam: A flexible, multi-purpose and extendable hpc cluster. *Journal of Physics: Conference Series*, 8, 082039.
- Aldinucci, M., Rabellino, S., Pironi, M., Spiga, F., Viviani, P., Drocco, M., Guerzoni, M., Boella, G., Mellia, M., Margara, P. et al. (2018). Hpc4ai: An ai-on-demand federated platform endeavour. In *Proceedings of the 15th ACM International Conference on Computing Frontiers* (pp. 279–286).
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609.
- Altman, N., & Krzywinski, M. (2018). The curse (s) of dimensionality. *Nature Methods*, 15(6), 399–400.
- Audretsch, D. B. (1995). Innovation, growth and survival. *International Journal of Industrial Organization*, 13(4), 441–457.
- Bargagli-Stoffi, F. J., Niederreiter, J., & Riccaboni, M. (2021). Supervised learning for the prediction of firm dynamics. In *Data Science for Economics and Finance* (pp. 19–41). Springer.
- Barnard, C. I. (1938). *The functions of the executive*. Cambridge, MA: Harvard University.
- Basile, R., Pittiglio, R., & Reganati, F. (2017). Do agglomeration externalities affect firm survival? *Regional Studies*, 51(4), 548–562.
- Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial education*, 33, 1–42.
- Bender, A., & Scheipl, F. (2018). Pammtools: Piece-wise exponential additive mixed modeling tools. arXiv preprint [arXiv:1806.01042](https://arxiv.org/abs/1806.01042)
- Carota, C., Durio, A., & Guerzoni, M. (2015). An application of graphical models to the innobarometer survey: A map of firms' innovative behaviour. *Italian Journal of Applied Statistics*, 25(1), 61–79.
- Cefis, E., & Marsili, O. (2005). A matter of life and death: Innovation and firm survival. *Industrial and Corporate Change*, 14(6), 1167–1192.
- Cefis, E., & Marsili, O. (2019). Good times, bad times: Innovation and survival over the business cycle. *Industrial and Corporate Change*, 28(3), 565–587.
- Cefis, E., Bettinelli, C., Coad, A., & Marsili, O. (2021). Understanding firm exit: A systematic literature review. *Small Business Economics*, 59, 423–446.
- Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3), 462–467.
- Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 28(3), 591–605.

- Chuard, P. J., Vrtílek, M., Head, M. L., & Jennions, M. D. (2019). Evidence that nonsignificant results are sometimes preferred: Reverse p-hacking or selective reporting? *PLoS Biology*, *17*(1), e3000127.
- Coad, A., Frankish, J., Roberts, R. G., & Storey, D. J. (2013). Growth paths and survival chances: An application of gambler's ruin theory. *Journal of Business Venturing*, *28*(5), 615–632.
- Coeurderoy, R., Cowling, M., Licht, G., & Murray, G. (2012). Young firm internationalization and survival: Empirical tests on a panel of 'adolescent' new technology-based firms in Germany and the UK. *International Small Business Journal*, *30*(5), 472–492.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.
- Crosato, L., Domenech, J., & Liberati, C. (2021). Predicting sme's default: Are their websites informative? *Economics Letters*, *204*, 109888.
- de Abreu, G. C., Labouriau, R., & Edwards, D. (2009) High-dimensional graphical model search with graphd r package. arXiv preprint [arXiv:0909.1234](https://arxiv.org/abs/0909.1234).
- De Jay, N., Papillon-Cavanagh, S., Olsen, C., El-Hachem, N., Bontempi, G., & Haibe-Kains, B. (2013). mrmre: an r package for parallelized mrmr ensemble feature selection. *Bioinformatics*, *29*(18), 2365–2368.
- Delmar, F., McKelvie, A., & Wennberg, K. (2013). Untangling the relationships among growth, profitability and survival in new firms. *Technovation*, *33*(8–9), 276–291.
- Delmar, F., Wallin, J., & Nofal, A. M. (2022). Modeling new-firm growth and survival with panel data using event magnitude regression. *Journal of Business Venturing*, *37*(5), 106245.
- Edwards, D., De Abreu, G. C., & Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal aic or bic forests. *BMC Bioinformatics*, *11*(1), 18.
- Eklund, J., Karlsson, S., et al. (2007). An embarrassment of riches: Forecasting using large panels, Central Bank of Iceland Working Paper series Nr. 34/2007.
- Eshima, N., & Tabata, M. (2007). Entropy correlation coefficient for measuring predictive power of generalized linear models. *Statistics & Probability Letters*, *77*(6), 588–593.
- Eshima, N., & Tabata, M. (2010). Entropy coefficient of determination for generalized linear models. *Computational Statistics & Data Analysis*, *54*(5), 1381–1389.
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, *46*(4), 1–37.
- Gasparrini, A. (2014). Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics in Medicine*, *33*(5), 881–899.
- Geroski, P. (1992). Entry, exit and structural adjustment in European industry. In *European industrial restructuring in the 1990s* (pp. 139–161). Springer.
- Geroski, P. A. (1995). What do we know about entry? *International Journal of Industrial Organization*, *13*(4), 421–440.
- Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT Press.
- Giot, P., & Schwienbacher, A. (2007). Ipos, trade sales and liquidations: Modelling venture capital exits using survival analysis. *Journal of Banking & Finance*, *31*(3), 679–702.
- Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526.
- Grazzi, M., Piccardo, C., & Vergari, C. (2021). Turmoil over the crisis: Innovation capabilities and firm exit. *Small Business Economics*, *59*, 537–564.
- Guerzoni, M., Nava, C. R., & Nuccio, M. (2020). Start-ups survival through a crisis. Combining machine learning with econometrics to measure innovation. *Economics of Innovation and New Technology*, *30*, 468–493.
- Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, *69*(3), 553–566.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106.
- Holtz-Eakin, D., Joulfaian, D., & Rosen, H. S. (1994). Sticking it out: Entrepreneurial survival and liquidity constraints. *Journal of Political Economy*, *102*(1), 53–75.
- Hyytinen, A., Pajarinen, M., & Rouvinen, P. (2015). Does innovativeness reduce startup survival rates? *Journal of Business Venturing*, *30*(4), 564–581.
- Jordan, M. I., et al. (2004). Graphical models. *Statistical Science*, *19*(1), 140–155.
- Jung, H., Hwang, J., & Kim, B.-K. (2018). Does r & d investment increase SME survival during a recession? *Technological Forecasting and Social Change*, *137*, 190–198.



- Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle. *The American Economic Review*, 86(3), 562–583.
- Koller, D., Friedman, N., Džeroski, S., Sutton, C., McCallum, A., Pfeffer, A., Abbeel, P., Wong, M.-F., Meek, C., Neville, J., et al. (2007). *Introduction to statistical relational learning*. MIT Press.
- Kratzer, G., & Furrer, R. (2018) varrank: An r package for variable ranking based on mutual information with applications to observed systemic datasets. arXiv preprint [arXiv:1804.07134](https://arxiv.org/abs/1804.07134).
- Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1), 48–50.
- Kyle, R. A., Gertz, M. A., Greipp, P. R., Witzig, T. E., Lust, J. A., Lacy, M. Q., & Therneau, T. M. (1997). A trial of three regimens for primary amyloidosis: Colchicine alone, melphalan and prednisone, and melphalan, prednisone, and colchicine. *New England Journal of Medicine*, 336(17), 1202–1207.
- Lauritzen, S. (1996). *Graphical models, ser. Oxford statistical science series*. Oxford University Press.
- Lewis, P. M., II. (1959). Approximating probability distributions to reduce storage requirements. *Information and Control*, 2(3), 214–225.
- Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2), 561–572.
- Malerba, F., & Orsenigo, L. (1997). Technological regimes and sectoral patterns of innovative activities. *Industrial and Corporate Change*, 6(1), 83–118.
- Mogos, S., Davis, A., & Baptista, R. (2021). High and sustainable growth: Persistence, volatility, and survival of high growth firms. *Eurasian Business Review*, 11, 135–161.
- Musso, P., & Schiavo, S. (2008). The impact of financial constraints on firm survival and growth. *Journal of Evolutionary Economics*, 18(2), 135–149.
- Ortiz-Villajos, J. M., & Sotoca, S. (2018). Innovation and business survival: A long-term approach. *Research Policy*, 47(8), 1418–1436.
- Pérez, S. E., Llopis, A. S., & Llopis, J. A. S. (2004). The determinants of survival of Spanish manufacturing firms. *Review of Industrial Organization*, 25(3), 251–273.
- Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4), 841–867.
- Riso, L., & Guerzoni, M. (2022). Concept drift estimation with graphical models. *Information Sciences*, 606, 786–804.
- Riso, L., Zoia, M. G., & Nava, C. R. (2023). Feature selection based on the best-path algorithm in high dimensional graphical models. *Information Sciences*, 649, 119601.
- Santarelli, E., & Vivarelli, M. (2007). Entrepreneurship and the process of firms' entry, survival and growth. *Industrial and Corporate Change*, 16(3), 455–488.
- Savin, I., & Novitskaya, M. (2023). Data-driven definitions of gazelle companies that rule out chance: Application for Russia and Spain. *Eurasian Business Review*, 13, 507–542.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1), 239–241.
- Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician*, 40(2), 106–108.
- Sternberg, R., & Litzengerger, T. (2004). Regional clusters in Germany—Their geography and their relevance for entrepreneurial activities. *European Planning Studies*, 12(6), 767–791.
- Sternberg, R., et al. (2009). Regional dimensions of entrepreneurship. *Foundations and Trends in Entrepreneurship*, 5(4), 211–340.
- Suárez, F. F., & Utterback, J. M. (1995). Dominant designs and the survival of firms. *Strategic Management Journal*, 16(6), 415–430.
- Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6), 363–377.
- Useche, D., & Pommet, S. (2021). Where do we go? VC firm heterogeneity and the exit routes of newly listed high-tech firms. *Small Business Economics*, 57(3), 1339–1359.
- Zhang, D., Zheng, W., & Ning, L. (2018). Does innovation facilitate firm survival? Evidence from Chinese high-tech firms. *Economic Modelling*, 75, 458–468.
- Zhou, H., & van der Zwan, P. (2019). Is there a risk of growing fast? The relationship between organic employment growth and firm exit. *Industrial and Corporate Change*, 28(5), 1297–1320.