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

[1], Luigi Riso[1,*], Zakaria Babutsidze[2,3], and Marco Guerzoni[4]

[1]Università Cattolica del Sacro Cuore
[2]SKEMA Business School, Université Côte d'Azur (GREDEG) and
[3]OFCE, Sciences Po Paris
[4]DEMS, University of Milan-Bicocca

Abstract

We address the issue of variable selection in the context of survival analysis. The availability of very rich firm-level data-sets present two challenges. First, manual variable selection in the context of high-dimensional data might go beyond cognitive abilities of the researcher and could create fertile ground for scientific malpractices (like reverse p-hacking) and for generating ex-post justified opaque models influenced by cognitive biases. Second, the presence of a time dimension in the data, which is becoming longer with digitized bookkeeping, requires explicit treatment of evolving socio-economic context as the stability of the data-generation process is fast becoming implausible. We present a hybrid algorithm which guides the researcher in the process of variables selection over time, but preserves the possibility of including her expertise in the process of model-building. As a test-bed for this methodology we use the data-set of Italian startups funded in 2009. We study their survival over the period of 10 years. Besides demonstrating significant volatility in the set of variables explaining firm exit over the years, by using this novel methodology we are able to challenge conventional view on on the key role played by industrial sector and geographical location in firm survival.

Keywords: Firm survival; Model drift; Variable selection.
JEL codes: C10; D22; L29.

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 straight-forwardly 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' [Eklund et al., 2007, Altman and Krzywinski, 2018] 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.

In this paper we present a data-driven alternative to conventional survival analysis models. 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 and Guerzoni, 2021]. In presence of drift, the variable selection is re-evaluated at every period (length of which needs to be specified by the researcher). The second stage constitutes the fitting of econometric survival model before which the researcher has the possibility to revise the relevant variables list generated by the machine at the first stage.

They key characteristic of the proposed framework is the use of data science algorithms to empower and complement, rather than completely substitute out the researcher. With this aim in mind, Section 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, Suárez and 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 [Pérez et al., 2004, Giot and Schwienbacher, 2007, among others]. One can distinguish six fundamental elements in traditional survival analysis which have emerged over the past decades:

- Age and size [Pérez et al., 2004, Audretsch, 1995, Geroski, 1995];
- Sector or Industry of belonging; [Malerba and Orsenigo, 1997, Klepper, 1996, Geroski, 1995]
- Geographical localization; [Acs et al., 2007, Sternberg and Litzenberger, 2004, Sternberg et al., 2009]
- Profitability; [Delmar et al., 2013]
- Liquidity constraints; [Holtz-Eakin et al., 1994, Musso and Schiavo, 2008]

- Innovativeness and Entrepreneurship. [Guerzoni et al., 2020, Cefis and Marsili, 2005, Santarelli and Vivarelli, 2007]

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 [Hyytinen et al., 2015, Cefis et al., 2021]. 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 6 variables including size, proxy for innovative performance, productivity, capital intensity, and industry dummies. Ortiz-Villajos and Sotoca [2018] use variables on innovative perfomance, size, corporate social responsibility, entrepreneurs psychological traits, but do not 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 BvD data 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] the effect of agglomeration economies on firms survival, and Guerzoni et al. [2020] the survival of innovative start-ups. Useche and Pommet [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. 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 trail-and-error. 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 and 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 allow 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 is 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 $\mathbf{V}$ 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, $G = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = \{v_1, ..., v_p\}$ is the set of vertices and $\mathbf{E}$ is the set of edges. An edge $e = (u, v) \in \mathbf{E}$ indicates that the variables associated to u and v are conditionally dependent [Jordan et al., 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 repre-

sented 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 Likehood Estimator for our 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 and 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. Appendix A.1 exposes the details of the GM methodology in the current paper.

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

As graphical models 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 in different time periods. In this paper we borrow from Riso and Guerzoni [2021] 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 model as a single graphical model spanning the whole study period cannot deliver correct estimates [Riso and Guerzoni, 2021].

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 to their relevance of information for the target variable [Kratzer and Furrer, 2018] (survival, in our case) by considering the conditional dependence as defined by GM. The automatically generated list of variables is not directly employed in the econometric analysis. Instead it is viewed as a suggestion to aid the researcher in making the final choice for model variables. Researcher can add or remove variables, but motivating the choice based on the variable quality, its meaningfulness for the analysis or on the basis of theoretical reasons. Appendix A.3 outlines the steps in the algorithm.

Further technical details and the comparison with other feature selection methods can be found in Riso [2021]. 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 graphical model and theory-driven transparent intervention removing or adding variables and subsequent transformation of the graphical model;
- automatic variables selection on the transformed graphical model;
- econometric analysis;

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

4 Empirical application

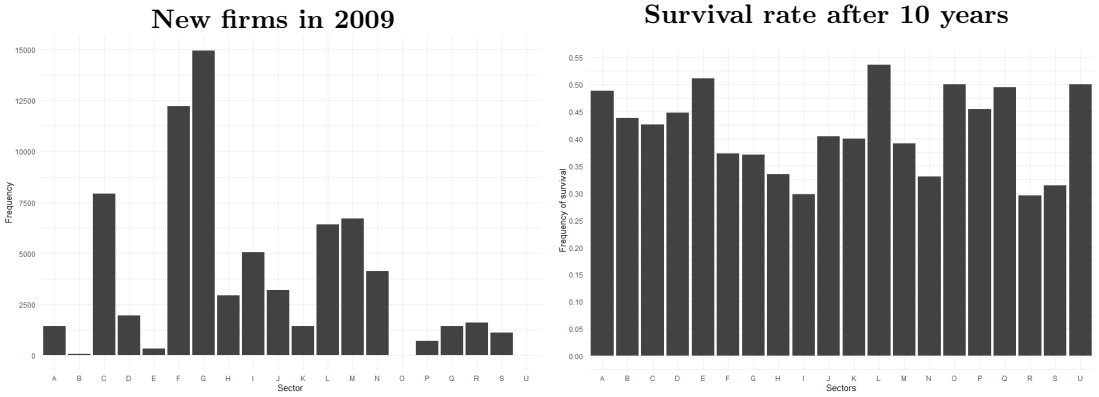
4.1 Data

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 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 2 in Appendix B. Since there is still a percentage of missing variables, we use *Random Forest Missing Algorithm* as the missing data imputation strategy.[1]

[1] 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 and 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].

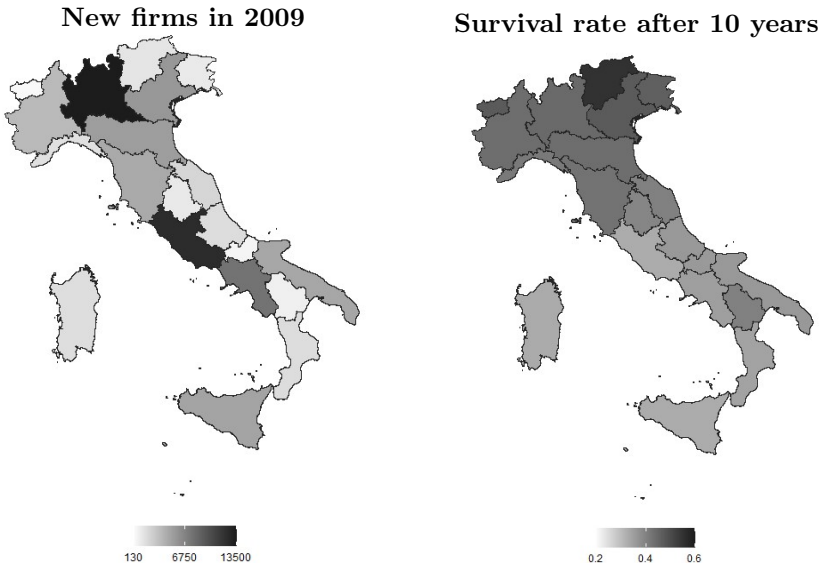
4.2 Traditional Survival Analysis

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 Section 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. The maps in Figure 2 describe geographical distribution and corresponding 10-year survival rates in the population of Italian firms created in 2009. More details are reported in Appendices in the Tables 3 and 4.



(a) Frequency (b) Survival Rate

Figure 1: Histograms of the Sectors in the *start-ups*



(a) Map of the Frequency (b) Map of the Survival Rate

Figure 2: Descriptive look at Italian startups.

We can define $X_i = \{X_{i,1}, ..., X_{i,p}\}$ 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:

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

It is worth nothing that some variables are time variant. Following the standard approach to survival analysis, we consider the time dimension according to:

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

where the covariate $X(t)$ is the value of time-varying covariate for the i_{th} subject at time t, with

$t = 1, ..., T$. The partial likelihood, in general, can be written out as

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

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 3 shows the survival curve for the firms born in 2009, while Figure 4 shows the survival curs with the stratification for *Macro-Region*.

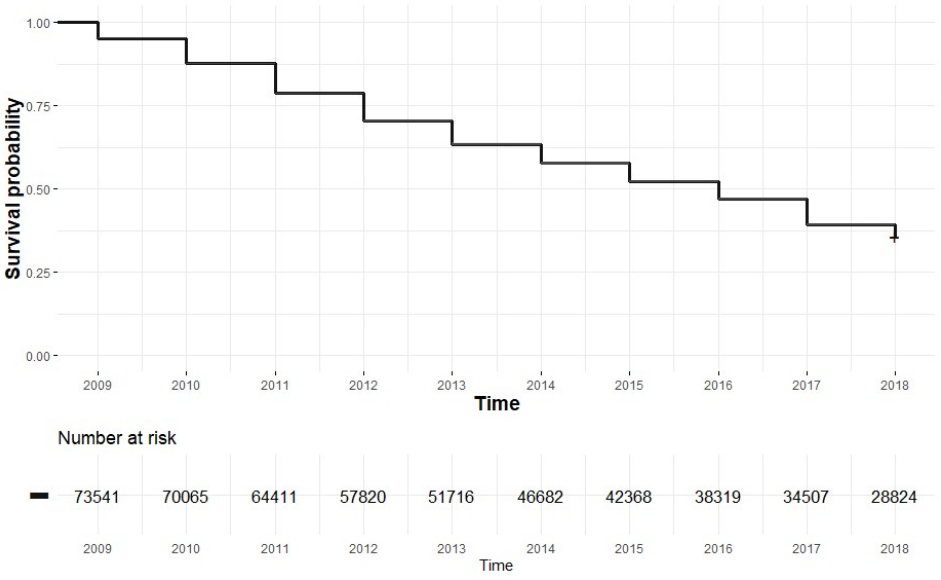


Figure 3: Survival function for the firms born in 2009 and number of firms at risk.

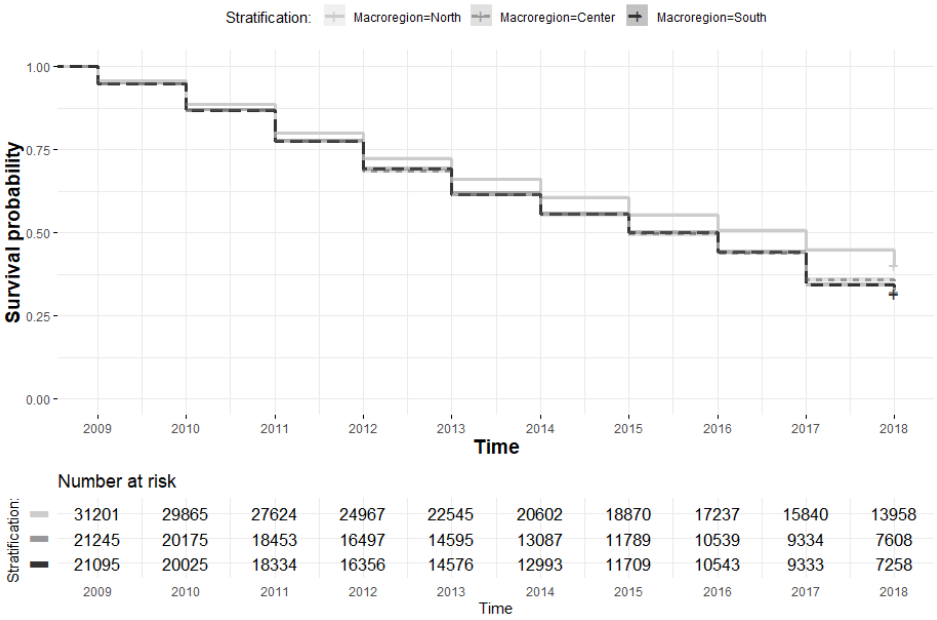


Figure 4: Survival Function Stratification for Macro-Region and number of firms at risk.

Figure 4 highlights the role played by the regions in Italy: a socio-economic divide between the North and the rest of Italy is readily noticeable. Table 1 reports the *log-rank* test on macro-regions that corroborates the our intuition from Figure 4.[2]

Table 1: Log-rank test for Macro-Region

Macro-Region	**N**	**Observed**	**Expected**	$\frac{(O-E)^2}{E}$
North	31201	18634	20876	241
Center	21245	14366	13315	83
South	21095	14409	13218	107
$\chi^2 = 480$, on 2 degrees of freedom, $p\text{-}value \leq 2e-16$				

Table 5 in Appendix C 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.

[2]The *log-rank* is the most widely used method for comparing different survival curves. It is approximately distributed as a χ^2 test statistic and is a non-parametric test, which makes no assumptions about the survival distributions.

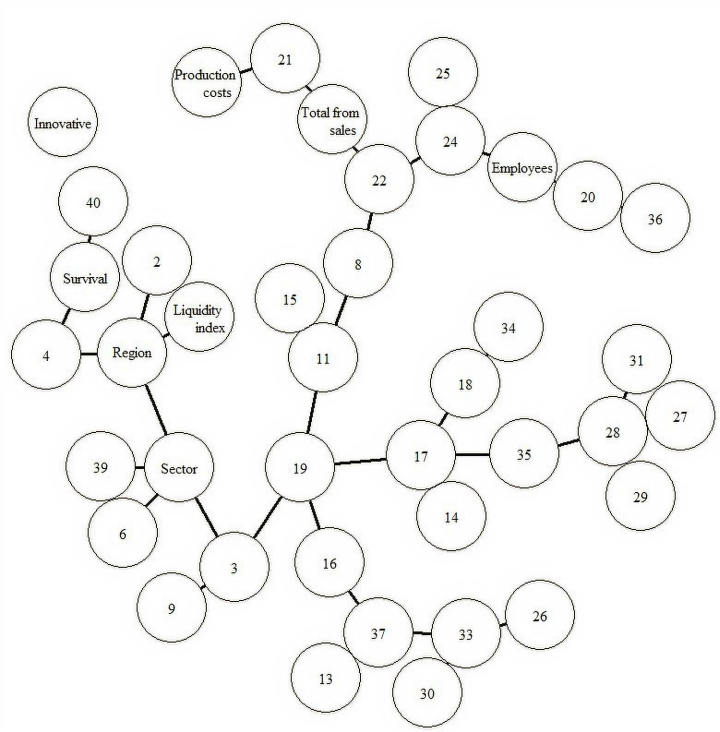
Firms in most of the regions and sectors have a lower chance to survive vis á vis firms in *Lombardia* and in the automotive, respectively.

These results assume both a stable relation among variables over a time span of ten years, as well as stable coefficients. In the next section we apply the theoretical framework presented in Section 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.

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 section 4.2.

Spanning Tree 2009



Spanning Tree 2010

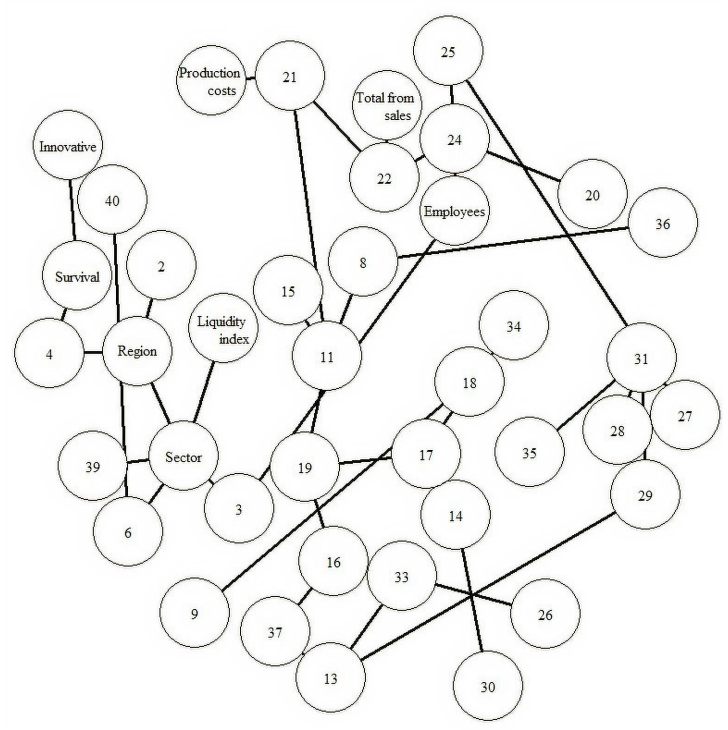


Figure 5: Graphical Models for Italian startup over 2009-2010

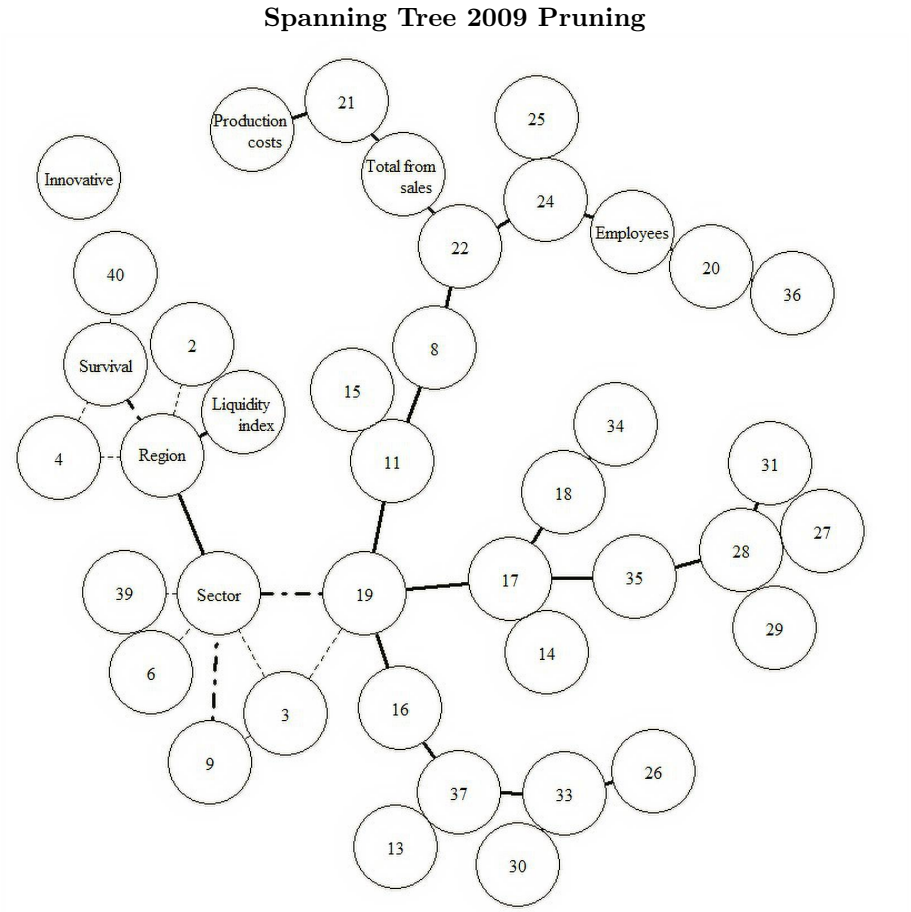


Figure 6: Pruning GM 2009, solid black lines indicate the original connection, while dot-dash lines indicate the connection between the variables after the pruning

4.3.1 A graphical model of Italian Start-ups

Figure 5 zooms on the inferred graphical models for the years 2009 and 2010. Alternatively, Figure 8 presents the model inferred for the whole study period. In Figure 5, for simplicity of exposition, we labeled some variables while labels for the others can be found in table 2. The node object of the study is the dummy variable describing the survival of a firm, node 7 *Survival*. Notice that graphs are not identical overtime. For instance, the variable *Innovative*, that is a dummy variable to identify innovative start-up, is conditionally independent from survival in 2009, but not in 2010. To evaluate how changes in the relationship between the variables drift over time, we compute the *stability* of the inferred graphs and visualize it in Figure 7. This measure contrasts the GM inferred for a given year to the one inferred in year 2009.

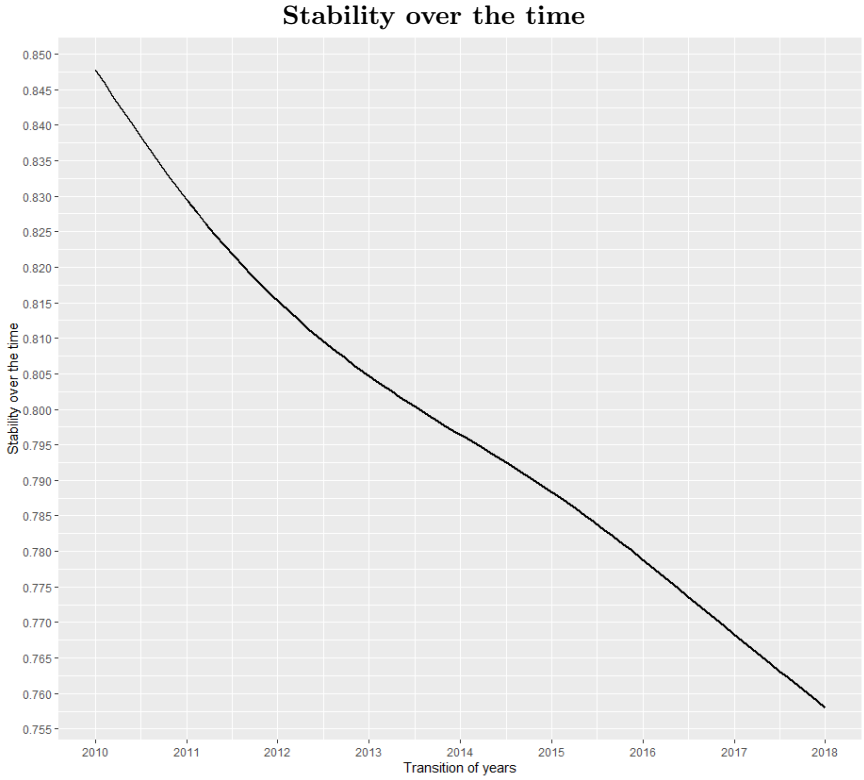


Figure 7: Stability

As we can see from the figure, *Stability* is monotonically decreasing supporting our conjecture that the assumption of stability in high-duration survival analysis is not appropriate. Based on this finding, we carry on with estimating a different GM for every year of observation in the dataset.

4.4 Variable selection

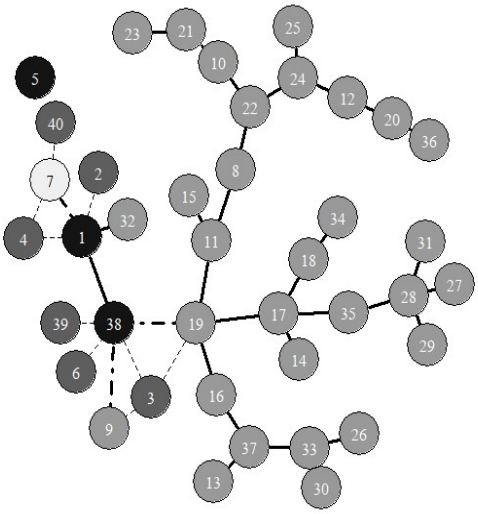
The automated process generating the Figure 6 can help the researcher in identifying variables to include in the econometric analysis. This step, which introduces "the-man-in-the-loop" in the otherwise automatic process, serves to include the researcher's knowledge with scientific judgment in the process. Reasons for exclusion could be based on economic theory, or on statistical features of the considered variable that could introduce undesirable assumption violations in econometric analysis. In the case at hand we decided the 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.

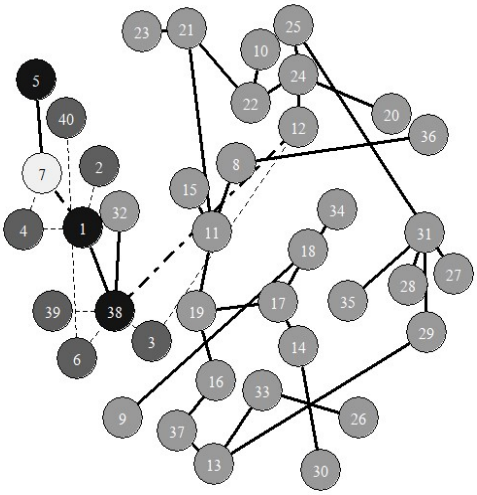
This process in which we remove variables from the model and generate new link to account for the omitted variable is called pruning. For instance the node 4 (legal status) mediates the impact of *Region* on *Survival*. This means that the survival is conditionally independent from the region once we 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. Naturally, pruning is applied to very year-specific graphical model.

Figure 8 presents the graphical models for ten years after the pruning process. Different sets of variables are color-coded. The variable label correspondence is presented in Table 2.

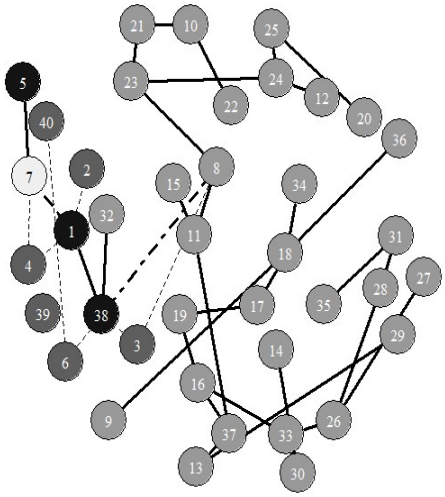
Spanning Tree 2009



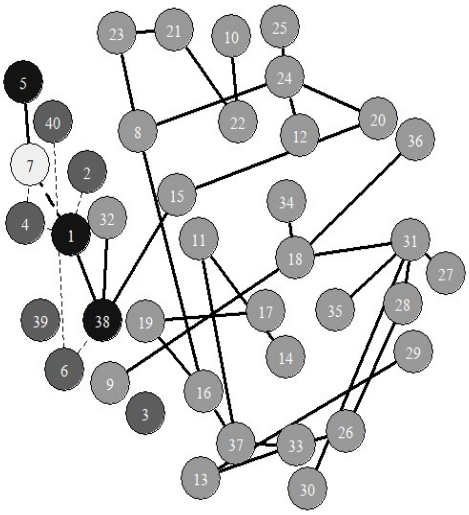
Spanning Tree 2010



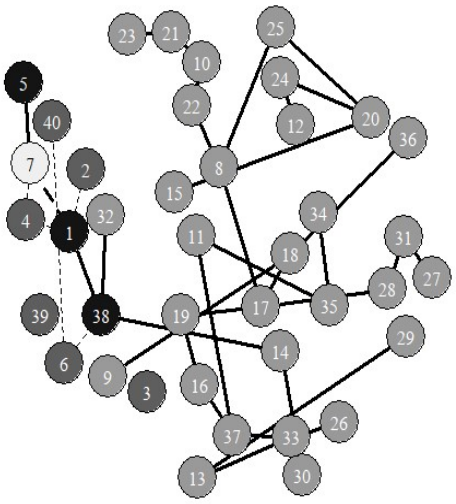
Spanning Tree 2011



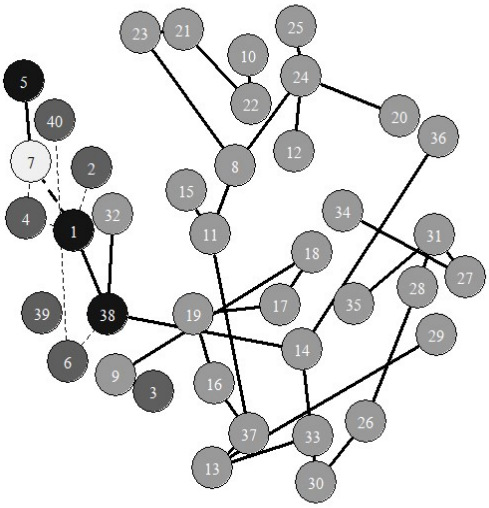
Spanning Tree 2012



Spanning Tree 2013



Spanning Tree 2014



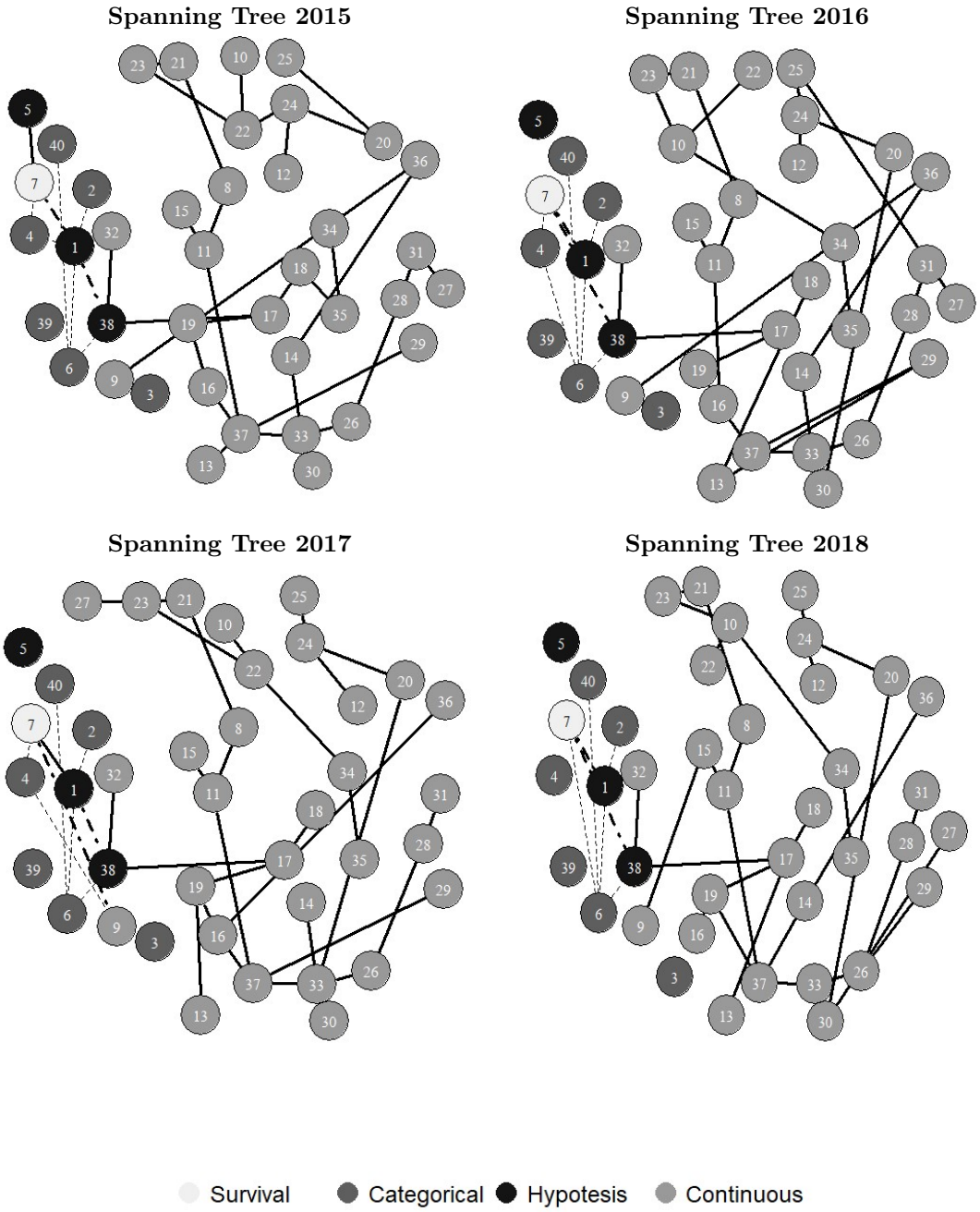


Figure 8: Graphical Models for Italian startup survival over ten years. Solid black lines indicate the original connection, while dot-dash lines indicate the connection between the variables after the pruning.

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 the 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 A.3 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:

$$\log\left(\frac{\rho_{i,t}}{1-\rho_{i,t}}\right) = \beta_0 + \sum_{j}^{N} \beta_{j,t} x_{j,t} \quad \forall \quad t = 2009, ..., 2018, \tag{4}$$

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.

4.4.1 Results

In this section we present the results of the variables selection carried out through the algorithm presented in the Section 3.2. The Figure 9 shows the selected variables and the sign of the coefficients, while figures 10 and 11 focus on the results for all the levels of the variables *Sector* and *Region* respectively. In Figure 9, gray color denotes coefficients with positive coefficients, black signifies coefficients with negative coefficients, while white corresponds to variables not selected for the analysis. In Figures 10 and 11, gray and black shading has the same interpretation as in figure 9, while white denotes variables that are included in the analysis but are estimated to be statistically insignificant. In the three figures, the last column summarizes the result of the Cox regression for comparison.

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. The 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.

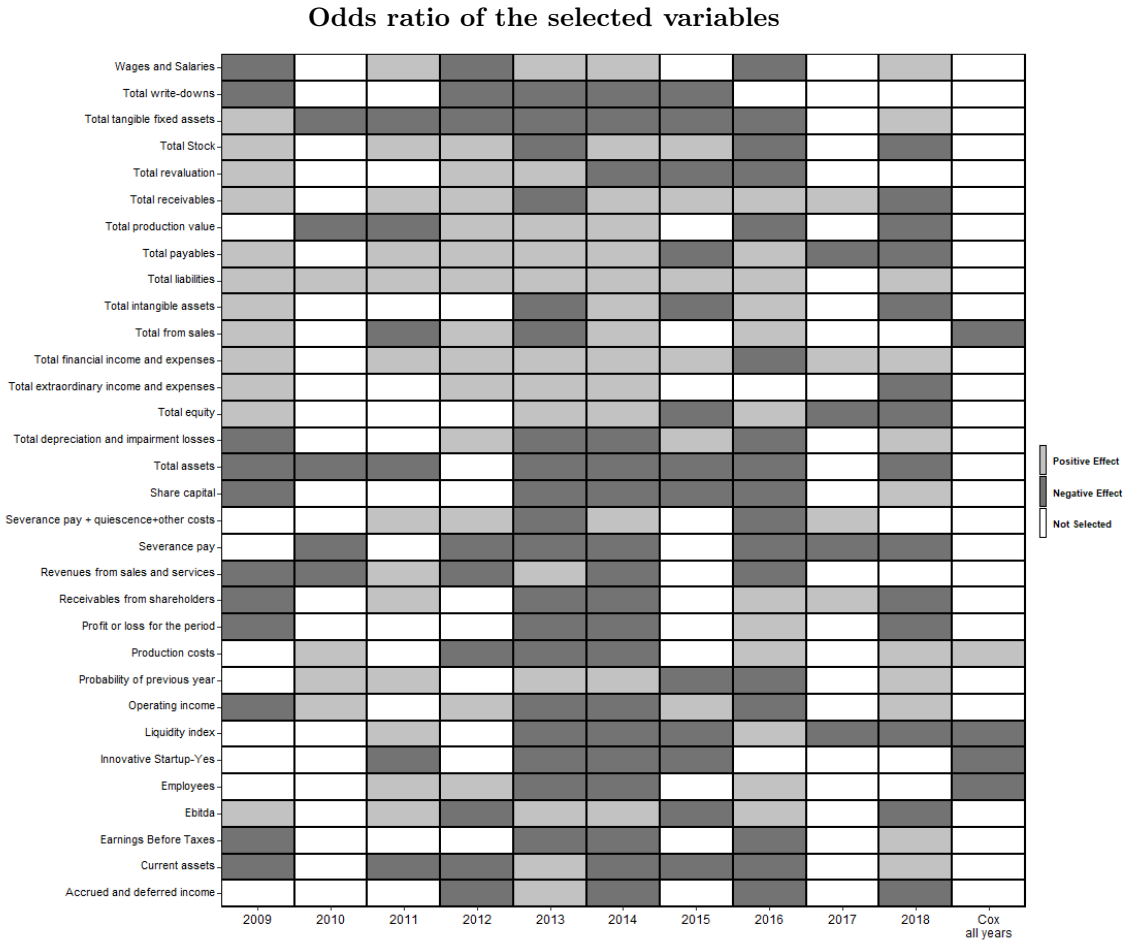


Figure 9: Impact on probability of survival

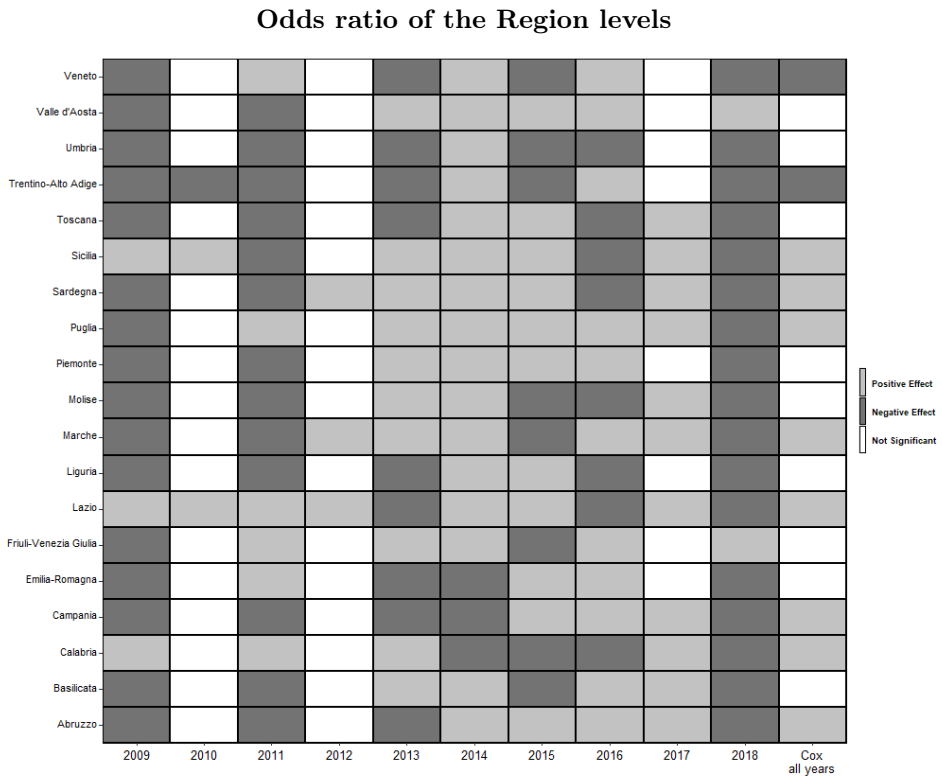


Figure 10: Impact on probability of survival of *Regions*
Reference level: region *Lombardia*

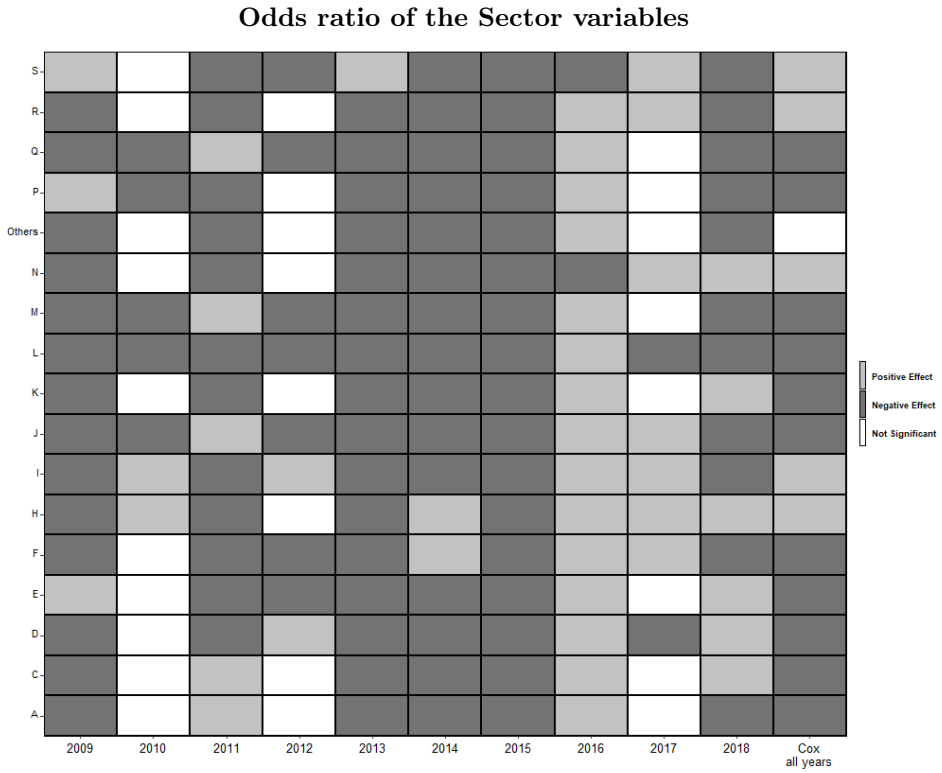


Figure 11: Impact on probability of survival of *Sector*
Reference Level: sector G

5 Summary and conclusion

This paper contributes to the ongoing attempt in the literature of combining data science tools with 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 otherwise automated process. The researcher carries the burden of modifying and approving the final set of model variables. She is responsible for making sure the model is built on sound economic (as well as econometric) theory.

The proposes 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. 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 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.

References

Z. J. Acs, C. Armington, and T. Zhang. 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, 2007.

M. Agostino, D. Scalera, M. Succurro, and F. Trivieri. Research, innovation, and bankruptcy: evidence from european manufacturing firms. *Industrial and Corporate Change*, 2021.

M. Aldinucci, S. Bagnasco, S. Lusso, P. Pasteris, S. Rabellino, and S. Vallero. Occam: a flexible, multi-purpose and extendable hpc cluster. In *Journal of Physics: Conference Series*, number 8, page 082039, 2017.

M. Aldinucci, S. Rabellino, M. Pironti, F. Spiga, P. Viviani, M. Drocco, M. Guerzoni, G. Boella, M. Mellia, P. Margara, et al. Hpc4ai: an ai-on-demand federated platform endeavour. In *Proceedings of the 15th ACM International Conference on Computing Frontiers*, pages 279–286, 2018.

E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

N. Altman and M. Krzywinski. The curse (s) of dimensionality. *Nat Methods*, 15(6):399–400, 2018.

D. B. Audretsch. Innovation, growth and survival. *International journal of industrial organization*, 13(4):441–457, 1995.

F. J. Bargagli-Stoffi, J. Niederreiter, and M. Riccaboni. Supervised learning for the prediction of firm dynamics. In *Data Science for Economics and Finance*, pages 19–41. Springer, Cham, 2021.

C. I. Barnard. Thefunctions of the executive. *Cambridge, MA: Harvard University*, 1938.

R. Basile, R. Pittiglio, and F. Reganati. Do agglomeration externalities affect firm survival? *Regional Studies*, 51(4):548–562, 2017.

J. L. Bellovary, D. E. Giacomino, and M. D. Akers. A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial education*, pages 1–42, 2007.

C. Carota, A. Durio, and M. Guerzoni. An application of graphical models to the innobarometer survey: A map of firms'innovative behaviour. *Italian Journal of Applied Statistics, 25 1*, 25 (1), 2015.

E. Cefis and O. Marsili. A matter of life and death: innovation and firm survival. *Industrial and Corporate change*, 14(6):1167–1192, 2005.

E. Cefis and O. Marsili. Good times, bad times: innovation and survival over the business cycle. *Industrial and Corporate Change*, 28(3):565–587, 2019.

E. Cefis, C. Bettinelli, A. Coad, and O. Marsili. Understanding firm exit: a systematic literature review. *Small Business Economics*, pages 1–24, 2021.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

P. J. Chuard, M. Vrtílek, M. L. Head, and M. D. Jennions. Evidence that nonsignificant results are sometimes preferred: Reverse p-hacking or selective reporting? *PLoS biology*, 17(1): e3000127, 2019.

L. Crosato, J. Domenech, and C. Liberati. Predicting sme's default: Are their websites informative? *Economics Letters*, 204:109888, 2021.

G. C. de Abreu, R. Labouriau, and D. Edwards. High-dimensional graphical model search with graphd r package. *arXiv preprint arXiv:0909.1234*, 2009.

N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains. mrmre: an r package for parallelized mrmr ensemble feature selection. *Bioinformatics*, 29(18): 2365–2368, 2013.

F. Delmar, A. McKelvie, and K. Wennberg. Untangling the relationships among growth, profitability and survival in new firms. *Technovation*, 33(8-9):276–291, 2013.

D. Edwards. *Introduction to graphical modelling*. Springer Science & Business Media, 2012.

D. Edwards, G. C. De Abreu, and R. Labouriau. Selecting high-dimensional mixed graphical models using minimal aic or bic forests. *BMC bioinformatics*, 11(1):18, 2010.

J. Eklund, S. Karlsson, et al. *An embarrassment of riches: Forecasting using large panels*. Citeseer, 2007.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

P. Geroski. Entry, exit and structural adjustment in european industry. In *European industrial restructuring in the 1990s*, pages 139–161. Springer, 1992.

P. A. Geroski. What do we know about entry? *International Journal of Industrial Organization*, 13(4):421–440, 1995.

G. Gigerenzer and R. Selten. *Bounded rationality: The adaptive toolbox*. MIT press, 2002.

P. Giot and A. Schwienbacher. Ipos, trade sales and liquidations: Modelling venture capital exits using survival analysis. *Journal of Banking & Finance*, 31(3):679–702, 2007.

M. Grazzi, C. Piccardo, and C. Vergari. Turmoil over the crisis: innovation capabilities and firm exit. *Small Business Economics*, pages 1–28, 2021.

M. Guerzoni, C. R. Nava, and M. Nuccio. Start-ups survival through a crisis. combining machine learning with econometrics to measure innovation. *Economics of Innovation and New Technology*, pages 1–26, 2020.

M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. The extent and consequences of p-hacking in science. *PLoS Biol*, 13(3):e1002106, 2015.

S. Højsgaard, D. Edwards, and S. Lauritzen. *Graphical models with R*. Springer Science & Business Media, 2012.

D. Holtz-Eakin, D. Joulfaian, and H. S. Rosen. Sticking it out: Entrepreneurial survival and liquidity constraints. *Journal of Political economy*, 102(1):53–75, 1994.

A. Hyytinen, M. Pajarinen, and P. Rouvinen. Does innovativeness reduce startup survival rates? *Journal of business venturing*, 30(4):564–581, 2015.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

M. I. Jordan et al. Graphical models. *Statistical science*, 19(1):140–155, 2004.

H. Jung, J. Hwang, and B.-K. Kim. Does r&d investment increase sme survival during a recession? *Technological Forecasting and Social Change*, 137:190–198, 2018.

S. Klepper. Entry, exit, growth, and innovation over the product life cycle. *The American Economic Review*, pages 562–583, 1996.

D. Koller, N. Friedman, S. Džeroski, C. Sutton, A. McCallum, A. Pfeffer, P. Abbeel, M.-F. Wong, C. Meek, J. Neville, et al. *Introduction to statistical relational learning*. MIT press, 2007.

G. Kratzer and R. Furrer. varrank: an r package for variable ranking based on mutual information with applications to observed systemic datasets. *arXiv preprint arXiv:1804.07134*, 2018.

J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

R. A. Kyle, M. A. Gertz, P. R. Greipp, T. E. Witzig, J. A. Lust, M. Q. Lacy, and T. M. Therneau. 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, 1997.

S. Lauritzen. Graphical models, ser. *Oxford Statistical Science Series. Oxford University Press*, 1996.

S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The annals of Statistics*, pages 31–57, 1989.

P. M. Lewis II. Approximating probability distributions to reduce storage requirements. *Information and control*, 2(3):214–225, 1959.

D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2):561–572, 2016.

F. Malerba and L. Orsenigo. Technological regimes and sectoral patterns of innovative activities. *Industrial and Corporate Change*, 6(1):83–118, 1997.

P. Musso and S. Schiavo. The impact of financial constraints on firm survival and growth. *Journal of Evolutionary Economics*, 18(2):135–149, 2008.

J. M. Ortiz-Villajos and S. Sotoca. Innovation and business survival: A long-term approach. *Research policy*, 47(8):1418–1436, 2018.

S. E. Pérez, A. S. Llopis, and J. A. S. Llopis. The determinants of survival of spanish manufacturing firms. *Review of Industrial Organization*, 25(3):251–273, 2004.

L. Riso. Use of high dimensional modeling for automatic variables selection: the best path algorithm. *arXiv preprint arXiv:2105.03173*, 2021.

L. Riso and M. Guerzoni. Drift estimation with graphical models. *arXiv preprint arXiv:2102.01458*, 2021.

E. Santarelli and M. Vivarelli. Entrepreneurship and the process of firms' entry, survival and growth. *Industrial and corporate change*, 16(3):455–488, 2007.

R. Sternberg and T. Litzenberger. Regional clusters in germany–their geography and their relevance for entrepreneurial activities. *European Planning Studies*, 12(6):767–791, 2004.

R. Sternberg et al. Regional dimensions of entrepreneurship. *Foundations and Trends® in Entrepreneurship*, 5(4):211–340, 2009.

F. F. Suárez and J. M. Utterback. Dominant designs and the survival of firms. *Strategic management journal*, 16(6):415–430, 1995.

F. Tang and H. Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

D. Useche and S. Pommet. Where do we go? vc firm heterogeneity and the exit routes of newly listed high-tech firms. *Small Business Economics*, 57(3):1339–1359, 2021.

D. Zhang, W. Zheng, and L. Ning. Does innovation facilitate firm survival? evidence from chinese high-tech firms. *Economic Modelling*, 75:458–468, 2018.

H. Zhou and P. van der Zwan. Is there a risk of growing fast? the relationship between organic employment growth and firm exit. *Industrial and Corporate Change*, 28(5):1297–1320, 2019.

Appendices

A Methods

A.1 Framework of the Chow-Liu algorithm

Graphical Models are used to specify the conditional independence relationships between random variables of a dataset. Graphically, these relationships are depicted as a networks of variables in a graph. A graph is a mathematical object $\mathbf{G} = (V, E)$, where V is a finite set of nodes in a one-to-one correspondence with the random variables of the dataset, and $E \subset V \times V$, is a subset of ordered couples of V, that defines edges or links representing the interactions between nodes [Jordan et al., 2004]. Two generic node u and v in a graph $\mathbf{G} = (V, E)$, are connected if there is a sequence $u = v_1, .., v_k = v$ of distinct nodes such that $(v_{i-1}, v_i) \in E$, $\forall i = 1, ..., k$. The sequence $u = v_1, .., v_k = v$ is called *path* [de Abreu et al., 2009].

In this analysis we make use of *High-Dimensional Graphical Models*, which prove useful to represent the relationships between a large set of variables. Indeed, we consider a dataset composed by n observations on p random variables $\mathbf{X}_p$ that are collected in a $N \times p$ matrix $\mathbf{X}$. We assume that the set of these p variables is split into two sets: a set of d discrete $\mathbf{Z} = (Z_1, ..., Z_d)$ and a set of q continuous $\mathbf{Y} = (Y_1, ..., Y_q)$ variables. Accordingly, the i-observation of the dataset $\mathbf{X} = (\mathbf{Z}, \mathbf{Y})$ can be expressed as $(\mathbf{z}_i, \mathbf{y}_i)$. Given the one-to-one correspondence between variables and nodes, we can write the sets of nodes V as $V = \{\Delta \cup \Gamma\}$ where Δ and Γ are the nodes corresponding to the variables in $\mathbf{Z}$ and $\mathbf{Y}$, respectively.

In the following, we restrict our attention to discrete random variables, and denote with $z = (z_1, ..., z_d)$ the generic observation (or cell) of $\mathbf{Z}$. In this case the set, $\mathcal{Z}$, of the possible cells of the variables $\mathbf{Z}$, also called *levels*, may be labelled as $1, ..., |Z_v|$. Now, we assume that the cell probabilities factorize according to a an unknown tree τ as follows $\mathbf{G}_Z = (\Delta, E_\Delta)$, where Δ and E_Δ are the vertex and the edges set, respectively. Accordingly, the cell probabilities can be written as follows

$$p(z) = \prod_{e \in E_\Delta} g_e(z) \tag{5}$$

for a given function $g_e(z)$ that depends on the variables included in the edges set e.

Should $e = (Z_u, Z_v)$, then $g_e(z)$ would be a function of z_u and z_v [Edwards et al., 2010] and the cell probabilities would take the form (Chow and Liu [1968])

$$p(\mathbf{z}|\tau) = \frac{\prod_{u,v \in E_\Delta} p(z_u, z_u)}{\prod_{v \in V} p(z_v)^{d_v - 1}} = \prod_{v \in V} p(z_v) \prod_{u,v \in E_\Delta} \frac{p(z_u, z_u)}{p(z_u)p(z_v)} \tag{6}$$

where d_v is the degree of v, that is the number of edges incident to the node v.
In light of (6), the maximized log-likelihood, up to a constant, turns out to be

$$\sum_{(u,v)\in E_\Delta} I_{u,v} \tag{7}$$

The quantity $I_{u,v}$, called *mutual information*, is defined as follows

$$I_{u,v} = \sum_{z_u, z_v} n(z_u, z_v) \ln \frac{n(z_u, z_v)}{n(z_u)n(z_v)}$$

where $n(z_u, z_v)$ is the number of observations with $Z_u = z_u$ and $Z_v = z_v$.
Lewis II [1959] defined the mutual information between two variables as a measure of their closeness.
The maximum likelihood tree for the entire set Z of the d discrete random variables can be obtained by computing $I_{u,v}$ as edges weights on the complete graph with vertex set Δ using a maximum spanning tree algorithm [Chow and Liu, 1968] .
It can be proved that $I_{u,v}$ is one half the usual statistic G of the likelihood ratio test for marginal independence of Z_u from Z_v, that is

$$2Iu, v = G \tag{8}$$

which is calculated by using the table of count $\{n(z_u, z_v)\}$ built by cross-tabulating Z_u and Z_v. Under the marginal independence G has an asymptotic $\chi^2_{(k)}$ distribution, where the degrees of freedom $k = (|X_u| - 1)(|X_v| - 1)$ is the number of parameters tested under the null [Edwards, 2012].
Following a specular approach, we can obtain the maximum likelihood tree $\mathbf{G}_Y = (\Gamma, E_\Gamma)$ for a set $\mathbf{Y}$ of continuous random variables that can be assumed to have a multivariate Gaussian distribution. Here the sample mutual information between two margins Y_u and Y_v is given by

$$I_{u,v} = -N \, \frac{ln(1 - \hat{\rho}^2_{u,v})}{2}$$

where $\hat{\rho}_{u,v}$ is the sample correlation between Y_v and Y_u. As before, the mutual information is related to the likelihood ratio test statistic as in (8). Under marginal independence, G has a $\chi^2_{(1)}$ distribution [Edwards, 2012]
The Chow and Liu [1968] algorithms, which are finalized to find the maximum weight spanning tree of a arbitrary undirected connected graph with positive edge weights, have been studied thoroughly . In this regard, the algorithm due to Kruskal [1956] provides a simple and efficient solution to this problem. Starting with a null graph, it proceeds by adding at each step the edge with the largest weight that does not form a cycle with the ones already chosen. Edwards et al. [2010] proposed an extension of the Chow-Liu Algorithm that can be applied with mixed

dataset $\mathbf{X}$. This algorithm relies on the use of mutual information between a discrete variable Z_u and a continuous variable Y_v and it is characterized by the fact that the marginal model is a simple ANOVA model (section 4.1.7 [Edwards, 2012]). In order to find the mutual information $I(z_u, y_v)$ between each couple of variables in the mixed case it is important to distinguish two possible cases that depend on the variance of Y_v with respect to the levels of the discrete variable Z_u. Indeed the variance of Y_v can be homogeneous or heterogeneous across different levels of the discrete variable Z_u [Edwards, 2012]. For a couple of variables (Z_u, Y_v), we can write the sample cell count, mean, and finally the variance, respectively, $\{n_i, \bar{y}_v, s_i^{(v)}\}_{i=1,\dots,|Z_u|}$. An estimator of mutual information, in the homogenous case is give by:

$$I(z_u, y_v) = \frac{N}{2} \log\left(\frac{s_0}{s}\right),$$

where $s_0 = \sum_{k=1}^{N}(y_v^{(k)} - \hat{y}_v)/N$, $\quad s = \sum_{i=1}^{|Z_u|} n_i s_i/N$, and $k_{z_u,y_v} = |Z_u| - 1$ are the degrees of freedom of the test for marginal independence between the discrete variable Z_u and the continuous variable Y_v. While, in the heterogeneous case an estimator of the mutual information is equal to

$$I(z_u, y_v) = \frac{N}{2} \log(s_0) - \frac{1}{2} \sum_{i=1,\dots,|Z_s|} n_i \log(s_i)$$

where $k_{z_u,y_v} = 2(|Z_u| - 1)$ are degrees of freedom of the test for the marginal independence between the discrete variable Z_u and continuous variable Y_v, with statistic specified as in (8) and with a $\chi^2_{(k_{u,v})}$ distribution.
Edwards et al. [2010] suggested also the use of one of the following measures to avoid the inclusion of links not supported by data

$$I^{AIC} = I(x_i, x_j) - 2k_{x_i,x_j} \tag{9}$$

or

$$I^{BIC} = I(x_i, x_j) - \log(n)k_{x_i,x_j} \tag{10}$$

where k_{x_i,x_j} are the degrees of freedom associated with the pair of variables, that are defined according to the nature of the variables involved.
The above measures can be employed in an algorithm that finds the best-spanning tree. The algorithm stops once the maximum number of edges has been included in the graph. It is worth remembering that High-Dimensional Graphical Models are strong decomposable (section 7.4 of [Højsgaard et al., 2012]). Strongly decomposabilty is a useful property of graphical models [Lauritzen and Wermuth, 1989] as it allows to restrict the analysis to sub-regions or sub-models of the graph which are local strongly decomposable models with minimal AIC/BIC [Lauritzen and Wermuth, 1989, Edwards et al., 2010].
A mixed graphical model is strongly decomposable if and only if it is acyclic and contains no

forbidden paths [Lauritzen and Wermuth, 1989]. A forbidden path is a path between two non-adjacent discrete vertices passing through continuous vertices (more details [Lauritzen, 1996], p 7-12). Now, High-Dimensional Graphical Models are strong decomposable as, on the one hand, trees and forests are acyclic, on the other hand, the [Edwards et al., 2010] algorithm avoids the presence of forbidden paths.

A.2 Graphical models and stability

In the context of time dependent variables, such as the case for survival analysis, the (in)dependence relationship among variables, as encoded in a graph does not remain necessarily stable. Riso and Guerzoni [2021] is devoted to the analysis of the drift, that is the estimation of the stability of connection of a graph over time.

First, given a collection of datasets composed by p variables, t different time of observation and n_t observations that correspond to the number of observations, we compute a random variable *stability*, based on the change of connections among the variable described by a *transition matrix* (TM). Starting from the adjacency matrix (AM), obtained from the graphical models for each year, *Transition matrix process* is a function that maps possible change of state between two variables, that is a change of state between AM_{t-1} to AM_t with $t = 1, 2, ..., T$. Specifically, the generic element $w_{i,j;t}$ of TM_t, with dimension $V \times V$, registers the evolution of the conditional dependency of any couple of nodes V_i and V_j in T period. Specifically, the TM takes the following form:

$$TM_t = \sum_{t=1}^{T} 2^{(T-t)} AM_t \tag{11}$$

Riso and Guerzoni [2021] shows how we use the $w_i, j; t$ to identify the distribution of stable connections between two variables and the describe it as a Bernoulli distribution:

$$Y_i|\theta_i \overset{ind}{\sim} Bern(\theta_i), \quad i = 1, .., n$$

where Y is the long form vector of all possible connection over the T years and takes value of 1 if the connection is stable up to that period and 0, otherwise. On this basis, it is possible to implement a Bayesian logistic regression model that is the odds of Y as a linear function of both the history of past connections encoded in W and T.

Since W has 2^t levels, we regress $2^t - 1$ dummy variable and keep $W = 0$ (stable absence of

connections) as the reference category, it that way the logistic regression is:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 \times T + \sum_{j}^{2^t-1} \beta_j \mathbf{w}_{j,i} \tag{12}$$

We can express the Eq.12 in term of probability of stability of θ_i:

$$\theta_i = \frac{\exp\{\beta_0 + \beta_1 \times T + \sum_j^{2^t-1} \beta_j \mathbf{w}_{j,i}\}}{1 + \exp\{\beta_0 + \beta_1 \times T + \sum_j^{2^t-1} \beta_j \mathbf{w}_{j,i}\}} \tag{13}$$

The estimated intercept of the model β_0 is an empirical measure for the stability of data generative process overtime. β_0 closes to zero means that between two periods the data generative process is stable, while low values indicate the absence of a drift.

A.3 Graphical models and variable selection

We briefly explain here the strategy used to implement an automatic feature selection, in which we use strong decomposable graphs for the development of an algorithm of variable selection. It is important to underline that we applied this algorithm after the pruning of GMs. This algorithm belongs to *mRMRe* approach [Kratzer and Furrer, 2018] and ranks the variables according their relevance penalized by a redundancy measure. In this case, the algorithm at **Step 0** produces a rank of the importance for all variables according to the distance of the node of interest. This means that we organize the variables in k possible subsets, where k corresponds to the maximum distance from the node of interest to other nodes inside to the spanning tree. In order to explain better the concept of distance d, we can consider Figure 12: in this case the node of interest *Survival* is at distance $d = 3$ from the node 19 that corresponds to the variable *Total liabilities*. After computing d for each node from the node of interest, we organize the variables in subset and define them w_i, where $i = 1, ..., k$. In each subsets w_i, there are all variables with distance from the node of interest $d \leq i$. From **Step 1** to **Step 4**, the algorithm maximizes the relevance and at **Step 5** it minimizes the redundancy of information. Specifically, the algorithm operates according the following steps:

- **Step 0**: the algorithm finds the best spanning tree or forest, and after the pruning we call this model $\mathcal{M}_0$;
- **Step 1**: the algorithm identifies all subsets of variables w_i, starting form the variable of interest identified by the researcher;
- **Step 2**: the algorithm divides the dataset in training set(75%) and validation set (25%)

- **Step 3**: the algorithm for $i = 1, ..., k$:

  (a) fits the model with the variables present in the subset w_i

  (b) computes the predicted probability of the outcome of interest in the test set;

  (b) measures the overall performance with the area under the ROC curve (AUC) [James et al., 2013] .

- **Step 4**: Select the best among k models (largest AUC), and call it $\mathcal{M}_w$

- **Step 5**: Fit the model $\mathcal{M}_w$ and select only the significant variables. Call this model $\mathcal{M}_f$

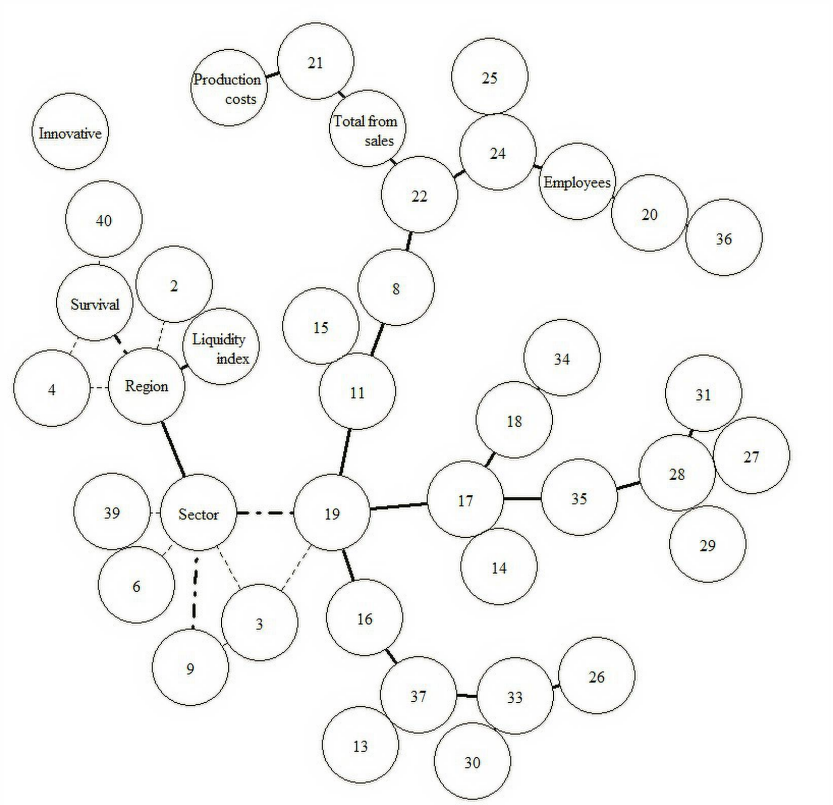


Figure 12: Spanning Tree after the pruning process, solid black lines indicate the original connection, while dot-dash lines indicate the connection between the variables after the pruning

It is simple to note that $\mathcal{M}_f \subseteq \mathcal{M}_w \subset \mathcal{M}_0$. The output of the model $\mathcal{M}_f$ can be written as:

$$y_i = f(\mathbf{w}_i) + \epsilon_i$$

where $\mathbf{w}_i$ represent a vector of the variables selected by the algorithm.

B Data

In this section are reported the details of missing data (Table 2), the frequencies distributions for *Regions* and *Sector* (Table 3 and 4). Table 5 shows coefficients of the Cox Regression.

Table 2: 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%
Province	2	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Legal form	3	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Legal State	4	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Innovative Startup	5	Dichotomous	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Artisan Companies	6	Dichotomous	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Death (Survival)	7	Dichotomous	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total receivables	8	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Receivables from shareholders	9	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total from sales	10	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%
Current assets	11	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Employees	12	Continue	28%	26%	12%	8%	7%	4%	4%	4%	3%	3%
Total tangible fixed assets	13	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total intangible assets	14	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total Stock	15	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total assets	16	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total equity	17	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Share capital	18	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total liabilities	19	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Severance pay	20	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total production value	21	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%
Revenues from sales and services	22	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%
Production costs	23	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Wages and Salaries	24	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%
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%	4%	3%	3%	2%	1%	0%
Total extraordinary income and expenses	27	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%
Earnings Before Taxes	28	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total financial income and expenses	29	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total depreciation and impairment losses	30	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Profit or loss for the period	31	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%
Liquidity index	32	Continue	27%	10%	9%	9%	9%	8%	8%	7%	7%	6%
Ebitda	33	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total revaluation	34	Continue	45%	30%	24%	17%	9%	3%	2%	2%	1%	0%
Total write-downs	35	Continue	20%	5%	4%	4%	4%	3%	3%	2%	1%	0%
Accrued and deferred income	36	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Total payables	37	Continue	17%	4%	4%	4%	4%	3%	3%	2%	1%	0%
Sector	38	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Sae description	39	Discrete	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%
Constitution quarter	40	Discrete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 3: Distribution of the firms for Region

Region	**Number of fimrs born in 2009**
Lombardia	13430
Veneto	5453
Emilia-Romagna	5079
Piemonte	3552
Liguria	1408
Trentino-Alto Adige	1204
Friuli-Venezia Giulia	943
Valle D'Aosta	132
North	**31201**
Lazio	12217
Toscana	4435
Marche	1975
Abruzzo	1631
Umbria	987
Center	**21245**
Campania	7645
Sicilia	4766
Puglia	4580
Calabria	1614
Sardegna	1583
Basilicata	590
Molise	317
South	**21095**

Table 4: Frequency of the sector

Label	Frequency	Description
A	1426	agriculture, forestryand 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	12214	construction
G	14955	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

C Statistical results

Table 5: 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	0.000307	***
Region Basilicata	-8.21E-03	9.92E-01	5.29E-02	5.15E-02	-0.159	0.873346	
Region Calabria	1.26E-01	1.13E+00	3.23E-02	3.30E-02	3.809	0.00014	***
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	0.847482	
Region Friuli-Venezia Giulia	-6.11E-02	9.41E-01	4.39E-02	4.39E-02	-1.393	0.163483	
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	0.122827	
Region Marche	6.59E-02	1.07E+00	3.02E-02	3.07E-02	2.144	0.032037	*
Region Molise	8.57E-02	1.09E+00	6.96E-02	6.76E-02	1.268	0.204727	
Region Piemonte	1.46E-02	1.01E+00	2.42E-02	2.45E-02	0.597	0.550658	
Region Puglia	1.28E-01	1.14E+00	2.13E-02	2.22E-02	5.78	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	8.8	2E-16	***
Region Toscana	-6.24E-03	9.94E-01	2.21E-02	2.23E-02	-0.28	0.779621	
Region Trentino-Alto Adige	-3.55E-01	7.01E-01	4.31E-02	4.28E-02	-8.292	2E-16	***
Region Umbria	5.44E-02	1.06E+00	4.13E-02	4.15E-02	1.312	0.189687	
Region Valle d'Aosta	-9.14E-02	9.13E-01	1.14E-01	1.17E-01	-0.78	0.435531	
Region Veneto	-8.76E-02	9.16E-01	2.10E-02	2.12E-02	-4.13	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	0.292157	
Sector C	-6.59E-02	9.36E-01	1.77E-02	1.93E-02	-3.42	0.000626	***
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	0.000564	***
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	0.031038	*
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	0.006913	**
Sector N	8.03E-02	1.08E+00	2.14E-02	2.68E-02	2.99	0.002788	**
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	0.006322	**
Innovative Startup	-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	0.000474	***
Total from sales	-1.13E-04	1.00E+00	6.52E-06	3.51E-05	-3.208	0.001339	**
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	0.017148	*

Signif. code: $p^{***} < 0.0001$, $p^{**} < 0.001$, $p^{*} < 0.05$, $p^{\cdot} < 0.1$

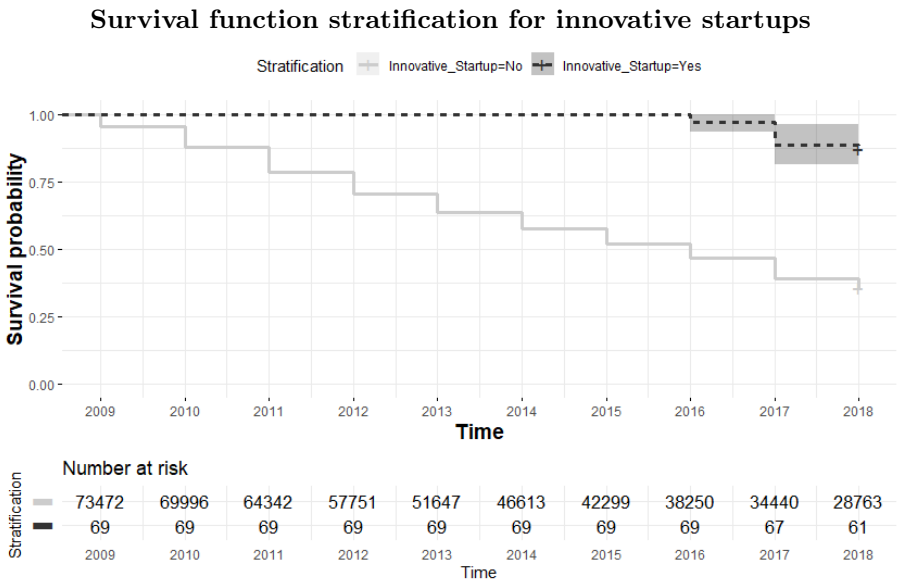


Figure 13: Survival Function Stratification for innovative startup and number of firm at risk