The survival of start-ups in time of crisis. Combining machine learning with econometrics to measure innovation

PREPRINT. Published in Economics of Innovation and new Technology.
https://www.tandfonline.com/doi/full/10.1080/10438599.2020.1769810

Marco Guerzoni[1, 5], Consuelo R. Nava[2, 1, 3], and Massimiliano Nuccio[4, 6, 1]

[1]Despina, Department of Economics and Statistics, University of Turin
[2]Department of Economics and Political Sciences, University of Aosta Valley
[3]Department of Economic Policy, Università Cattolica del Sacro Cuore di Milano
[4]Department of Management, Ca' Foscari University of Venice*
[5]ICRIOS, Bocconi University
[6]City REDI - University of Birmingham

January 7, 2021

Abstract

This paper shows how data science can contribute to improving empirical research in economics by leveraging on large datasets and extracting information otherwise unsuitable for a traditional econometric approach. As a test-bed for our framework, machine learning algorithms allow us to create a new holistic measure of innovation following a 2012 Italian Law aimed at boosting new high-tech firms. We adopt this measure to analyse the impact of innovativeness on a large population of Italian firms which entered the market at the beginning of the 2008 global crisis. The methodological contribution is organised in different steps. First, we train seven supervised learning algorithms to recognise innovative firms on 2013 firmographics data and select a combination of those models with the best prediction power. Second, we apply the latter on the 2008 dataset and predict which firms would have been labelled as innovative according to the definition of the 2012 law. Finally, we adopt

*Corresponding author: massimiliano.nuccio@unive.it

this new indicator as the regressor in a survival model to explain firms' ability to remain in the market after 2008. The results suggest that innovative firms are more likely to survive than the rest of the sample, but the survival premium is likely to depend on location.

JEL CODES: O30, D22, C52, C55.

Keywords: innovation, start-ups, economic crisis, machine learning, survival analysis.

1 Introduction

This paper shows how data science can contribute to empirical research in economics by leveraging on large datasets and extracting information otherwise unsuitable for a traditional econometric approach. Yet, research questions drawn from the economic theory, on the one hand, and assumptions behind the econometric modelling, on the other, guide our choice to exploit the richness of the science of data. As an exemplary case and further contribution to the literature, we apply this framework to evaluate the performances and survival rate of innovative start-ups (hereafter, INNs) vis-à-vis other types of newly funded firms (non-innovative start-ups, hereafter NOINNs) for which empirical evidence shows controversial results. More consensus can be found around the two major challenges which undermine a robust causal relationship between innovation and survival probability. First, most commonly selected proxies and measures for innovation have revealed serious limitations in capturing innovation (OECD, 2018). Second, firm survival may depend on many internal and external factors, and therefore the innovation effect is not easy to isolate and might suffer from confounding issues (Freeman, 1994). Nevertheless, this paper does not want to be just another study of the innovation effect on firm survival. Our contribution is primarily methodological. We adopt an alternative and holistic measure of innovation drawn from the Italian national regulation. Therefore, we analyse the effect of innovation on the survival probability of a large sample of Italian start-ups established in 2008, the very first year of the financial crisis that marked a strong acceleration of the Italian industrial decline. Assuming that the crisis exacerbated both market risks and financial constraints, this database offers an extraordinary opportunity to test the effect of a very strong selection mechanism. If there is any truth in the evolutionary framework, which describes industrial dynamics as triggered by the evolutionary mechanism of entry and selection, we should be able to observe it in a time of crisis.

Our empirical strategy is able to effectively relax some of the constraints imposed by the traditional inferential analysis by integrating a data science approach with econometrics, according to the following three steps.

First, we adopt a definition of "innovative start-up", built on the multiple criteria prescribed by the Italian regulation in 2012 aimed at boosting new high-tech firms through a program of incentives. Therefore, we extract all available new entrant firms in 2013 from AIDA, the Bureau Van Dijk database, including start-ups both registered and not registered as innovative according to the above regulation. After a data cleansing process, we implement a supervised machine learning approach based on the training of seven algorithms (namely recursive partitioning, classification and regression trees, logit regression, bagging, naïve Bayesian classifier, and artificial neural network) to predict the probability of being INNs using 124 firmographics

variables. Since the innovation literature considers sectors and locations as important confounding effects in explaining survival, we exclude them from the training-set of the machine learning algorithms. This allows us to eventually include these variables in an econometric framework, without the risk of describing spurious relationships.

Second, from the same database, we extract the sample of new firms entering the market in 2008 which faced the highly selective environment of the crisis, and we select a combination of the above algorithms able to predict the probability of being INNs.

Third, once we can discriminate between INNs and NOINNs according to the above multi-criteria definition, we estimate, with a Cox proportional hazards model, firms' survival over ten years (2008-2018), now controlling for the impact of sectors and locations. Without the use of machine learning algorithms, this innovative measure of innovation could not have been created and, without a clear theoretical input from the literature and the econometric assumptions to guide the machine learning modelling, this new indicator would have been useless.

The paper proceeds as follows. In Section 2, we present our methodological approach and explain how it can contribute to economic empirical analyses in general, while Section 3 positions our contribution in the debate around the role of innovation in fostering survival in the market, as a specific case to test our methodology. Thereafter, in Section 4, we present the machine learning process which leads to the creation of the new indicator for innovation. In Section 5, we carry out a survival analysis, and, finally, in Section 6 we summarise and discuss the main results of the paper, as well as the new challenges ahead.

2 Data science: an opportunity for the creation of new variables

The data science paradigm is the result of the recombination and convergence of a few complementary technologies which allow the extraction of information and knowledge from a very large dataset: algorithms, computational power, collection and storage of digitised data (Estolatan et al., 2018). Along with Varian (2014), this paradigmatic change has provided economists with an expanded set of analytical tools to explore data and acquire information. In particular, we can recognise at least three types of approaches to data analysis which widely differ among each other in both their goal and test for uncertainty.

Econometrics is the most popular and oldest set of statistical methods aimed at highlighting causal relations between economic variables. The external validity of its results relies on statistical inference, which requires available observations to be a random sample of the population. Well-known techniques have been developed for non-random data or for the generation of truly random data in experimental settings. Nevertheless, estimator properties have been derived on a

limited class of mostly linear models with several statistical limits. Feedback between variables is difficult to handle and even prohibited between dependent and independent variables; the presence of heteroskedasticity and autocorrelation in the error terms of the econometric model need to be carefully addressed; and an excess of multicollinearity between covariates raises serious computational issues. All in all, the capability to highlight statistically robust causal relations heavily constrains the variety of models that can be implemented and this limitation impinges upon the explanatory power and the performance in out-of-sample predictions. Moreover, the complex reality represented by big data rarely fits into the required econometric assumptions, and nor does the data collection always happen in controlled settings. For this reason, econometrics lacks the capability of fully exploiting the information in big data. However, there is a rapidly growing approach in data science which is based on machine learning techniques for prediction and/or classification, also known as supervised machine learning (see, among others, Kotsiantis et al., 2007). Predictive models learn from historical data and make predictions on new data where we do not know the answer. Technically, predictive modelling is the problem of approximating a mapping function (f) given input data (X) to predict an output value (y).In this framework, algorithms are trained on large number of cases and variables (training-set) and learn from a target category to assign new observations. External validity, *i.e.* the variance of the estimates in out-of-sample predictions, is tested on a partition of the available data (test-set), which is hold aside and not employed in the learning process: namely, for the algorithm prediction over an unobserved category. For this reason and contrary to the econometric approach, any machine learning algorithm is not restricted by any assumption and the only objective function is to maximise the prediction power on the test-set. In this way, the explanatory power of the algorithm can be very high, since no limits to its functional form are imposed, but nothing can be said on the true impact of the single variable on the target one. A clear trade-off emerges between the adoption of models aimed at finding causal relations and models aimed at predicting or classifying a phenomenon (Shmueli, 2010). The former are cautious in the data selection and need to be relatively simple in the functional form to approximate data points and to minimise the mean square error of estimators in order to confirm the underlying theory. Extremely simple models tend not to fit data well enough (under-fitting) and their explanatory power remains limited to the few variables involved, which are not necessarily those which explain the total variability of the phenomenon outside the sample. In other words, they might be unable to account for the complex nature of social phenomena like innovation characterised by the interdependence and interaction of a variety of agents and factors (Antonelli, 2009; Fontana, 2010). The latter are meant to gain excellent performance in prediction, but they are blind to

spot any causal relationship and risk capturing the noise of data (over-fitting)[1].

A third approach is still based on machine learning, but in the form of unsupervised algorithms which create a partition of the data without any *a priori* restriction on the number and type of categories to be generated. Clustering algorithms (Macqueen, 1967), self-organising maps (Carlei and Nuccio, 2014) and, more recently, topic modelling for text analysis of the economic literature (Ambrosino et al., 2018) belong to this group. In unsupervised algorithms, the model validation is pursued by an *ex-post* educated interpretation of the result.

How can research in economics take advantage of the latter approaches? Even though some scholars have decreed the (possible) end of theory and the transition towards a pure data-driven type of science (Anderson, 2008; Prensky, 2009), other authors have suggested (Kitchin, 2014; Ambrosino et al., 2018; Nuccio and Guerzoni, 2019; Carota et al., 2014; Gould, 1981) that the large availability of data, which reveals the complexity of the relationships in the observed reality, actually calls for more theory. Data and its analysis can still act as a powerful hypotheses-mining engine (Jordan, 1998; Carota et al., 2014) and provide new theoretical ideas, which nevertheless need to be filtered by a theoretical interpretation effort. Unfortunately, the implementation of machine learning within economics is still not widespread, except for a few contributions which are mainly methodological (Athey, 2018; Varian, 2014). We claim there are both a high complementary and a dense feedback between theory, econometrics and data science. For example, within the traditional framework of economic theory and hypotheses testing, the large availability of data can be exploited to create new dependent and independent variables which fit into a standard regression analysis. Figure 1 shows the methodological conceptualisation behind our empirical exercise.

The test-bed of our approach is rooted in long-standing controversial evidence on the different survival rates between INNs and NOINNs, and on the extent to which this relationship is distorted by a failure to control for possible confounding variables. As possible weaknesses in previous works, we highlight both the type of indicator used to proxy innovation and the lack of consistent controls for sector and location. Following the broad literature on this topic, we further argue that the existence of a survival premium of INNs can be best tested during the 2008 crisis, when market selection mechanisms were more effective. Only as a second step do we turn to data and data science. We thus collect data about new firms in 2013 when a new Italian Law, enacted on 17 December 2012, provides incentives to start-ups to be identified as innovative firms. We employ a supervised machine learning approach to estimate the probability

[1] There is a stream of literature which tries to develop models that overcome this trade-off (Pearl and Mackenzie, 2018, for instance), but they are more concerned with the creation of artificial general intelligence. With regard to the statistical learning on data, the trade-off between the prediction error due to a simple model (bias in the sense that it could suffer from variable omission or violation of the underlying model assumptions), and the variance of estimates in out-of-sample predictions is still binding.

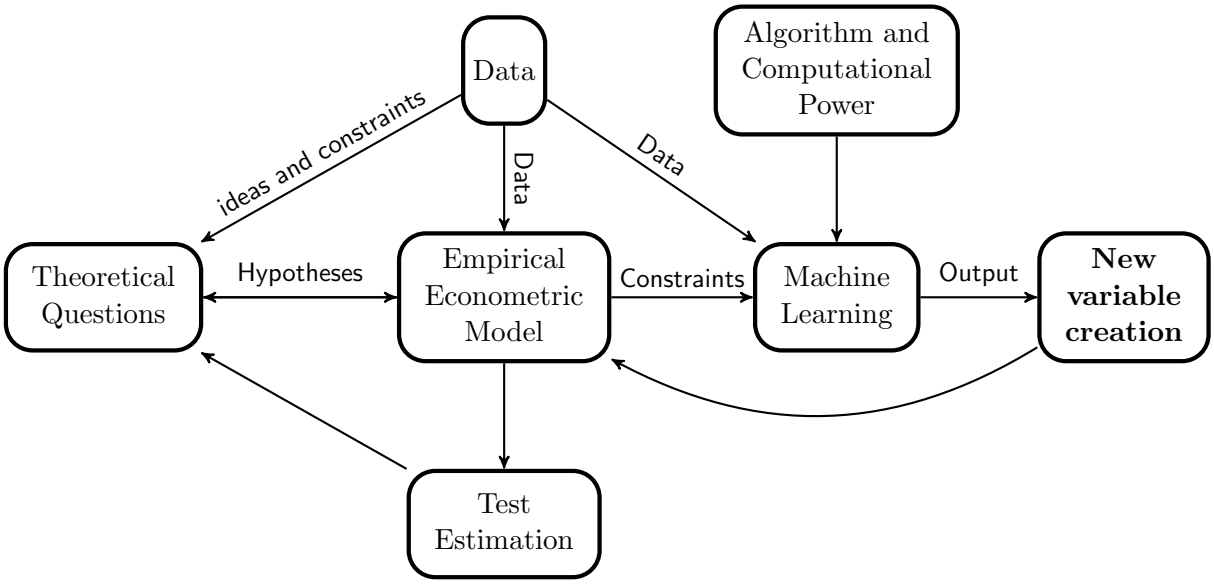


Figure 1: Data science for economics: the creation of a new variable

of firms in 2013 belonging to the given class of "innovative start-ups", and then we apply the same algorithm to predict which firms in 2008 could have been labelled as innovative according to the 2012 law. As explained before, we partition the 2013 start-up sample in the training- and the test-set. On the training-set, we apply a series of algorithms (see Appendix A.2) with different degrees of complexity and with the aim of maximising their prediction power on the test-set. The algorithm, or the combination of algorithms, with the best performance is then used to predict INNs on the 2008 sample. However, to include the new measure of innovation in an econometric model, the predictive algorithm has been trained on all available variables but sector and location, which will be used as covariates. The remainder of the paper discusses the details of this process, which is also briefly summarised in Table 1.

3 Innovation and survival

A key empirical stylised fact in industrial dynamics is the widespread heterogeneity of firms along many dimensions ranging over firms' distribution of size, productivity, their growth, and chance of survival (Griliches and Mairesse, 1995). While this stylised fact clashes with the mainstream economics narrative of the representative agent, it is fully aligned with an evolutionary economics framework in which the mechanism of selection among heterogeneous agents explains the development of industries. In recent decades, scholars in this field have made incredible

Table 1: Summary of the methodology based on econometrics and machine learning tools

Aim	Tool	Data	Process	Output
Measuring innovation	Supervised machine learning	Data on new Italian firms in 2013 by AIDA Bureau van Dijk	In the data, firms are flagged as innovative if they are registered as INNs according to the 2012 Italian Law 221/2012. The flag is used as a label to train a battery of models on a subset of the data. A second subset is used as a test-set to evaluate model performance	Model trained to detect INNs
Controlling for business cycle	Prediction	Data on new Italian firms in 2008 by AIDA Bureau van Dijk	The supervised algorithm, output of the previous line, is used on 2008 data, during the crisis, to detect innovative firms	Classification of 2008 Italian start-ups between innovative and non-innovative
Controlling for sector and location	Multivariate survival analysis	Data on a 10-year panel of new Italian firms in 2008 by AIDA Bureau van Dijk, enriched by the classification created with the model	We run Cox regressions using the classification of the previous output as the main independent variable. Sector, location and interaction terms are added as controls	Estimation of the effect of being an innovative, sector and location on the survival probability

progress in understanding the driving forces which trigger selection and determine which firms prosper and succeed or which, on the contrary, fail. In particular, great attention has been given to the entry of new firms and how selection shapes their chance of survival and subsequent growth. More specifically, scholars focus on the role of the innovativeness of new firms and whether innovation can explain, or at least improve, their fitness in the evolutionary landscape to improve survival chances or not.

3.1 Survival

This literature is rooted in the Schumpeterian idea of creative destruction generated by the entry of new firms (Schumpeter, 1912, 1942), which are more prone to catch both market and technological opportunities than a large incumbent since they are not locked into partly obsolete competencies (Malerba and Orsenigo, 1997; Breschi et al., 2000). For this reason, we mainly observe successful INNs in sectors with a high level of technological opportunity and where the cumulativeness of knowledge is low: that is, previous competencies are not a strategic asset but more likely a burden which hinders the possibility of exploiting new ideas (ibid.). This Schumpeterian view shaped the mythological idea of the entrepreneur who challenges the odds of the fortune and after many attempts eventually succeeds. When new innovative firms succeed,

they grow fast like gazelles (Acs and Mueller, 2008) and, in a few cases, they also become rare unicorns (Simon, 2016).

Beyond this narrative, innovative ventures come along with both advantages and disadvantages. Innovative firms can improve users' and consumers' utility by introducing better products and services (Guerzoni, 2010); they are less myopic and can focus on emerging markets (Bower and Christensen, 1996); they have less cognitive biases generated by previous activities (Aestebro et al., 2007); and they are more dynamic (Teece, 2012). At the same time, a high degree of uncertainty might undermine their innovative efforts and bring them quickly to failure. New markets are often characterised by high uncertainty about consumers' preferences (Guerzoni, 2010) and future development of the technology (Dosi, 1982). Further uncertainty is due either to path-dependency, which might impede the diffusion of novelty (Dosi, 1988) or to fierce competition, since other firms might take the lead in new markets (Fudenberg et al., 1983, among others). Finally, it might be more difficult to find investments (Stucki, 2013).

Lately, scholars have been forming a consensus which suggests an overall survival premium for innovative firms. However, this consensus does not seem to be rooted in strong empirical evidence. Consider for instance the very precise review on this issue by Hyytinen et al. (2015), who survey the most relevant works (to mention a few Arrighetti and Vivarelli, 1999; Audretsch, 1995; Calvo, 2006; Cefis and Marsili, 2005, 2006, 2012; Helmers and Rogers, 2010; Santarelli and Vivarelli, 2007; Wagner and Cockburn, 2010). Hyytinen et al. (2015) classify empirical works according to the sign of the effect of innovativeness on survival probability, specify the sample and proxy used for measuring firms' innovativeness and, eventually, conclude that the large majority of the works account for a positive impact. A rigorous reading of the paper shows, however, that the evidence of positive effects is rather weak. For instance, Cefis and Marsili (2005) find a close to zero effect, while Cefis and Marsili (2006), although reporting a more robust result, do not control for the sector. In a very detailed work, Colombelli et al. (2013) showed that the Kaplan-Meier survival function is virtually the same for innovators and non-innovators, while, on a sample of French start-ups, Colombelli et al. (2016) show that being innovative is not enough to have better survival chances than non-innovative firms, and yet a very small effect on survival emerges for process innovation only. Helmers and Rogers (2010) use patent activity as a proxy for innovation and find a mild positive and significant effect of patenting on survival, but, owing to the large sample, the simple use of the p-value can not really highlight anything conclusive. Indeed, when they repeat the analysis at the industry level and, hence, with less observations for regression, the coefficients are still negligible in size and the p-value is statistically significant only for some sectors. On a sample of U.S. listed companies, Wagner and Cockburn (2010) find just a very small impact of owning patents on firm survival and the coefficients are also mildly

significant and only for a few specifications. There are also works included in the survey which show a negative impact of innovativeness on survival such as Boyer and Blazy (2014). All in all, and to mention just a few, the works surveyed by Hyytinen et al. (2015) do not provide robust evidence that '[...] [T]the prevailing view in the empirical literature appears to be that there is a positive association between the innovativeness of firms and their subsequent survival' (ibid. p.12). Furthermore, there is also a mild disagreement on the effect that different types of innovation might have. For instance, Cefis and Marsili (2019) show that non-technological innovation can be detrimental for survival, while Expósito and Sanchis-Llopis (2019) find that the non-technological innovation can improve business performance.

We highlight three main issues with the present empirical literature which might explain the disparate effect of innovation on the performance of entrants (Audretsch, 1995).

The measurement of innovation The research community of innovation studies has always acknowledged a number of shortcomings in the measurement of innovation, but this is rarely addressed in empirical works and mostly relegated in footnotes. Even the Oslo Manual 2018 (OECD, 2018) spends just a few words on the limits on the measurement of innovation that we would like to recall in the next few paragraphs. The proxies adopted in empirical research for measuring innovation can be roughly divided into two groups: proxies for innovation input and proxies for innovation output. The input of the innovation process are typically R&D investments and high-skilled labour, while the innovation output is usually the number of products or process innovations or patent applications. Figures on R&D expenses and personnel costs usually come from register data; patents are easily identified in the patents office, and there is an extensive literature on their uses; and information about the number and nature of new products or processes can be found in self-reported surveys such as the Community Innovation Survey.

Each of these empirical proxies has proved to have important downsides, which are even more severe for recently established firms. R&D expenses in register data are not always representative of real R&D activity especially in small enterprises for which R&D is not pursued in a formal way or in high-tech start-ups for which, conversely, the R&D activity is spread out across any firm operation. The number of product and process innovations are biased towards the misrepresentation of the respondent's concept of innovation (OECD, 2018). Moreover, surveys do not cover all the population of firms, especially small firms, as start-ups typically are, and, thus, the process of sample selection can induce bias, reduce the possibility of panel data, reduce the degree of freedom of the model, and, thus, raise problems with the inference. As for patents, there is clear evidence on the extreme variance in the propensity to patent both between and within sectors, since, in many cases, especially for process innovation, the appropriability of

the economic returns of intellectual property rights (IPR) can be achieved by means of secrecy (Harabi, 1995). In addition to that, patents are an indicator of the inventing activity, and only rarely do they turn out to be commercially valuable, since the patenting activity is pursued for a vast array of purposes[2].

These measurement issues are even more stringent for start-ups since the balance sheets in the first years are rarely a precise representation of the firm's business, and, as for patents, start-ups might still be in the application process or decide not to patent since in some contexts time-to-market might be much more important than a strong IPR.

Business cycles as confounding effect New firms can prosper or fail for many reasons which do not necessarily relate to economic or technological conditions at the micro-level. For instance, vulnerable firms might survive in a growing economy even if they are not profitable, while selection mechanisms become stricter in downturns. The literature on the economic and financial crisis agrees that recessions usually hinder survival for existing firms. Peric and Vitezic (2016) review the literature on the adverse effect of crisis on existing firms and highlight the main channels such as: production and product lines (Liu, 2009); sales (Cowling et al., 2014); employment (Rafferty et al., 2013); investments (Campello et al., 2011; Bucă and Vermeulen, 2017); performance (Akbar et al., 2013); risk tolerance (Hoffmann et al., 2013; Inklaar and Yang, 2012); and business confidence (Zenghelis, 2012; Geels, 2013; Peric and Vitezic, 2016, p.3). However, entrepreneurial studies have also stressed the positive effects that can be produced by an economic crisis (Bartlett, 2008). This is especially true for those firms that can identify changes in the market and react promptly to exploit new opportunities (Hodorogel, 2009). For this reason, if there are clear differences between innovative and non-innovative firms' survival growth, we should be able to spot them more precisely from this cohort of firms born in 2008, when business constraints became more binding.

Sectors and location as confounding effects Since the work by Pavitt (1984), it has been widely acknowledged that the sector specificity plays a crucial role in explaining the performance of a sector especially in terms of innovation. Along the same lines, the work by Malerba and Orsenigo (1997) developed a theory and provided strong empirical evidence that the technological base underlying the activities of a sector is a key driver of the innovative performance of firms. Sectors characterised by high technological opportunities, low appropriability, and a low cumulativeness of the technological knowledge experience a high entry rate of innovative firms, but also a high rate of exits. On this premise, Malerba et al. (1999) introduced the idea that

[2]The debates on the use of patents dates back at least to Pavitt (1985).

the sector is the proper unit of analysis for the development of models able to replicate the history of industrial dynamics in the computer industry. Sector-based history-friendly models have thereafter been applied to various industries such as the pharmaceutical (Malerba and Orsenigo, 2002), textile (Fontana et al., 2008), and DRAM (Kim and Lee, 2003)[3].

Along the same line, the industry life cycle approach theorised and showed that the early stages of new industries attract most of the entries, but at the same time experience the highest rates of failures (Klepper, 1996; Geroski, 1995). Within an evolutionary perspective this can be framed as the costly process of trial and error at the industry level in which many enter, but most do not survive: survivors thereafter exhibit a more than proportionate growth base on their performance (ibid.). Thus, Pavitt (1984)'s taxonomy, Malerba and Orsenigo (1997)'s classification, and Klepper (1996)'s industry life cycle approach suggest that there exists a bias for innovative firms towards specific sectors, and survival rates might differ between innovative and non innovative firms because of the self-selection of innovative firms in specific sectors with specific patterns of survival.

Similarly, since the distribution of economic activities is very uneven across space, region-specific fixed effects can introduce a further confounding effect when analysing the survival rate. The impact of a region on economic performance is heavily determined by the spatial distribution of economic activity at the industry level. However, Acs et al. (2007) show that, even after controlling for both the industry mix of an area and its degree of specialisation, there is still an effect of location on survival. Indeed Sternberg and Litzenberger (2004); Sternberg et al. (2009) recall and show that entrepreneurship is mainly "a regional event" (Feldman, 2001) for many other reasons which can be broadly defined as agglomeration economies (Leone and Struyk, 1976; Sorenson and Audia, 2000) of the regional system of innovation (Howells, 1999) and include, among other things, local government policies, specific user-producer interactions (Rothwell, 1994), the presence of an entrepreneurial atmosphere (Ciccone and Hall, 1996), the role of cities (Lee et al., 2004), industrial clusters (Rocha, 2004), and the presence of higher tertiary education institutions or research centres (Fetters et al., 2010): knowledge spillovers are the key input in the complex process of innovation especially for new entrants (Audretsch and Feldman, 1996, 2004). Nevertheless, there is no consistent use of the control for industries and regions which the theoretical literature has suggested as being the most important. For instance, none of the work discussed (among others Arrighetti and Vivarelli, 1999; Audretsch, 1995; Calvo, 2006; Cefis and Marsili, 2005, 2006; Colombelli et al., 2016; Helmers and Rogers, 2010; Santarelli and Vivarelli, 2007; Wagner and Cockburn, 2010) controls for the location, and not all of them control for the sector.

[3]For a review and future perspectives on history-friendly models see Capone et al. (2019).

Both theoretical and empirical considerations trigger the necessity for a novel approach to the survival problem. In this paper we aim to provide a solution to the three issues discussed above, and therefore we look for evidence of different survival rates between INNs and NOINNs by: (i) clearly introducing a new broad measure of innovativeness; (ii) focusing on the population of new Italian firms in time of crisis; and (iii) controlling for sectors and location, as suggested by the theory.

The contextual achievement of these three goals calls for the necessity to develop a challenging methodology, which is the main contribution of this paper. More in detail, we provide:

- a new way to detect innovative firms with a scope larger than the simple question about survival;
- new evidence on the survival of innovative start-ups;
- a methodological framework to combine data science and econometrics.

4 Data and methodology

The AIDA (Analisi Informatizzata delle Aziende) database, provided by the Bureau van Dijk, contains comprehensive information on all Italian-owned companies required to file their accounts, including whether they have registered as INNs, or not, according to the Government Decree 179/2012[4].

Each firm in AIDA is described by 427 variables belonging to the following macro categories: (i) identification codes and vital statistics; (ii) activities and commodities sector; (iii) legal and commercial information; (iv) index, share, accounting, and financial data; (v) shareholders, managers, company participation. Only variables in category (iv) are observed for different years. In the construction of our sample, we excluded category (v), since the nature of this data is very specific to each observation and not suitable for prediction analysis, nor for econometrics. Despite its considerable dimension (Table 2), the AIDA database does not cover the entire population of Italian firms as does, for instance, the database of the Italian Board of Trade (IBT), and, amongst others, banks, insurance companies, and public bodies are not included. Still our sample varies from a minimum of 62,934 observations in 2009 (about 21.8% of new firms) to a maximum of 74,508 in 2010 (about 28% of new firms). The dataset is restricted to 276 variables for all firms entering the Italian market from 2008 to 2015 and 262 variables are observed from the starting year of activity until 2015. For new Italian firms established in 2008, we have a balanced panel with ten selected variables up to 2018. Here we focus on two cohorts of firms, which entered the market in 2008 and 2013, respectively. Since not

[4]The Decree was thereafter incorporated in the 221/2012 Law, in force since 17 December 2012.

all information is mandatory for each category of firms, the dataset is characterised by many missing values. Therefore, we conduct a careful missing value analysis which brought us to exclude some variables and observations and obtain two samples of 45,576 (2013) and 39,295 (2008) observations. Appendix A.1 includes details on our cleansing methodology.

4.1 Measuring innovation: innovative start-ups and Law 221/2012

In the previous section, we suggested that we cannot rule out the possibility that the weak evidence in the empirical literature on survival and innovation depends on measurement issues. Usually, the literature assesses the innovativeness of firms looking either at the inputs of the innovation process, such as R&D expenses, or the number of employed researchers, or the outputs of the process, such as the patent pool or the number of innovations. For this reason, we adopt a new definition of "innovative start-ups" which allows several dimensions of their innovative activity to be grasped simultaneously.

Starting from 2013, a new Italian class of firms named "innovative start-ups" has been identified through the Law 221/2012. The policy enforced by this law encourages the creation of companies with specific characteristics, defining the level of incentives, and setting up a dedicated section in the Italian companies register[5]. Start-ups applying for these incentives must meet the following requirements:

- be new or have been operational for less than five years;
- have their headquarters in Italy or in another EU country, but with at least a production site branch in Italy;
- have a yearly turnover lower than 5 million euros;
- do not distribute profits;
- produce, develop and commercialise innovative goods or services of high technological value;
- not be the result of a merger, split-up or selling-off of a company or branch;
- be of an innovative character, which can be identified by at least one of the following criteria:
  - at least 15% of the company's expenses can be attributed to R&D activities (satisfied by 64.97% of the INNs);
  - at least 1/3 of the total workforce are PhD students, the holders of a PhD, or researchers; alternatively, 2/3 of the total workforce must hold a Master's degree (satisfied by 29.68% of the INNs);

[5]See website: http://startup.registroimprese.it/startup/index.html

– the enterprise is the holder, depositary or licensee of a registered patent (industrial property) or the owner of a program for original registered computers.

According to the actual composition of the INNs, only 2.7% satisfied all those three requirements and 11.08% are characterised by two to three requirements. However, from AIDA, we do not know which specific criteria they satisfy in order to be registered as innovative. We only have aggregate data from the IBT, presented in Table 3 for 2013.

The Law 221/2012 provides us with a new tool to identify INNs with some advantages over previous indicators of innovativeness:

- we focus on small firms, which are very likely to be truly new entities and not subsidiaries or foreign green-field entrants;
- all innovative firms are focused on innovative goods or services;
- they need to have at least one of the usual proxies for innovative input and output, but not necessarily a specific one, as in the other measures.

Table 2 shows the number of INNs in the total number of sampled firms and the percentage of firms in the data over the entire population (source: see Footnote 5) of new Italian firms. AIDA covers about a fifth of new Italian firms, since the firms of the self-employed, professional and other minor activities are not required to file their accounts. In 2013, firms registered as innovative start-ups were about 1.5%.

Table 2: INNs and NOINNs in the collected sample according to their initial year of activity

	2008	2009	2010	2011	2012	2013
INNs	0	4	51	320	531	1,010
NOINNs	65,088	62,930	74,457	71,599	65,653	67,306
Total	65,088	62,934	74,508	71,919	66,184	68,316
% Italian Start-ups (IBT)	22.7%	21.8%	28.1%	27.2%	24.0%	24.7%
% INNs (AIDA)	0%	0.01%	0.07%	0.4%	0.8%	1.5%
% INNs in AIDA/ INNs in IBT	0%	21.05%	43.22%	124.51%	112.03%	100.50%

Note: the value can exceed 100% for a different account of firms which ceased to exist.
Source: AIDA and IBT.

Table 3: Number of 2013 INNs satisfying the three possible requirements

	First Req.		Second Req.		Third Req.	
No	360	(35.82%)	709	(70.55%)	787	(78.31%)
Yes	645	(64.18%)	296	(29.45%)	218	(21.70%)

Source: IBT, March 2016.

Table 4 describes the distribution of INNs across the ATECO2007 sector classification and shows that INNs are principally active in service and manufacturing (code J and C, respectively). The

Table 4: One digit ATECO2007 of the 2013 INNs ($n = 1,010$) and NOINNs (67306)

ATECO	NOINNs start-ups		INNs start-ups	
A	1,039	(1.54%)	6	(0.59%)
B	46	(0.06%)	0	(0.00%)
C	7,112	(10.56%)	161	(15.94%)
D	703	(0.10%)	10	(0.10%)
E	326	(0.48%)	4	(0.39%)
F	8,290	(12.32%)	14	(1.39%)
G	15,415	(22.90%)	59	(5.84%)
H	2,640	(3.92%)	3	(0.30%)
I	6,072	(9.02%)	0	(0.00%)
J	3,113	(4.63%)	431	(42.67%)
K	1,309	(1.94%)	1	(0.10%)
L	3,193	(4.74%)	0	(0.00%)
M	4,963	(1.54%)	261	(25.84%)
N	4,260	(7.37%)	37	(3.66%)
O	6	(0.00%)	0	(0.00%)
P	666	(0.99%)	7	(0.69%)
Q	1,307	(1.94%)	6	(0.59%)
R	1,865	(2.77%)	3	(0.30%)
S	1,098	(1.63%)	4	(0.40%)
T	1	(0.00%)	0	(0.00%)
Total	67,306		1,010	

Source: AIDA.

map of INNs in 2013 (see Table 5 and Figure A.4 in Appendix A.4) shows a striking concentration in two regions, Lombardia and Lazio, and their capital cities, Milan and Rome, which attract about one out of three INNs.

Finally, we conclude the presentation of INNs main features by presenting some summary tables on their state of activity after 2 years (the 98% were still active in 2015, see Table 6). It should be noted that in the survival analysis we will consider as a firm's death only the negative exits, such as closing or failures, excluding mergers and takeovers.

4.2 Isolating the innovators' premium from confounding effects

Unfortunately for the purpose of this paper, which aims to classify and study the survival of firms born in 2008, the Law was introduced in 2012 and only since 2013 has it been consistently exploited by new firms. In this subsection, we explain how a machine learning algorithm can be trained and tested on 2013 data to identify INNs in 2008 without specific information on model assumptions, but based on a vast array of other firmographics. We can use any type of variable, with the only restriction provided by the requirements of the theory and the econometric model to be applied. Namely, we excluded from the analysis those variables related to the industry sector and the geographical location. Otherwise, the new indicator would have not been suitable

Table 5: Regional distribution (NUTS2) of the 2013 INNs ($n = 1,010$)

Italian Region	Number of innovative start-up		Italian Region	Number of innovative start-up	
Abruzzo	16	(1.58%)	Molise	0	(0.00%)
Basilicata	4	(0.40%)	Piemonte (Torino - 57)	72	(7.13%)
Calabria	14	(1.39%)	Puglia	51	(5.05%)
Campania	65	(6.44%)	Sardegna	30	(2.97%)
Emilia-Romagna	111	(10.99%)	Sicilia	39	(3.86%)
Friuli-Venezia Giulia	26	(2.57%)	Toscana	53	(5.25%)
Lazio (Roma - 102)	111	(10.99%)	Trentino-Alto Adige	27	(2.67%)
Liguria	15	(1.49%)	Umbria	13	(1.29%)
Lombardia (Milano - 155)	229	(22.67%)	Valle d'Aosta	3	(0.30%)
Marche	47	(4.65%)	Veneto	84	(8.32%)

Source: AIDA.

Table 6: Activity state of the 2013 INNs after 2 years ($n = 1010$)

Status	Number of INNs	
Active	994	(98.42%)
Close down	0	(0.00%)
Failed	1	(0.1%)
Liquidation	15	(1.49%)

Source: AIDA.

as the regressor in a model in which sector and location appear as other covariates. At the same time, we also discard from the training-set the investment in R&D and IPR. This exclusion will serve us later as an evaluation tool for prediction in 2008, for which we do not observe the target variable, as explained in Section 4.2.2. We focus on 2013 data since it is the closest available year to 2008 with the relevant information about the target variable (INNs). Predictive modelling learning from historical data is assumed to be static, but data evolves and must be analysed in near real time. The change over time of the statistical properties of the target variable, which the model is trying to predict, is also known as "concept drift" (Žliobaitė, 2010). Therefore, to prevent deterioration of the prediction accuracy, one effective solution is to minimise the time interval between input data and prediction.

4.2.1 Training, test, and model selection to predict INNs

In this section, we apply different algorithms to classify firms as INNs and, thereafter, we compare their predictive power to select the most performing one. We have deployed seven widely used classifiers, which are described analytically in Appendix A.2:

- Recursive Partitioning (RPART);
- Classification Tree (TREE);
- Conditional Inference Tree (CTREE);

- Bagging (BAG);
- Logit Regression (LOGIT);
- Naïve Bayesian classifier (NB);
- Artificial Neural Network (ANN).

We train these algorithms on a 2013 random subset of 80% of the cleansed sample (36,401 firms including 563 INNs), and we test them on the remaining 20% of the sample (9,175 firms including 150 INNs). The dataset is unbalanced since the target variable (INNs) is underrepresented in the sample. The SMOTE algorithm (Chawla et al., 2002) is a well-known technique to address this problem because it artificially generates new examples of the minority class (here INNs) using the nearest neighbours of these cases. Furthermore, the majority class examples are also under-sampled, leading to a more balanced dataset. Eventually, we synthetically increase the number of INNs cases in the training-set only, while we keep the test-set unchanged to evaluate the performance on factual data.

Each algorithm predicts the probability of a start-up in the test-set to be an INN. The predicted probability, which maps from 0 to 1, collapses to NOINN or INN according to a threshold (or cut-off) chosen by the researcher on the basis of a model performance assessment. Our toolbox for comparing algorithms and selecting thresholds includes the analysis of receiver-operating characteristics (ROC) curves (Figure 2), and the density function of the predicted probability for both INNs and NOINNs. In detail, the ROC curves represent pairs of true positive and false positive rates of a classifier for a continuum of probability thresholds, and they can be used to compare different classifiers. Specifically, the highest performing classifier is the one with the ROC curve closest to the upper-left corner (*i.e.* true positive rate close to 1 and false positive rate close to 0). If two classifiers are characterised by intersected ROC curves, it means that the two classifiers are better under different loss conditions[6]. For each algorithm, we define the optimal cut-off as the one associated with the point which minimises the Euclidean distance between the ROC curve and the (0,1) point (see Appendix A.3 for further details). Once a cut-off is set, confusion matrices (Table 7) summarise the number of correctly classified cases and classification errors for each algorithm.

Also the density function for both NOINNs (or negative) and INNs (or positive) predicted probabilities (Figure 3) generated by the seven algorithms on the 2013 test-set can provide some insights on the model performance. In the ideal scenario, represented in Figure 3a, each distribution for the predicted probability of INNs and NOINNs is skewed, respectively, towards 1

[6]Alternatively, as measure of performance, we can compare the area under the ROC curve (AUC). For further details on the interpretation of ROC curves, see Alpaydin (2014).

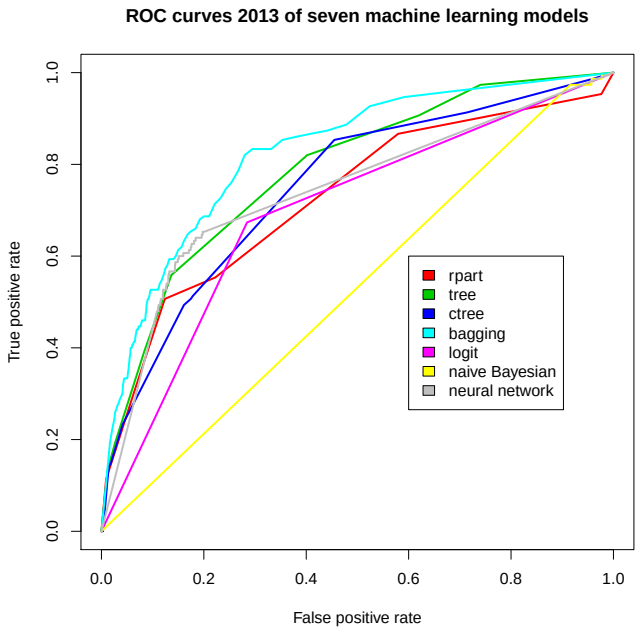
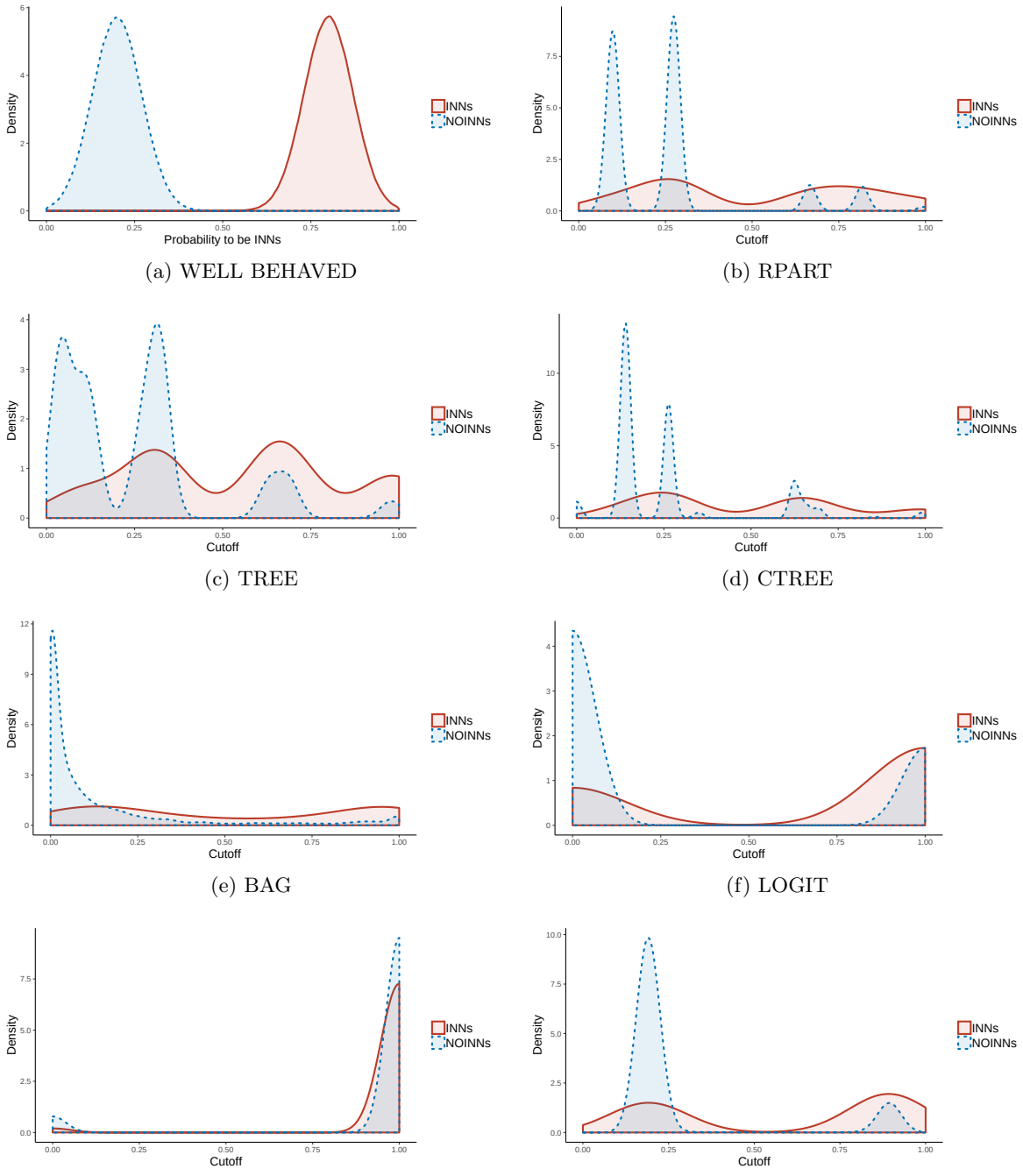


Figure 2: ROC curves on the 2013 start-ups and given the SMOTE technique

Table 7: Confusion matrix of the seven algorithms with SMOTE (optimal cut-offs in parentheses)

Real Data	RPART (0.2817)		TREE (0.3210)		CTREE (0.2632)		BAG (0.1200)		Total
	NOINNs	INNs	NOINNs	INNs	NOINNs	INNs	NOINNs	INNs	
NOINNs	7,908	1,117	7,779	1,246	6,857	2,168	6,612	2,413	9,025
INNs	74	76	66	84	42	108	32	118	150
		50%		56%		72%		79%	
Real Data	LOGIT (1)		NB (1)		ANN (0.1905)				Total
	NOINNs	INNs	NOINNs	INNs	NOINNs	INNs			
NOINNs	9,025	0	9,023	2	7,240	1,785			9,025
INNs	150	0	150	0	53	97			150
		0%		0%		65%			



(a) WELL BEHAVED (b) RPART

(c) TREE (d) CTREE

(e) BAG (f) LOGIT

(g) NB (h) ANN

Figure 3: INNs and NOINNs predicted probabilities for the 2013 start-up according to the seven machine learning algorithms with SMOTE and the well-behaved example, as a benchmark

and 0 and without a common support. Unfortunately, this is not the case and, in most empirical analyses, I and II type misclassification errors can be relevant.

ROC curves offer an interesting comparative insight on the performance of the same model when used for estimating a causal relation instead of predicting a category, as discussed in Section 2. For example, the widely used Logit model is among the worse classifiers in terms of its performers (Figure 3e), confirming the theoretical framework developed in Section 2 and the main limitations of econometrics used to fit such types of data. Among the algorithms tested, BAG (with SMOTE) stands out as having the best predictive power. In fact, when considering the optimal cut-off in the 2013 sample (0.12), this algorithm classifies 6,644 observations as NOINNs and 2,531 (38.1%) as INNs. Unfortunately, the predicted probability distributions associated with BAG do not correctly identify many INNs (see Figure 3e), which is not unexpected since it exhibits a bi-modal distribution with a large variance across its domain. We also consider the second best performing algorithm, ANN, whose distribution of the predicted probability for INNs shows a peak close to 1, although it maintains a second small peak close to 0 (Figure 3h).

In order to further increase the performance, instead of using only one algorithm, we consider a weighted mix of BAG and ANN, and we calculate its predicted probability as a convex linear combination of the predicted probabilities originating from the two algorithms independently[7]. Despite the overall improvement, a large area of common support still remains between INNs and NOINNs densities (Figure 4), and this issue is particularly severe for intermediate values of the predicted probability. Since there is not much difference between the two densities, prediction in that area inevitably leads to a high number of both type I and type II errors. The continuum nature of the empirical problem helps to explain the poor performance of a categorical classifier. Categories like "innovativenes" are often the result of an artificial dichotomisation of an otherwise continuous variable: firms can be more or less innovative on a continuum scale. For this reason, when using the model for prediction, instead of introducing only one cut-off, which separates a predicted INN from a predicted NOINN, we identify two cut-offs which identify three intervals in the (0,1) domain of the predicted probability. Firms with a predicted probability smaller than the first cut-off (0.2) are classified as NOINNs, while firms with a predicted probability higher than the second cut-off (0.8) are classified as INNs. We consider as unclassified those firms with a predicted probability in-between the two cut-offs, and we drop them from the analysis. The algorithm turns out to perform extremely well in correctly classifying INNs (Table 8): most of the misclassification errors are indeed false negatives. This type of error reduces the differences among groups: if we find a difference between innovative and non-innovative firms, the result

[7]We assigned weights 0.77 and 0.23, respectively, to BAG and ANN, according to a function which maximises the separation between the predicted probabilities for INNs and NOINNs and the area under the ROC curve (AUC). As a robustness check, we also tested the mix of different algorithms but there was no substantial improvement in the performance. See Appendix A.3 for further details.

would hold *a fortiori* with a better algorithm.

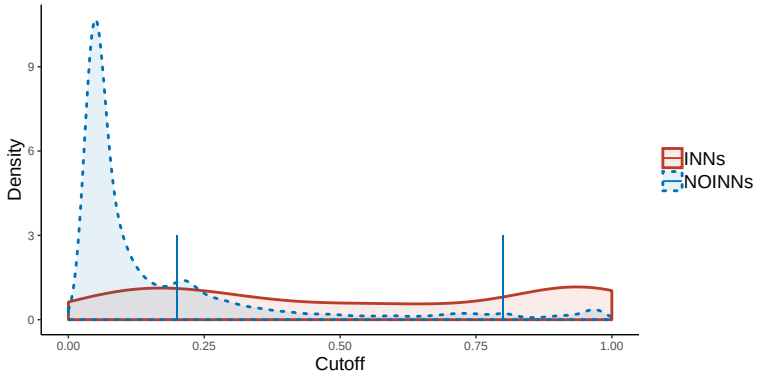


Figure 4: Density of the predicted probabilities for the 2013 start-ups according to the BAG-ANN mixture with SMOTE

Table 8: BAG-ANN mixture performances on the SMOTE 2013 start-up with different predicted probability cut-offs

Real data	Prediction					
	cut-off 0.2		cut-off 0.8		Final subsample	
	NOINNs	INNs	NOINNs	INNs	NOINNs	INNs
NOINNs	6,751	2,274	8,698	327	6,751	327
	75%		96%		95%	
INNs	40	110	99	51	40	51
		73%		34%		56%

4.2.2 Predicting the past: innovative firms in 2008

The mixed model BAG-ANN can now be leveraged to predict which firms would have been innovative in the 2008 sample and generate a new binary variable ($Inno$). According to Table 9 the sample is still very unbalanced with 87.8% INNs and 1.9% NOINNs.

Table 9: BAG-ANN mixture classification of NOINNs (predicted probability ≤ 0.2) and INNs (predicted probability ≥ 0.8) on the 2008 start-ups

	Predicted Probability		Total	%		%
	$\leq$0.2	$\geq$ 0.8		$\leq$0.2	$\geq$ 0.8	$\geq 0.2 \leq 0.8$
2008	34,487	763	39,295	87.8%	1.9%	10.3 %

We are aware that it is impossible to directly evaluate the performance of the 2008 prediction, and we can only reasonably assume that the true and unknown model which generates the data in 2013 is similar to the one in 2008. Nevertheless, we can indirectly provide some statistics on the predicted INNs and NOINNs for a qualitative evaluation of the BAG-ANN performance.

For example, Table 10 shows the percentage of 2008 firms involved in R&D and IPR investment, and the average investment for the period 2008-2018 for INNs and NOINNs, and the values are significantly higher for the former.

Table 10: Qualitative evaluation of the prediction

	INNs	NOINNs
% of firms with positive R&D investment over 10 years	6%	4%
% of firms with positive IPR investment over 10 years	10%	4%
average R&D investment over 10 years, if positive (€)	612K	346K
average IPR investment over 10 years, if positive (€)	7,056K	0,776K

Note: Differences between groups are statistically significant at the 1% level.

Table 11: ATECO classification according to the BAG-ANN mixture classification of NOINNs and INNs on the 2008 sample

ATECO	predicted $\leq$0.2	probability $\geq$ 0.8	Total	ATECO % $\leq$ 0.2	ATECO % $\geq$ 0.8
A	743	12	805	92.23%	1.49%
B	41	1	44	93.18%	2.27%
C	3,222	109	3,903	82.55%	2.79%
D	805	17	880	91.48%	1.93%
E	145	1	170	85.29%	0.58%
F	6,936	136	7,869	88.14%	1.73%
G	5,885	195	7,014	93.90%	2.78%
H	946	30	1,146	82.55%	2.62%
I	1,721	56	2,013	85.49%	2.78%
J	1,537	37	1,754	87.62%	2.11%
K	575	7	627	91.71%	1.12%
L	4,763	48	5,088	93.36%	0.94%
M	3,294	49	3,646	90.35%	1.34%
N	1,754	32	1,982	88.50%	3.85%
O	2	0	5	40.00%	0.00%
P	372	5	401	92.77%	1.25%
Q	594	8	681	87.22%	1.17%
R	689	13	762	90.42%	1.71%
S	423	6	467	90.58%	1.28%
NA	37	1	38	97.37%	2.63%
Total	34,487	763	39,295		

5 Econometric analysis

In this section we test the hypothesis of a survival premium for 2008 firms classified as INNs with respect to NOINNs.

Univariate analysis We first employ the Kaplan-Meier estimator (KME) to show short- and long-term differences in the probability of survival of the 2008 firms after the crisis.

The KME is a non-parametric estimator classically used to, among the other things, estimate survival distribution functions (see, e.g. Fleming and Harrington, 1991; Andersen et al., 2012). In general, this analysis studies the time to death for a population with survival distribution function $S(t)$: namely, the probability that a start-up will be still alive at time t. Let us consider a sample from the population with dimension n (note that here we are dealing with a right-censoring problem). Denote with $t_1 < t_2 < \cdots$ the years when start-ups definitely close their activities on the Italian market. Let d_i be the number of start-ups who close at t_i. The Kaplan–Meier estimator $\hat{S}(t)$ for $S(t)$ is:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{r_i}\right), \tag{1}$$

where r_i is the number of start-ups in the risk set just before time t_i, *i.e.* those firms that had survived, and d_i is the number of failures at time t_i[8].

The KME allows for direct comparisons across the survival probability of samples with different sizes. Figure 5 shows the two Kaplan-Meier curves with time in years on the horizontal axis and probability of surviving, or proportion of firms surviving, on the vertical one. Lines represent the survival curves stratified by INNs and NOINNs within their shadows of confidence intervals. At time zero, the survival probability is 1.0 (*i.e.* 100% of the firms are alive). After ten years, the survival probability is approximately 0.687 (or 68.7% - standard deviation 0.002497) for NOINNs and 0.790 (or 79.0% - standard deviation 0.01474) for INNs. INNs enjoy a survival premium and their survival curve always lies above NOINNs. The associated confidence intervals are wider for INNs than for NOINNs, suggesting the higher uncertainty around innovative ventures. Nevertheless, there is always a statistically significant difference in the two groups, as the rejection of the null-hypothesis of the log-rank suggests.

Multivariate analysis We now seek to address the last issue raised in the theory, that is whether the survival premium of innovative firms persists even when controlling for sectors and locations. We perform this task by adding the one-digit ATECO2007 classification for economic activities, the NUTS3 region classification (namely, "provincia") for the location effect, and the interaction variables of being INNs with both sector and region as controls in a Cox proportional hazards model. The interaction effect can be interpreted as the positive or negative survival premium linked with the specific sector and region. With the Cox model in Eq. (2), we simultaneously estimate the impact of several variables on survival. More precisely, we estimate how the effect of being an INN in a specific sector and in a given location influences the exit

[8]We estimate the variance with Greenwood's formula using the Delta method, and we use log-minus-log transformation for the confidence interval (Borgan and Liestøl, 1990).

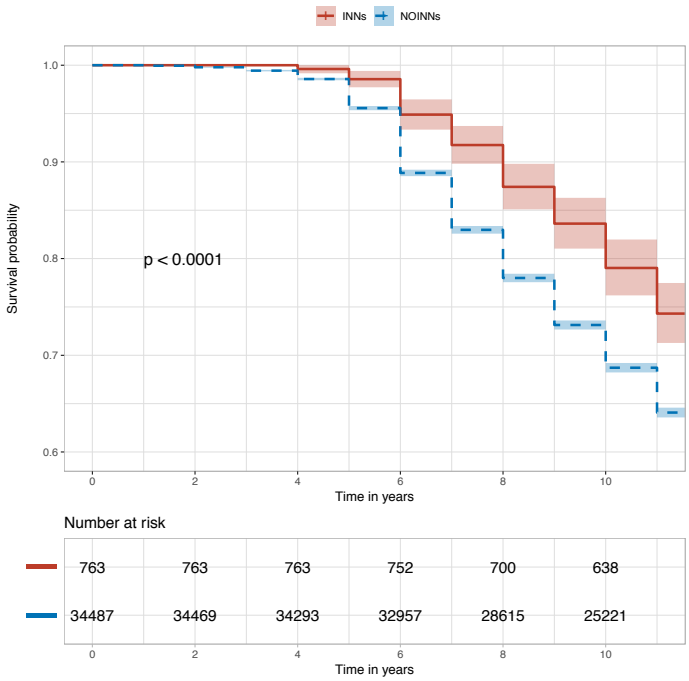


Figure 5: Survival curves of INNs and NOINNs.
Note: the p-value for the log-rank test is reported. Log-minus-log transformation is applied for confidence intervals

rate from the market in a particular year, given that a firm survived up to that year. *Id est*, the hazard rate of failure at time t is;

$$h(t) = h_0(t) \exp\{\beta_1 * Inno + \beta_2 * Sector + \beta_3 * Location + \beta_4 * Interactions\} \quad (2)$$

where:

- t is the survival time;
- $h(t)$ is the hazard function;
- β_i are the coefficients. Since the Cox model can be written as a linear regression model of the logarithm of the hazard, it is possible to interpret the $\exp(\beta_i)$ as the hazard ratio of the i^{th} covariate;
- h_0 is the baseline hazard when all the covariates are set equal to zero. It is possible to estimate the β_i without any consideration of the hazard function only under the assumption of proportional hazard, validated both visually and with the log-rank test (see Table 13;
- *Inno*, *Sector* and *Location* are categorical variables summarised in Table 12, while *Interactions* are the interaction terms between *Inno* and the remaining variables.

Table 12: Description of the variables

Name	**Description**	**Categories**	**Reference**	**Observation**
Inno	Dummy variable for being an INN or a NOINN	2	NOINNs	35,250
Sector	ATECO classification of sectors	20	Manufacturing	35,212
Location	Italian Province (NUTS3 region) in which firm is located	110	Milan	35,250

Table 13 summarises the results for five different models and estimated coefficients. Model (1) uses just the dummy variable for INNs. The coefficient value -0.428 shows that being innovative has a negative and statistically significant effect on the probability of failure with respect to NOINNs. A straight interpretation of the effect is to compute the hazard ratio $= e^{-0.428} = 0.65$, *i.e.* at any given time, innovative firms almost double their chance of survival vis-à-vis NOINNs. Models (2) and (3) add industrial sector and regional controls, respectively, while models (4) and (5)also consider their interaction effects with INNs. When adding interaction effects for the location, the significance of being innovative fades. This evidence suggests that, as pointed out by the theoretical consideration (Feldman, 2001), a large part of the survival premium experienced by INNs depends on the self-selection by innovative firms for locations in which any firm, and not only innovative ones, is more likely to survive. But we would like to make few considerations. This result does not imply that being innovative is irrelevant. For instance, being innovative in a specific region can still lead to a survival premium. By looking at the

interaction of location with the innovative dummy, we can rank Italian provinces according to the survival premium for being innovative. Figure 6 shows the hazard ratio of the interaction effects when they are statistically significant. The higher the value, the higher is the positive effect of innovation on the chance of survival. Second, it might seem counter-intuitive that sector controls do not absorb the explanation power of INNs, whereas location does. But, in fact, NUTS3 regions can capture a much larger effect, which includes, on the one hand, the mix of sectors which characterise a geographical area and, on the other hand, the dynamics discussed above, such as entrepreneurial atmosphere, agglomeration economies, university roles, and so on. However, also for sectors, we can compute the magnitude of the interaction effect as plotted in Figure 6. Nevertheless, an inquiry on the causes that make INNs more likely to survive in some locations or sectors is outside the scope of this work, but there is certainly room for new research questions. Note that, at least theoretically, a further model based on the joint estimates of both sectors and locations is possible. Unfortunately, here, especially for the 2008 firms classified as INNs, we suffer from the complete separation problem, which does not allow the estimation of some interaction effects (Albert and Anderson, 1984).

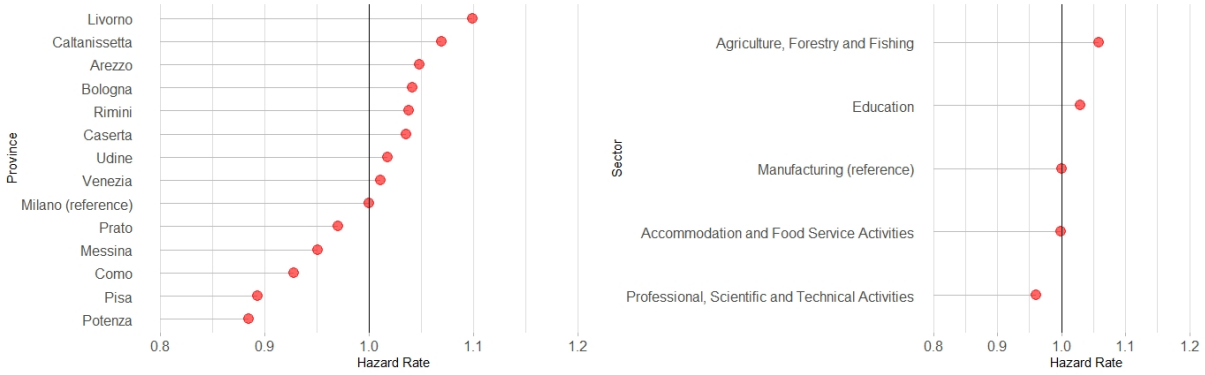
Table 13: Cox regressions: summary

	Dependent variable:				
	Failure				
	(1)	(2)	(3)	(4)	(5)
Inno	-0.428^{***}	-0.459^{***}	-0.438^{***}	-0.512^{***}	-0.122
	(0.072)	(0.072)	(0.072)	(0.198)	(0.246)
Sector		YES		YES	
Location			YES		YES
*Inno * Sector*				YES	
*Inno * Location*					YES
Observations	35,250	35,212	35,250	35,212	35,250
Log Likelihood	$-129{,}172.400$	$-128{,}598.200$	$-128{,}991.200$	$-128{,}591.100$	$-128{,}940.600$
Wald Test	35.340^{***}	483.320^{***}	381.580^{***}	490.800^{***}	386.730^{***}
LR Test	40.756^{***}	493.846^{***}	403.239^{***}	508.001^{***}	504.494^{***}
Score (Logrank) Test	35.885^{***}	493.376^{***}	388.083^{***}	504.257^{***}	443.604^{***}
Df	1	19	110	36	209

Note: *p<0.1; **p<0.05; ***p<0.01

6 Conclusion

This paper has contributed to designing a new research framework which combines data science within econometric models. In particular, we use machine learning algorithms to extrapolate information from a large source of data, which could have not been otherwise employed in



(a) NUTS3 regions

(b) ATECO sector

Figure 6: Hazard Ratio for the interaction term

standard regression models. We stress, and demonstrate, how this exercise needs feedback between theory, econometrics and data science, in order to design the desirable properties of the variable created by machine learning algorithms.

Applying this methodological approach to the longstanding debate around the survival premium of innovative start-ups, we employed machine learning to develop a new indicator of innovation at firm level which removes some drawbacks of previous proxies. During this process, we carefully considered specific constraints to make the new variable suitable for the use in an econometric model. The machine learning algorithm is trained on all possible information (except variables on location and sector, as well as on R&D and IPR). In this way, it has been possible to run a survival model which, as suggested by the theory, includes sector and location as *controls* and does not suffer from any form of endogeneity.

This framework can be considered as a weak integration of econometrics and data science since the two approaches are connected via feedback, but they still run separately. Also, it is possible to imagine different scenarios either with a stronger methodological integration or in which data science completely supersedes econometrics. However, since the state of the art of economics is focused on causal relationships, we believe that such research frameworks are yet to be designed.

As a second contribution, we introduce a new indicator for innovation which is a suitable candidate to be used in various analyses since it overcomes many of the main drawbacks of other innovation proxies: it blends together different aspects of both the inputs and the outputs of the innovation process. However, its nature is very much connected with the Italian case. In this paper, we use the model to predict the innovativeness of Italian start-ups in the past, but the same exercise can be done to predict the innovativeness of foreign firms in the present. Indeed, the AIDA Bureau van Dijk database used to train and test the algorithm is consistent with

the ORBIS-AMADEUS database which collects the same information as AIDA on European firms. Nevertheless, the application of a prediction algorithm on a sample in a different country requires considerable efforts in evaluating the results in relation to other measures of innovation, such as patent applications or survey data. However, for the largest European economies, a match between AMADEUS-ORBIS with PATSTAT and CIS data is already in place. Thus, a major line for future work opened up by this paper is the extension of this new measurement to other countries.

As a third contribution, we provide new empirical evidence on the survival of INNs on the basis of the new indicator. When controlling for sector specificity, INNs seem to maintain their survival premium, which, conversely, fades out when controlling for the location at the NUTS3 regional level. This result challenges the previous literature which formed a weak consensus on the positive effect of innovativeness on survival. We find that INNs have a survival premium only in relation to specific locations. Probably, the specific attributes of a location, which also include the composition of the local economy in term of sectors, might be more or less suitable for a newly established innovative firm. For many locations in the dataset, the effect is not statistically significant, while for others it could not even be estimated due to the small number of start-ups in those areas; however, for some regions a clear-cut effect exists. Understanding the determinants of survival at the regional level could be a question to be addressed in further work.

References

(1989). *Foundations of Cognitive Science*. MIT Press, Cambridge.

Acs, Z. J., Armington, C., and 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.

Acs, Z. J. and Mueller, P. (2008). Employment effects of business dynamics: Mice, gazelles and elephants. *Small Business Economics*, 30(1):85–100.

Aestebro, T., Jeffrey, S. A., and Adomdza, G. K. (2007). Inventor perseverance after being told to quit: The role of cognitive biases. *Journal of Behavioral Decision Making*, 20(3):253–272.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Akbar, S., ur Rehman, S., and Ormrod, P. (2013). The impact of recent financial shocks on the financing and investment policies of uk private firms. *International Review of Financial Analysis*, 26:59–70.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Alpaydin, E. (2014). *Introduction to Machine Learning*. The MIT Press Cambridge, Massachusetts London, England.

Ambrosino, A., Cedrini, M., Davis, J. B., Fiori, S., Guerzoni, M., and Nuccio, M. (2018). What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology*, 25(4):329–348.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07.

Antonelli, C. (2009). The economics of innovation: from the classical legacies to the economics of complexity. *Economics of Innovation and New Technology*, 18(7):611–646.

Arrighetti, A. and Vivarelli, M. (1999). The role of innovation in the postentry performance of new small firms: Evidence from italy. *Southern Economic Journal*, pages 927–939.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Audretsch, D. B. (1995). Innovation, growth and survival. *International Journal of Industrial Organization*, 13(4):441–457.

Audretsch, D. B. and Feldman, M. P. (1996). R&d spillovers and the geography of innovation and production. *The American Economic Review*, 86(3):630–640.

Audretsch, D. B. and Feldman, M. P. (2004). Knowledge spillovers and the geography of innovation. In *Handbook of regional and urban economics*, volume 4, pages 2713–2739. Elsevier.

Bartlett, D. (2008). Fallout of the global financial crisis. In *World Economic Forum*.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Borgan, Ø. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics*, pages 35–41.

Bower, J. L. and Christensen, C. M. (1996). Disruptive technologies: Catching the wave. *The Journal of Product Innovation Management*, 1(13):75–76.

Boyer, T. and Blazy, R. (2014). Born to be alive? the survival of innovative and non-innovative french micro-start-ups. *Small Business Economics*, 42(4):669–683.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth.

Breschi, S., Malerba, F., and Orsenigo, L. (2000). Technological regimes and schumpeterian patterns of innovation. *The Economic Journal*, 110(463):388–410.

Bucă, A. and Vermeulen, P. (2017). Corporate investment and bank-dependent borrowers during the recent financial crisis. *Journal of Banking & Finance*, 78:164–180.

Calvo, J. L. (2006). Testing gibrat's law for small, young and innovating firms. *Small Business Economics*, 26(2):117–123.

Campello, M., Giambona, E., Graham, J. R., and Harvey, C. R. (2011). Access to liquidity and corporate investment in europe during the financial crisis. *Review of Finance*, 16(2):323–346.

Capone, G., Malerba, F., Nelson, R., Orsenigo, L., and Winter, S. (2019). History friendly models: retrospective and future perspectives. *Eurasian Business Review*, 9(1):1–23.

Carlei, V. and Nuccio, M. (2014). Mapping industrial patterns in spatial agglomeration: A som approach to italian industrial districts. *Pattern Recognition Letters*, 40:1–10.

Carota, C., Durio, A., and Guerzoni, M. (2014). An application of graphical models to the innobarometer survey: A map of firms' innovative behaviour. *Department of Economics and Statistics "Cognetti de Martiis" Working Paper Series*.

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

Cefis, E. and Marsili, O. (2006). Survivor: The role of innovation in firms' survival. *Research Policy*, 35(5):626–641.

Cefis, E. and Marsili, O. (2012). Going, going, gone. exit forms and the innovative capabilities of firms. *Research Policy*, 41(5):795–807.

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ciccone, C. and Hall, R. (1996). Productivity and the density of economic activity. *The American Economic Review*, 86(1):54–70.

Colombelli, A., Krafft, J., and Quatraro, F. (2013). Properties of knowledge base and firm survival: Evidence from a sample of french manufacturing firms. *Technological Forecasting and Social Change*, 80(8):1469–1483.

Colombelli, A., Krafft, J., and Vivarelli, M. (2016). To be born is not enough: The key role of innovative start-ups. *Small Business Economics*, 47(2):277–291.

Cowling, M., Siepel, J., Liu, W., and Murray, G. (2014). Are highly innovative firms also high growth firms? and what are the causal events that deliver high sales growth? *Frontiers of Entrepreneurship Research*, 34(14):7.

Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy*, 11(3):147–162.

Dosi, G. (1988). Sources, procedures, and microeconomic effects of innovation. *Journal of Economic Literature*, pages 1120–1171.

Estolatan, E., Geuna, A., Guerzoni, M., and Nuccio, M. (2018). Mapping the evolution of the robotics industry: A cross country comparison. *White Paper Series 2018/8 - Munk School of Global Affairs and Public Policy*.

Expósito, A. and Sanchis-Llopis, J. A. (2019). The relationship between types of innovation and smes' performance: A multi-dimensional empirical assessment. *Eurasian Business Review*, 9(2):115–135.

Feldman, M. P. (2001). The entrepreneurial event revisited: Firm formation in a regional context. *Industrial and Corporate Change*, 10(4):861–891.

Fetters, M., Greene, P. G., and Rice, M. P. (2010). *The development of university-based entrepreneurship ecosystems: Global practices*. Edward Elgar Publishing.

Fleming, T. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Fontana, M. (2010). Can neoclassical economics handle complexity? the fallacy of the oil spot dynamic. *Journal of Economic Behavior & Organization*, 76(3):584–596.

Fontana, R., Guerzoni, M., and Nuvolari, A. (2008). Habakkuk revisited: A history friendly model of american and british technology in the nineteenth century. Technical report, Jena economic research papers.

Freeman, C. (1994). The economics of technical change. *Cambridge Journal of Economics*, 18(5):463–514.

Fudenberg, D., Gilbert, R., Stiglitz, J., and Tirole, J. (1983). Preemption, leapfrogging and competition in patent races. *European Economic Review*, 22(1):3–31.

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Geels, F. W. (2013). The impact of the financial–economic crisis on sustainability transitions: Financial investment, governance and public discourse. *Environmental Innovation and Societal Transitions*, 6:67–95.

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

Gould, P. (1981). Letting the data speak for themselves. *Annals of the Association of American Geographers*, 71(2):166–176.

Griliches, Z. and Mairesse, J. (1995). Production functions: The search for identification. Technical report, National Bureau of Economic Research.

Guerzoni, M. (2010). The impact of market size and users' sophistication on innovation: The patterns of demand. *Economics of Innovation and New Technology*, 19(1):113–126.

Harabi, N. (1995). Appropriability of technical innovations an empirical analysis. *Research Policy*, 24(6):981–992.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Helmers, C. and Rogers, M. (2010). Innovation and the survival of new firms in the uk. *Review of Industrial Organization*, 36(3):227–248.

Hodorogel, R. G. (2009). The economic crisis and its effects on smes. *Theoretical & Applied Economics*, 16(5).

Hoffmann, A. O., Post, T., and Pennings, J. M. (2013). Individual investor perceptions and behavior during the financial crisis. *Journal of Banking & Finance*, 37(1):60–74.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Howells, J. (1999). Regional systems of innovation. *Innovation Policy in a Global Economy*, pages 67–93.

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

Inklaar, R. and Yang, J. (2012). The impact of financial crises and tolerance for uncertainty. *Journal of Development Economics*, 97(2):466–480.

Jordan, M. I. (1998). *Learning in graphical models*, volume 89. Springer Science & Business Media.

Kim, C. and Lee, K. (2003). Innovation, technological regimes and organizational selection in industry evolution: A "history friendly model" of the dram industry. *Industrial and Corporate Change*, 12(6):1195–1221.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):2053951714528481.

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

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, 160:3–24.

Lee, S. Y., Florida, R., and Acs, Z. (2004). Creativity and entrepreneurship: A regional analysis of new firm formation. *Regional Studies*, 38(8):879–891.

Leone, R. A. and Struyk, R. (1976). The incubator hypothesis: Evidence from five smsas. *Urban Studies*, 13(3):325–331.

Liu, X. (2009). Impacts of the global financial crisis on small and medium enterprises in the people's republic of china. *ADBI Working Paper*.

Macqueen, J. (1967). Some methods for quantification of the multivariate observations, western management science institute, university of california. Technical report, Working paper 96.

Malerba, F., Nelson, R., Orsenigo, L., and Winter, S. (1999). History-friendly models of industry evolution: the computer industry. *Industrial and Corporate Change*, 8(1):3–40.

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

Malerba, F. and Orsenigo, L. (2002). Innovation and market structure in the dynamics of the pharmaceutical industry and biotechnology: Towards a history-friendly model. *Industrial and Corporate Change*, 11(4):667–703.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Nuccio, M. and Guerzoni, M. (2019). Big data: Hell or heaven? digital platforms and market power in the data-driven economy. *Competition & Change*, 23(3):312–328.

OECD (2018). Guidelines for collecting, reporting and using data on innovation.

Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy*, 13(6):343–373.

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: Possibilities and problems. *Scientometrics*, 7(1-2):77–99.

Pearl, J. and Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Peric, M. and Vitezic, V. (2016). Impact of global economic crisis on firm growth. *Small Business Economics*, 46(1):1–12.

Perlich, C., Provost, F., and Simonoff, J. F. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, 4:211–255.

Prensky, M. (2009). H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: Journal of Online Education*, 5(3).

Rafferty, A., Rees, J., Sensier, M., and Harding, A. (2013). Growth and recession: Underemployment and the labour market in the north of england. *Applied Spatial Analysis and Policy*, 6(2):143–163.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Rocha, H. O. (2004). Entrepreneurship and development: The role of clusters. *Small Business Economics*, 23(5):363–400.

Rothwell, R. (1994). Issues in user–producer relations in the innovation process: The role of government. *International Journal of Technology Management*, 9(5-7):629–649.

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

Schumpeter, J. A. (1912). Theorie der wirtschaftlichen entwicklung. leipzig: Dunker & humblot. *The Theory of Economic Development*.

Schumpeter, J. A. (1942). *Socialism, capitalism and democracy*. Harper and Brothers.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Simon, J. P. (2016). How to catch a unicorn: An exploration of the universe of tech companies with high market capitalisation. Technical report, Joint Research Centre (Seville site).

Sorenson, O. and Audia, P. G. (2000). The social structure of entrepreneurial activity: Geographic concentration of footwear production in the united states, 1940–1989. *American Journal of Sociology*, 106(2):424–462.

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

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

Strasser, H. and Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, 8:220–250.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14:323–348.

Stucki, T. (2013). Success of start-up firms: The role of financial constraints. *Industrial and Corporate Change*, 23(1):25–64.

Teece, D. J. (2012). Dynamic capabilities: Routines versus entrepreneurial action. *Journal of Management Studies*, 49(8):1395–1401.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, New York.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28:3–28.

Wagner, S. and Cockburn, I. (2010). Patents and the survival of internet-related ipos. *Research Policy*, 39(2):214–228.

Zenghelis, D. (2012). A strategy for restoring confidence and economic growth through green investment and innovation. *Policy Brief*.

Žliobaitė, I. (2010). Learning under concept drift: an overview. Technical report, Vilnius University, arXiv:1010.4784.

A Appendix

A.1 Missing Value analysis

The AIDA database is a valuable source of information but reporting is mandatory for only few variables only. Hence, a missing value analysis (MVA) is needed to avoid information loss when applying machine learning algorithms, which immediately discard all observations containing missing values (NAs). We propose an MVA to identify variables and observations containing the highest number of NA and to choose which ones to delete. It is a semi-automatic approach which balances the loss of information with the introduction of a source of extra variability. No imputation of missing data is undertaken to avoid introducing potential bias in variables with too many NAs[9].

The MVA starts with the 2013 sample and only afterwards evaluates the status of the 2008 one. Since, in 2013 sample, the variables measured in 2015 had not yet been fully incorporated into AIDA at the time of the inspection (July 2016), we chose to drop them immediately[10]. Hence, we start with 800 available variables (174 do not change over time, while 786 are the results of the firm observation over three years: namely, 262×3), and we discard 262 accounting variables, *i.e.* we still retain 538 variables for the 68,316 start-ups of 2013. We also control for the presence of duplicated variables. Subsequently, we define the number of NAs for each variable and for each firm, obtaining the distributions shown in Table A.1 and Figure A.1. We observe that the NAs affect both INNs and NOINNs in a similar way. We choose to drop observations with a number of NAs higher than those in the third quartile, *i.e.* with more than 290 missing over the 538 variables. We obtain a dataset composed of 51,496 observations (including 796 innovative start-ups).

Then, we drop variables with a number of NAs exceeding those in the first quartile (3,968). We retain 51,496 firms observed on 127 variables. In order to employ a model trained on 2013 data for predicting values in 2008, the 2008 and 2013 samples need to have the same variables. We also undertake the same analysis on the 2008 sample so as not to lose too many firms from the 2008 sample. After removing variables already discarded in the 2013 sample, three new variables containing more than the 30% of the missing values are identified. Hence, we drop them from both samples, and we discard all observations still containing NAs. We obtain two final datasets with 124 variables: that of 2013 contains 45,576 firms while that of 2008 one

[9] Note that NAs are much too diffused among the variables and observations, and therefore multiple imputation will add an extra variability to observed variables not justified. Even if we limit the multiple imputation to some crucial variables, we do not have enough complete observations in the dataset to finalise the NA completion.

[10] Note that management variables, which contain a huge amount of unstandardised text, are discarded from the beginning of the data construction process.

contains 39,295 firms[11]. The 2008 sample is, then, enriched with further economic variables, such as EBITDA, R&D investments, employees, and IPR investments observed yearly from 2009 to 2018. After the MVA, the proportion of INNs/NOINNs in the 2013 sample is consistent with the original one, slightly growing from 1.5% to 1.59%.

Table A.1: Missing value distribution observed in 538 variables for the 68,316 observed 2013 start-ups according to the INNs and NOINNs classification

Missing value	INNs/NOINNs	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Variables	INNs+NOINNs	0	11,170	11,830	19,530	23,620	68,320
	INNs	0.00	88.25	213.00	274.90	297.00	1010.00
	NOINNs	0	11,050	11,620	19,260	23,370	67,310
Firms	INNs+NOINNs	24.0	70.0	98.0	153.8	290.0	530.0
	INNs	32.0	71.0	92.0	146.4	284.0	360.0
	NOINNs	24.0	70.0	98.0	153.9	290.0	530.0

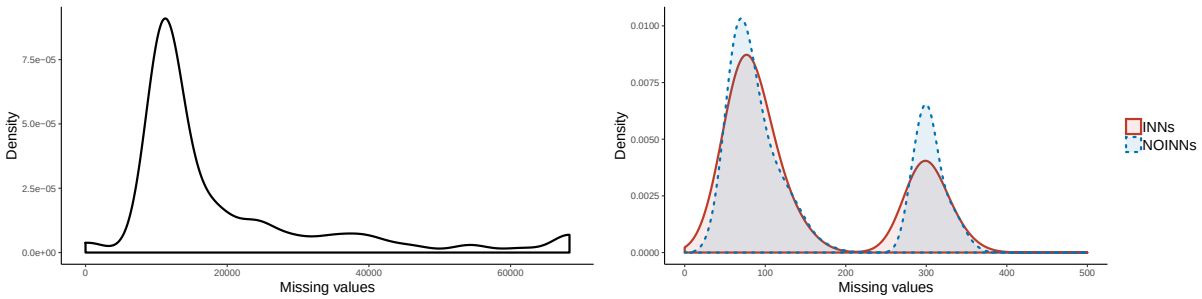


Figure A.1: The left panel shows the missing value distribution in 538 variables for the 68,316 observed 2013 start-ups. The right panel shows missing value distributions in observations separated into INNs ($n = 1,010$) and NOINNs ($n = 67,306$), over 538 variables

Table A.2: Description of starting and final sample dimension

Sample	2008	2013
Total	65,088	68,316
NA	25,793	22,740
Final sample	39,295	45,576

A.2 Algorithms

For the training of the model, we select seven algorithms known in the machine learning, neural network and econometric literature, briefly presented in what follows.

[11]Note that, without doing this last MVA step, in the 2008 sample only 18,078 firms would be left, representing less than the 28% of the initial amount of 2008 start-ups.

1. The binary recursive partitioning algorithm (RPART)(Breiman et al., 1984) is a tree-based method (Hastie et al., 2008) grounded in a top-down approach in which the partition starts at the top of the tree. Starting from all observations in a single region, the algorithm only successively splits the space via a two further branches of the tree. The Gini's coefficient method is used for the tree variable selection. RPART defines the best split at each step, and predictions are easily interpretable, differently from other classification algorithms (Gareth et al., 2013). We use RPART in R with the function `rpart()` in the package `rpart`.

2. The classification tree (TREE) (Breiman et al., 1984; Ripley, 1996) is based on binary recursive partitioning, given the classification INNs/NOINNs. It recursively chooses splits from the selected independent variables. Numerical variables are split at a given value α in each node, while categorical variables are split according to two non-empty sets of unordered levels. At each step TREE selects the split which minimises classification impurity. We use TREE in R with the function `tree()` in the package `tree`.

3. The conditional inference tree (CTREE) (Hothorn et al., 2006; Strasser and Weber, 1999) estimates a regression relationship by binary recursive partitioning in a conditional inference framework. It uses a permutation test in order to select the set of variables that maximises the Gini coefficient, differently from other tree-based methods that just select one variable at each step. We use CTREE in R with the function `ctree()` in the package `party`.

4. The bagging algorithm (BAG) (Breiman et al., 1984; Strobl et al., 2009; Gareth et al., 2013), or bootstrap aggregation, is based on the necessity to reduce the variance of the statistical learning tree previously described. It is based simply on the idea that the variance can be reduced if, instead of only one training set, we use the average of more training sets. For this reason, it is based on the aggregation of many decision trees. BAG generates M different bootstrapped training data sets (with an increment in computation time), and then it trains the method on the M bootstrapped sets in order to average all the obtained predictions. Here we use RPART as the basis of BAG. We use BAG in R with the function `bagging()` in the package `ipred`.

5. The Logit regression model (LOGIT) (McCullagh and Nelder, 1989) is used here as the benchmark and widely-used econometric model. It can be seen as a generalised linear model based on a Logit link function (Agresti, 2002; McCullagh and Nelder, 1989) or a random utility model for discrete choices (Train, 2009). It estimates a linear relation between the independent variables and the logit of the INNs probability. Its accuracy

suffers in the presence of huge datasets, as in our case (Perlich et al., 2003). We use LOGIT in R with the function `glm()`.

6. The naïve Bayesian classifier (NB) (Alpaydin, 2014), in its particular binary version, is based on the estimation of the conditional a-posterior probabilities of INNs, given the selected independent predictors, using the well-known Bayes rule. We use NB in R with the function `naiveBayes()` in the package `e1071`.

7. The artificial neural network (ANN)(Bishop, 1995; Ripley, 1996) is a single hidden layer back-propagation network (Hastie et al., 2008). It is based on the artificial reproduction of the functioning of the brain (Pos, 1989), therefore ANN is a nonlinear statistical model based on a two-stage estimator. We use ANN in R with the function `nnet()` in the package `nnet`.

A.3 Optimal cut-off and mixture weight optimisation

Part of the methodology introduced in this contribution is new, therefore new R functions have been coded to undertake the analysis. A first built-in function implements three criteria for the selection of the optimal cut-off in each algorithm:

1. the Youden index (J) method, which defines the optimal cut-point as the point maximising the difference between the true positive rate and the false positive rate (namely, the Youden function) over all possible cut-point values;

2. the point closest to the (0,1) corner in the ROC plane method, which defines the optimal cut-point as the point which minimises the Euclidean distance between the ROC curve and the (0,1) point;

3. the optimal cut-point method which selects as the optimal cut-off the point which maximises the product of sensitivity and specificity.

The confusion matrix in Table 7 is the result of the application of the second criterion. Similar results have been obtained applying the other two approaches.

A second built-in R function finds the optimal mixture weights, following the approach below. First, we select two (or more) candidate algorithms to compose the mixture (here $alg1$ =BAG and $alg2$ =ANN), according to their performance emerging from the study of the ROC curve and of the confusion matrix. Second, we retain the predicted probabilities (*pred.prod*), under the selected algorithms, for INNs (positive) and NOINNs (negative). Third, we select a mixture of weights α and $1-\alpha$ in the support (0,1) according to an optimisation process. The latter

simultaneously maximises, for all α in the support, i) the Euclidean distance between the INNs and the NOINNs predicted probabilities; and ii) the area under the ROC (AUC). A unique solution of this maximisation process exists and selects α, such that the predicted probability of the mixture is defined as follows:

$$pred.prob.mixture = \alpha * pred.prod.alg1 + (1 - \alpha) * pred.prod.alg2.$$

A.4 Further descriptive statistics on the data

Further descriptive statistics on the 2013 sample are here proposed. We study the distribution of employments (Figure A.2) and EBITDA (Figure A.3 and Table A.4) in the two first years of activity (2013 and 2014) according to the inscription (or not) on the special section in the Italian companies register. We also compare the geographical distribution, at NUTS3 level, of INNs and NOINNs in 2013 (see Figure A.4).

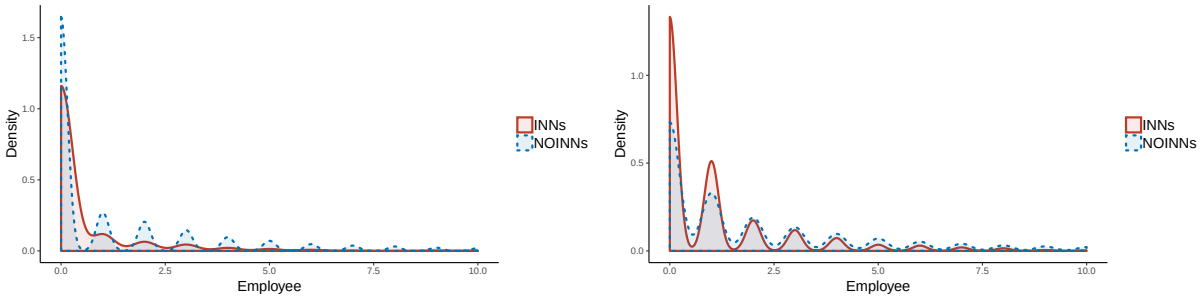


Figure A.2: Employees distributions in 2013 and 2014 of INNs and NOINNs in the 2013 sample

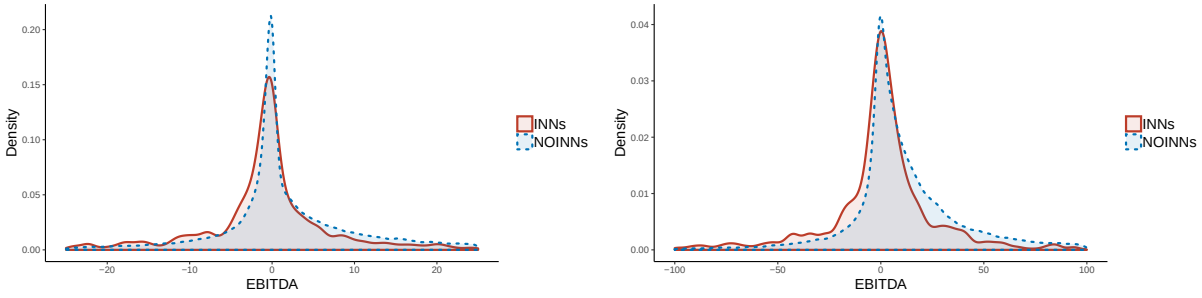


Figure A.3: EBITDA distributions in 2013 and 2014 of INNs and NOINNs in the 2013 sample

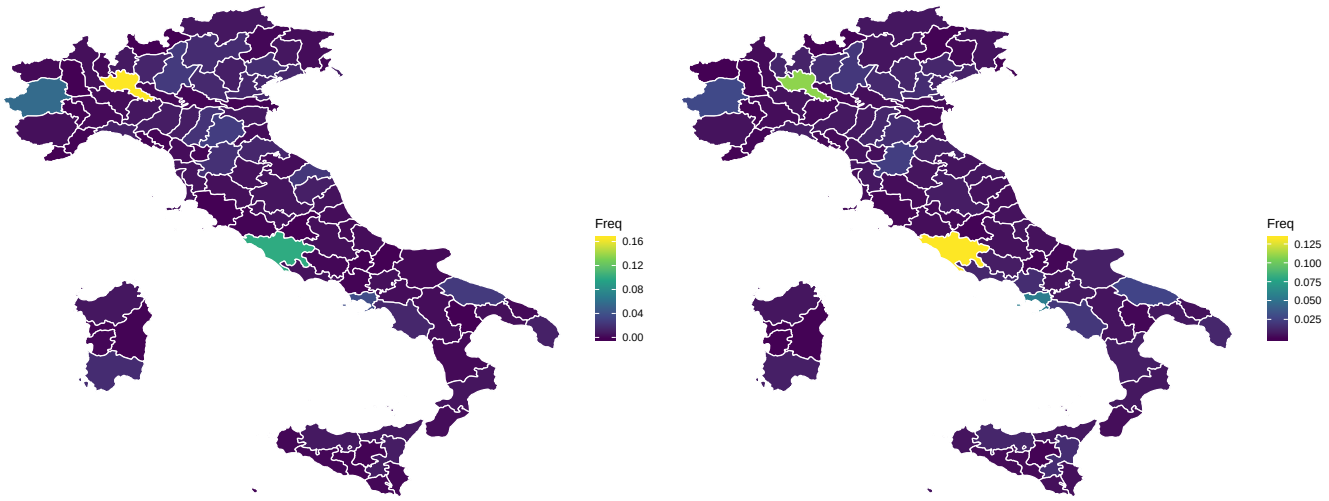


Figure A.4: Geo-localisation of the 2013 sample. The INNs are presented in the left panel, and the NOINNs in the right one

Table A.3: Juridical form of the 2013 innovative start-ups ($n = 1010$)

Juridical Form	Number of innovative start-ups
S.C.A.R.L.P.A.	14 (1.39%)
S.P.A.	13 (1.29%)
S.R.L.	820 (81.19%)
S.R.L. a capitale ridotto	11 (1.09%)
S.R.L. a socio unico	45 (4.46%)
S.R.L. semplificata	106 (10.50%)
Società consortile a responsabilità limitata	1 (0.1%)

Source: AIDA.

Table A.4: 2013 innovative start-ups - EBITDA distribution observed in the first (2013) and the second (2014) year of activity

Var		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	t-stat	p-val
EBITDA	INNs	-537.40	-3.68	-0.39	-3.02	1.95	207.20	213		
								(21.09%)		
2013	NOINNs	-12360	-1.53	0.01	11.73	7.51	120,000	9,946		
								(14.78%)		
									4.9949	0.00
EBITDA	INNs	-895.20	-10.22	0.09	-11.37	9.18	379.40	78		
								(7.72%)		
2014	NOINNs	-18090	-1.18	6.43	37.66	25.43	261,000	11,110		
								(16.51%)		
									7.7044	0.00

Source: AIDA.