This is an author version. The final version is published in *Online Information Review*, with DOI 10.1108/OIR-11-2016-0321

Do corporate websites' changes reflect firms' survival?

Desamparados Blazquez[1], Josep Domenech[1] and Ana Debón[2]

[1]Department of Economics and Social Sciences, Universitat Politècnica de València, Valencia, Spain
[2]Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València, Valencia, Spain

Abstract

Purpose: The objective of this paper is to analyze to what extent changes in corporate websites reflect firms' survival. Since keeping a website online involves some costs, it is likely that firms would invest resources on it only when they are active and healthy. Therefore, when a firm dies, this event is likely to be manifested on its website as lacking updates or being down.

Design/methodology/approach: Changes in the corporate websites of a panel of Spanish firms were tracked between 2008 and 2014 in order to evaluate our approach. The status of websites, classified according to the type of change undergone, was used to infer firms' activity status (active or inactive). Multi-period logistic regressions and a duration model were applied to study the relationship among the website status and the firm's status.

Findings: Results showed that changes in website contents clearly reflect the firm's status. Active firms were mainly associated with updated corporate websites, while inactive firms were more associated with down websites. In fact, results confirmed that the firms' death hazard increases when the website activity lowers.

Originality/value: Although online information is increasingly being used to monitor the economy, this is the first study to connect online data to firms' survival. Our results revealed a new source of information about business demography and evidenced corporate websites as a fresh source of high granularity business data.

Keywords: business demography statistics, corporate websites, online economic information, firms' survival

Article classification: Research paper

Introduction

Business demography is one of the economic aspects that attracts more attention from governments and policy makers. Indeed most official statistics institutions (e.g., Eurostat, Office for National Statistics of Spain, Australian Bureau of Statistics) carry out detailed surveys to monitor the active population of firms, their birth, survival and death. The interest shown in business demography statistics relies on the important role they play in economic growth, productivity and employment (Eurostat and OECD, 2007).

In the Digital Era, the prominent role of the Internet in economy and society, along with the development of advanced computer systems and architectures, opens up new ways of monitoring economic activities (Blazquez and Domenech, 2014; Vaughan, 2014) and, therefore, business demography. The Internet and the World Wide Web (WWW) have become basic tools for the daily activities of individuals and companies, whose importance is increasing in both developed and developing countries. For consumers, the WWW is a convenient instrument to find information on products and services, and if available, to purchase them online. For companies, the WWW is an inexpensive channel to not only offer information about their products, services and activities, but to also make transactions with customers more quickly and more flexibly. In this context, companies have massively developed their websites in order to be present on the digital channels. For instance, 75% of companies are present on the WWW in Spain (INE, 2016), which is the country on which our empirical analysis focuses.

Corporate websites constitute the most formal and official representation of firms on the Internet. Generally, firms describe their main activities, products and intended strategies on their websites. Therefore, corporate website contents are necessarily connected to business activity to some extent, which has been recently studied from different perspectives. For instance, evidence has been found for the relation of website contents to technology adoption (Youtie et al., 2012; Arora et al., 2013), innovation activities (Arora et al., 2016; Gök et al., 2015), firms' growth (Li et al., In press), and firms' export orientation (Blazquez and Domenech, 2014, 2017).

Once it has been proved that firms' activities emerge on their corporate website, the question as to whether firms' inactivity is also manifested on their website arises. Keeping a website online involves some costs, such as fees for domain name registrations or server hosting upkeep. Furthermore, costs increase when companies wish to keep website contents and related technologies up to date. Active companies are expected to regularly modify their website to include new products or services, renew its design and offer additional functionalities, or to inform potential customers about new offers or promotions. Since keeping the website updated requires firms to mobilize some resources (financial, working or both), it is plausible that only active and healthy firms would invest their resources to that end. Therefore if a company dies, this event is likely to be manifested on its website as lacking updates or simply as being down.

Most academic research conducted on firms' survival has focused on the factors that contribute to keep companies alive. The firm's age, size, productivity or profitability have been widely studied as determinant factors that contribute to firms' survival. Despite the important role that the WWW plays in today's business, there are no studies relating the WWW to firms' survival. Corporate websites are a fresh source of business information as they are publicly accessible and provide access to high granularity (company level) data which are generally updated regularly. For these reasons, they have been used to analyze some company behaviors or strategies. Nevertheless, the approach of employing corporate websites to analyze firms' survival is novel.

This paper analyzes to what extent changes in corporate websites reflect firms' survival. This work hypothesizes that if a firm dies, it is very likely that its website goes down, which could happen shortly before or after the firm's death. If this relationship exists, then the corporate website status (down, unchanged or updated), whose retrieval and tracking are inexpensive, could help monitor firms' survival. To evaluate our proposal, the corporate website changes and firm activity status of a panel of Spanish firms were monitored for 7 years and analyzed. Multi-period logistic regressions and survival analysis were run to infer the firms' activity status. The results showed that the corporate website status clearly reflects the firm's status.

The rest of the paper is organized as follows. Next section reviews the literature on firms' survival and the detection of economic information through web analysis. The subsequent section describes the data used and the methodology applied for the empirical analysis. The following section describes the results, including a data overview and a comprehensive analysis of model estimations. Finally, the last section draws some concluding remarks.

Theoretical background

This section provides background on firms' survival analysis and the detection of economic activities on the WWW. First, the related literature on firms' survival analysis is reviewed, reporting the firm-related variables which researchers have paid more attention to. Second, a review on the detection of firms' activities through web and online data is provided, motivating the exploration of whether firms' inactivity could also be detected through their websites.

Firms' survival analysis

Firms' survival is a hot topic for researchers because of its implications for business success, economic stability and growth. However, it was not until the 1990s, promoted by the increasing economic globalization, when the academic community started to focus on analyzing firms' survival. Firms started to face new challenges in a more complex and turbulent environment, which is the reason why determining which characteristics or actions could help them survive was more necessary than ever before.

The seminal work of Evans (1987), and other later ones like Audretsch (1991), Mata and Portugal (1994) and Geroski (1995), helped expand the field of firms' survival, focusing on a systematic analysis of which industry-specific and firm-specific factors affected companies' survival, and in which direction.

Regarding firm-related structural variables, firm size and age have been widely explored since they have been considered stylized facts related to firm survival (Geroski, 1995). Generally, firm size has been found to increase the likelihood of a firm's survival, especially for new entrants (Agarwal and Audretsch, 2001; Cefis and Marsili, 2005;

Geroski et al., 2010). Larger firms usually have more financial and human resources available as well as a solid structure after reaching a certain production level. These factors could help reduce the risk of mortality.

A similar pattern has been exhibited by firm age, which has been found to principally increase the likelihood of survival (Audretsch et al., 2000; Manjón-Antolín and Arauzo-Carod, 2008). Older firms have had the possibility of acquiring experience in how the market works and which strategies are more profitable for them. This could help them survive compared to newcomers. That is, the effect of experience on firms' survival is generally positive.

Other structural variables whose relationship with firms' survival has been consistently studied by researchers include the firm's debt structure, its level of productivity and its level of profitability (Audretsch et al., 2000; Delmar et al., 2013; Görg and Spaliara, 2014). These variables are closely related to the firm's level of success, stability and health, and are thus potentially influential for the likelihood of a firm to survive.

The technological intensity of the activity sector in which the company operates has also been considered in other firms' survival studies (Esteve-Pérez and Mañez-Castillejo, 2008; Giovannetti et al., 2011). The first findings pointed out that firms had more difficulties to survive in high-technological sectors. However, an opposite pattern was found later; providing highly technological products and services requires firms to develop sophisticated skills, to focus on innovation and knowledge, and these factors are potential contributors to firms' survival particularly within today's complex economic frame.

More recent studies continue providing insights into how the classical firm structural factors, e.g., size, age or financial ratios, and environmental factors like financial crises, location or the specific business life cycle, contribute to increase or decrease the likelihood of firms' survival (Basile et al., 2017; Gémar et al., 2016; Guariglia et al., 2016). The results of most reviewed literature works aim to serve mainly as guidelines in managers' decision-making processes, who could use this information to run or promote strategies that can contribute to firms' survival.

However, none of the studies in the literature has analyzed the relationship between firms' survival and corporate websites. The role of corporate websites in the firms' strategies is basic in today's digital society, and is expected to gain importance in the future. While accounting data have been useful for predicting firm's failure, they are not perfect measures of a firm's operational and financial status (Astebro and Winter, 2012), so complementing them with online data could offer a better idea of a firm's health.

For these reasons, it is relevant to confirm to what extent the corporate website status is related to the firm's status of activity, and to explore if the information provided on the website can be used for monitoring a firm's survival. The next subsection reviews the literature on the suitability of the WWW to reflect business activity, which motivated our study on checking whether it also reflects a firm's inactivity, i.e. a firm's death.

Capturing firms' economic activities through web data

Every minute of the day, thousands of individuals, companies and public organisms generate, post and share information through the Internet. These online activities leave a digital footprint behind that can be tracked and, if properly processed and analyzed, can help describe their economic and social behavior.

The detection of behavioral and consumer patterns, and economic and business activities, through online data is an incipient research field whose importance is starting to increase at the same time as the adoption of the Internet is expanding worldwide. This generalized expansion in Internet use is affecting the way companies do business, which are being enforced by the current digital context to go online. To do so, firms generally start by implementing websites, which are the most official representation of their image and could, at the same time, be used as a commercial channel.

Indeed websites are relevant sources of online data whose potential for detecting and monitoring economic activities has remained unexplored until quite recently. Websites have a complex structure that differs from one case to another, making the process of extracting, processing and analyzing information difficult to standardize and automate to allow massive data exploitation compared to traditional databases. However, websites also present many advantages, such as: they are publicly accessible, provide fresh information and can be analyzed at any time, which traditional databases generally do not. In particular, corporate websites are attracting more attention because they are being increasingly adopted by firms, which normally use them to reflect their characteristics, products and intended strategies. Therefore, websites have become rich sources of business information. For these reasons, specific technologies and methodologies for extracting and analyzing web data are being developed (Munzert et al., 2015).

The first works about detecting economic or business information on corporate websites were published a decade ago. Following a non-automated approach, Overbeeke and Snizek (2005) captured different corporate culture dimensions by analyzing the text and images available on a set of corporate websites, while Meroño-Cerdan and Soto-Acosta (2007) found that external web content related positively to firm performance.

Firms' corporate social responsibility and sustainability strategies, and their levels of adoption, have also been successfully detected in corporate websites contents (Gallego Álvarez et al., 2008; Tagesson et al., 2009; Tang et al., 2015). This has been done, for instance, by detecting the occurrence number of keywords related to green products (Albino et al., 2009). This measure has been extended and successfully used in other studies that have focused on novel technology industries. In their work, Libaers et al. (2010) found six types of business models for commercializing novel technology by automatically analyzing the frequency with which specific keywords were present on the corporate websites of the firms under study.

Following an automatic approach, Youtie et al. (2012) and Arora et al. (2013) applied web scraping and content analysis techniques to corporate websites, including the count of keywords, in order to track the technology adoption strategies of firms on emerging technology sectors. Innovation is another relevant topic which has been recently detected through web mining techniques. Gök et al. (2015) and Arora et al. (2016) successfully detected firms' R&D activities by analyzing corporate websites contents. For their part, Li et al. (In press) tracked firms' sales growth in a Triple Helix context.

The first attempt to generalize the automatic analysis of corporate websites to discover economic information was introduced by Domenech et al. (2012). This work presents a web data mining system architecture that manages the process of crawling and analyzing corporate websites, which was successfully tested for finding web-based indicators for firms' size. This system was adapted by Blazquez and Domenech (2017) to detect firms' export orientation by automatically analyzing their corporate websites since a previous manual analysis found that websites potentially reflect such business activity (Blazquez and Domenech, 2014).

Based on previous research, in which corporate websites were demonstrated to reflect economic information and business activities, this paper hypothesizes that detecting firms' inactivity by analyzing the data retrieved from corporate websites is also possible.

Data and methodology

This section first describes the structure of the data used herein and how it was obtained. Second, it reviews the methodology employed, which relies on multi-period logistic regression models to detect the ability of website status to predict firm's activity status, and a duration model to provide a deeper understanding of how the web status relates to a firm's survival.

Data

The initial study sample included 780 companies[1] established in Spain from manufacturing, services and other sectors (NACE Rev.2[2] codes 10-95), all of which were active and had a website in 2008. The sample was retrieved by a simple random sampling design from the SABI database (Bureau van Dijk, 2010), being eligible all firms in the database that met four criteria: being active, being located in Spain, belonging to any of the mentioned activity sectors, and having a website; all of them referred to year 2008, in which this study begins. The dataset consists of a panel of economic and online data for these firms for years 2008 to 2014. The economic information was retrieved from company financial records by accessing a more recent version of the SABI database in January 2016, and 2014 was the last year for which complete company economic records were found.

The online information was collected by accessing the corporate websites with the Wayback Machine of the Internet Archive (Kahle and Gilliat, 2016), which is a public and free repository of snapshots of about 484 billion web pages. The Internet Archive captures and stores websites on a daily basis, allowing users to access them and track their history and evolution over time. However, there are some limitations as to its use: its inability to capture websites that prevented themselves from being explored by web crawlers by means of the robots exclusion standard (i.e. robots.txt); its limited ability to capture Flash content; the fact that it does not crawl the whole WWW, so some websites are not captured and, therefore, their evolution over time cannot be tracked; and that not every website is frequently captured, even some of them less than once a year. These limitations prevented us from tracking the evolution of some corporate websites.

For these reasons, the firms whose websites were not found in the Wayback Machine were removed from the initial sample. This gave a final sample of 720 companies to be included in the study, of which 674 survived the whole time period, while the remaining 46 died at some time. Only the years from 2010 to 2014 were included in the data analysis presented below in order to track any website changes compared to the previous year and to align the website status to the time at which financial information is available. In order to take into account the different moments of time at which company data are available, information from the financial statements was lagged two periods in the empirical analyses. That is, it is possible to know the corporate website status at time t, but at this time the most recent financial statements available correspond to $t - 2$.

Some website captures in a particular year t of the study period were not available in the Wayback Machine. This resulted in an unbalanced panel with 3254 observations, of which 3152 corresponded to the firms that survived to 2014, and the remaining 102 to those that died during the study period.

To account for changes in the corporate websites, the procedure followed consisted in querying the Wayback Machine with the URL of each company's website and checking the homepage for each year studied. The observed changes were coded into the variable

Web_status, which could take five different values depending on the status or type of change experimented each year. These five levels are defined as:

- Code 1: the website is down. This includes the websites that do not work (e.g., HTTP Error 404 Not Found) or whose domain name has expired or is for sale.
- Code 2: the website remains unchanged. This includes the cases in which the website remains exactly the same as its previous year's version.
- Code 3: the website has undergone minor changes. These changes include the removal or addition of sections, options, pictures and contents.
- Code 4: the website has undergone major changes. These changes refer to a new website design, so that it completely differs from to the previous year's version; this may imply a change in the technology used to build the website.
- Code 5: the website has not been captured by the Wayback Machine. These cases were processed as missing data and were removed from the final sample as it was impossible to determine the website status.

The dataset also included the economic variables classically related to firms' survival according to the reviewed literature. These variables, together with the firm's status (active or inactive), were retrieved from the SABI database and complemented with the information taken from the Official Gazette of the Commercial Registry[3] to account for merges and acquisitions. The following economic variables were retrieved:

- $Active_{i,t}$: Dichotomous variable that takes a value of 1 if firm i is active in year t, and 0 otherwise[4].

- $Size_{i,t}$: Quantitative variable measured as the logarithm of the number of employees of firm i in year t. It is a proxy to firm size.

- $Age_{i,t}$: Quantitative variable measured as the number of years since firm i was established up to year t. It is a proxy to the firm's experience.

- $Debt_{i,t}$: Quantitative variable measured as the percentage of debt of firm i in year t.

- $Productivity_{i,t}$: Quantitative variable measured as the value added per employee (in millions of euros) of firm i in year t.

- $Profitability_{i,t}$: Quantitative variable measured as the ratio of economic profitability of firm i in year t. This ratio, known as 'Return on Assets (ROA)', is obtained from dividing the operating profit by total assets.

- $High\_tech_{i,t}$: Dichotomous variable that takes a value of 1 when the economic activity of firm i in year t is considered of high or medium-high technological intensity according to the Eurostat Classification (Eurostat, 2014), and 0 otherwise.

Multi-period logistic regression

In a first approach, firms' survival was studied by multi-period logistic regression models. These models are useful for examining how some independent variables are related to a dependent variable when the data used as input include individuals observed over time, which was the case of the present study, and have been applied successfully in existing firms' survival studies (Bridges and Guariglia, 2008; Jacobson and von Schedvin, 2015).

The dependent variable in this research is whether or not the firm's status is active ($Active_{i,t}$), which is a dichotomous variable that makes logistic regression suitable for analyzing the relation with covariates. The models used include fixed-time effects to account for the changing economic and politic situation that affects the baseline probability of being active each year. Analytically, the model is represented as:

$$\theta_{i,t} = \ln\left(\frac{P(y_{i,t}=1)}{1-P(y_{i,t}=1)}\right) = \beta' X_{i,t} + \gamma_t \tag{1}$$

where $\theta_{i,t}$ is the logit, $P(y_{i,t}=1)$ is the probability of occurrence of status '1' of the dependent variable $y_{i,t}$, β' is the vector of regression coefficients, $X_{i,t}$ is the vector of covariates for firm i in year t, and γ_t are the time specific parameters that reflect the unobservable events that affect all firms each year.

This model is used to first assess the relation between the WWW and firms' status, as it estimates the probability of a firm being active given its website status in a first specification, and given this website status and a number of economic variables in a second specification. Both model specifications controlled for the economic juncture or period effect by including dummies for each year considered in the study. Accordingly, the first model was defined as follows:

$$\theta_{i,t} = \ln\left(\frac{P(Active_{i,t}=1)}{1-P(Active_{i,t}=1)}\right) = \beta_0 + \alpha Web\_status_{i,t} + \gamma_t \tag{2}$$

where $P(Active_{i,t}=1)$ is the probability that firm i is active in year t, and the logit, $\theta_{i,t}$ is regressed on the explanatory variable $Web\_status_{i,t}$ and the fixed-effect of time, captured by γ_t.

An extended model was specified by including also the firms' economic variables that can affect firm survival according to the literature. The variables that were finally selected were those that varied with an admissible level of significance (p<0.05) between both groups of firms and did not highly correlate. This second specification was defined as follows:

$$\theta_{i,t} = \ln\left(\frac{P(Active_{i,t}=1)}{1-P(Active_{i,t}=1)}\right) = \beta_0 + \alpha Web\_status_{i,t} + \beta' Z_{i,t-2} + \rho High\_tech_{i,t-2} + \gamma_t \tag{3}$$

where $P(Active_{i,t}=1)$ is the probability that firm i is active in year t, and the logit, $\theta_{i,t}$, is regressed on the variable based on corporate website, $Web\_status_{i,t}$, the vector of economic quantitative variables $Z_{i,t-2}$ which includes $Size_{i,t-2}$, $Debt_{i,t-2}$, $Productivity_{i,t-2}$ and $Profitability_{i,t-2}$, the economic categorical variable $High\_tech_{i,t-2}$ and the fixed-effect of time, captured by γ_t.

Having confirmed the relation between the WWW and firms' status with both regressions, a duration model is applied to estimate the firm's probability of surviving one time period or more given the corporate website status.

Survival analysis

The relation of the firm's website status with its duration, the latter defined as the time elapsed (during the observed period) until a firm fails, was analyzed through survival models (also known as duration models (Lancaster, 1990)). These models are useful for predicting events like failures or deaths on a subject (e.g. firm, machine, system, product or patient). Specifically, time and other predictive variables are considered to estimate the hazard of failure or death during a particular time period.

In survival analysis, the hazard function $h(t)$ is the one used for conducting regressions. In this study, the hazard function was estimated through a cloglog generalized linear model, which is the equivalent to the discrete time version of the Cox Proportional Hazard Model (Jenkins, 1995). It has been successfully applied in previous firms' survival studies for data collected on an annual basis (Tsoukas, 2011; Görg and Spaliara, 2014; Guariglia et al., 2016), which is this case. The proportional hazard model assumes that the hazard rate depends only on the time at risk, $h_0(t)$ (the baseline hazard) and on the vector of explanatory variables, X. This is the rate at which firms die in year t, provided they survived the previous year, $t-1$. It is expressed as:

$$h(t, X) = h_0(t)\ \exp(\beta' X) \tag{4}$$

Particularly, the discrete-time hazard function (with period-specific effects) takes the following specification:

$$h(t, X) = 1 - \exp[-\exp(\beta' X + \gamma_t)] \tag{5}$$

where β' is the regression coefficient vector that describes how the hazard varies in response to explanatory vector X of the covariates, and γ_t captures the period-specific effects on the hazard.

For this study, this duration model was specified as follows:

$$h(t, Web\_status) = 1 - \exp[-\exp(\beta_0 + \alpha Web\_status + \gamma_t)] \tag{6}$$

where $h(t, Web\_status)$ is the hazard rate; that is, the rate at which firms become inactive at time t provided they were active in year $t-1$, which is modeled through the explanatory variable Web_status and the period-specific effect, γ_t.

Results

This section first shows some descriptive statistics and group comparisons to provide a data overview. Second, two multi-period logistic regressions are built and compared to evaluate to what extent the Web_status variable captures the company's activity status. Finally, these results are complemented with a duration model.

Descriptive statistics and group comparisons

The descriptive statistics of the whole dataset are reported in Table I. As we can see, the sample is dominated by active firms (96.9% of the sample) which operate in a low-technology sector (81%) and that have a moderate level of debt (61%). This table evidences the absence of high correlations among variables, which means that there was no high risk of information redundancy and multicollinearity when estimating the regression models.

The first column of Table II summarizes the behavior of corporate websites by showing the distribution of the Web_status variable across the sample. It indicates that most websites remained unchanged (37.7%) or underwent a moderate change (36.4%) compared with the previous year. Only 8.8% of the observations presented a down website, while 17.1% of them had totally changed. To illustrate the association with the other variables, the numeric value of Web_status is also included in Table I.

In order to test whether the variables behaved differently depending on the firm's status (active or inactive), statistical techniques of group differences were employed. Pearson's Chi-squared test was applied to the categorical variables, whose results are reported in Table II. Statistically significant differences were found for the technological intensity, being active firms more associated with technology-intensive sectors than those that became inactive (20.5% vs. 2.9%). For the website status, statistically significant differences were found between active and inactive firms. For the latter, most websites were down (50%) or remained unchanged (40.2%), while minor or major changes were found only in the remaining 9.8%.

– Insert Table I here –

In contrast, content changes were found in more than half of the active firms' websites, mainly minor changes (37.3%). This was expected because website design forms part of the firm's corporate image, which is not renewed yearly by most companies. Instead, minor changes to keep information up-to-date are frequently made by active firms. Moreover, down websites are not common among active companies (7.5%). The presence of unchanged websites (37.6%) was similar to the case of inactive firms, so this website status is not as indicative of the firm's status as the cases in which changes were made.

– Insert Table II here –

With the quantitative variables, normality and homogeneity of variance were checked both graphically and numerically. As none of the variables fulfilled both assumptions, the nonparametric Mann-Whitney U test was employed, which is based on the median (Anderson et al., 2014). These results are reflected in Table III. Most economic variables showed statistically different values for the active firms compared to the firms that had died during the observed period. The log number of employees was statistically higher for the active firms (3.761 vs. 3.401), so firm size relates to the firm's duration to some extent.

– Insert Table III here –

About the firm's age, no statistically significant differences were found, so firms seem to die with the same probability regardless of their age. The debt value was much higher for inactive firms (88.99% vs. 60.81%), which is indicative of the detrimental effect that high levels of debt have on firms' health, and thus on their continuity. Active firms were associated with statistically higher levels of productivity and profitability than inactive ones. High productivity levels are connected to overall better firm performance, which would contribute to having a higher profitability. Both these measures are related to the company's health so as expected, healthier companies continue with their activities more frequently.

Multi-period logistic regression models

In this section, we shed light on the role played by corporate websites status on firms' probability of being active. First, a multi-period logistic regression model based on the Web_status variable was built, as specified in equation (2).

Table IV provides the estimation results for this model, including the estimated regression coefficients (β), Odds Ratios (OR), Standard Errors (SE), z-values and p-values. The OR is a measure of the association between the different website statuses and the firm's status, and is calculated as the exponent of the coefficients. Thus, an OR over 1 indicates that the probability that a firm is active increases with a given independent variable (in this case, each particular website status). If it is lower than 1, it indicates that this probability decreases, while if it equals 1 then there is no association between the independent and dependent variable.

– Insert Table IV here –

For this web-based model, the results show that the observed web statuses have a statistically significant effect on the probability of a company being active. As website activity increases, the probability of a firm being active also increases. The estimate that corresponds to the website status 'Unchanged' (Code 2) is positive, which means that having a working website, even if its contents or look are not changed compared to the previous year, increases the probability of a firm being active with respect to having a down website (Code 1, which was taken as the baseline level). Indeed, the probability of a firm with an unchanged website being active is 5 times (or 409.4% higher) than that of a firm whose website is down, as indicated by the OR.

Updating websites to a minor (Code 3) or major (Code 4) extent increases the probability of a firm being active, as expected. Furthermore, the increase found is dramatically high in both cases. The probability of a firm being active when it moderately changes its website is 30-fold higher compared to a firm whose website is down, while it is more than 50-fold higher when a website has been completely renewed. These results are in line with what was hypothesized: healthy firms invest more in maintaining and updating their websites. Hence, the more activity they evidence on their website, the more likely they are active. It is noteworthy that this does not mean that updating websites helps firms remain active, but it strongly reflects firm's active status.

Once the relationship between the corporate websites' status and the firms' status was evidenced, the extended specification given by equation (3) was estimated. It included the website status variable and the structural variables selected for their potential relation to firms' survival, and for the significant variation across active and inactive firms.

As reported in Table V, the effect of each website status on the probability of being active remains positive, and statistically significant and high. Regarding the economic variables, only the firm's debt structure shows a statistically significant effect. Its negative coefficient indicates that as the amount of a firm's debt increases, its probability of being active decreases. Indeed, it decreases by 1.8% for each percentage point increase in debt.

– Insert Table V here –

Although the remaining economic variables showed differences in the univariate level, they do not help explain the firm's status at the multivariate level. On the one hand, such economic variables are related to a firm's status, but only to a limited extent as a large number of other factors, such as the firm's strategic decisions or specific market situations, can contribute to the death of firms with a wide range of characteristics (small or large, more or less productive, from any activity sector, etc.). On the other hand, website status has been revealed to be a clear indicator of a firm's status. So these economic variables were unable to complement the information displayed on the web.

Given that the relationship between corporate website status and firms' status was demonstrated, the next section went a step further to complement this analysis and to confer the study a different point of view. To do so, a survival analysis was conducted.

Survival analysis

This section describes the survival analysis conducted for modeling the hazard of a firm's death at certain times depending on website status. Since the data in this study were collected on an annual basis, a time-discrete duration model was built, as specified in equation (6).

Table VI offers the estimation results for this model, including the estimated regression coefficients (β), Hazard Ratios (HR), Standard Errors (SE), z-values and p-values. The hazard ratios, calculated as the exponent of coefficients, are a measure of how often an event happens in one group compared to how often it happens in another group over time. In this case, they measure how often the different website statuses happen in the groups of active and inactive firms. Thus an HR above 1 indicates that the hazard of death increases with the corresponding website status. If it is lower than 1, it means that this hazard decreases, while if it equals 1 then there is no difference in survival between the two groups being compared.

– Insert Table VI here –

The negative and statistically significant coefficient estimates indicate that the firms whose websites are unchanged (Code 2), or undergo minor (Code 3) or major changes (Code 4) compared to the previous year, are exposed to a significantly lower hazard than the firms whose corporate websites are down (Code 1, which is the baseline web status). Specifically, the hazard ratio for the 'Unchanged' website status (Code 2) indicates that the firms whose website contents are the same as the previous year have 0.301 times the hazard of death of the firms whose website is down; that is, their death hazard is 69.9% lower. The death hazard for firms which made minor changes in their websites (Code 3) is 91.6% lower than for those with down websites, while the percentage reaches 93.7% when the changes made are major (Code 4). As we can observe, the death hazard lowers at the same time as website activity increases. These results are consistent with those of the multi-period logistic regressions, and confirm the strong relationship between corporate website status and a firm's survival.

Conclusions

Business demography is a major area of interest for researchers and policy-makers because the creation and failure of companies have a huge impact on the production and employment in all the economies. In the current context, in which digital communications and Internet contents reflect society's main behavior, a new challenge arises: that of relating business demography with the WWW evolution.

This paper analyzed and confirmed the connection of a company's activity status to the corporate website's activity status. This was done by tracking corporate websites and statuses of firms for 7 years, and then analyzing their relation with logistic regressions and a survival model. Logistic regression estimates that major changes in the corporate website increase the odds of a firm being active by more than 50-fold compared to a down website. In terms of survival, corporate website changes are related to a predicted death hazard more than 90% lower. Since both methods gave similar results, this means that the corporate website is a robust indicator of a firm's activity status.

These results open up new possibilities to monitor business demography. Web data capture a firm's status, while access to corporate websites is open and inexpensive. This means that it is possible to build online indicators to nowcast and monitor business death rates. Unlike traditional official statistics methods, which rely on surveys done on a population sample that take time to be processed, monitoring the WWW could fast reach the entire population of companies with a website. This can be done in a very short period thanks to the digital nature of the WWW, which allows firms' information to be automatically retrieved and analyzed. This, in turn, allows policy-makers and other consumers of official statistics to obtain short-term estimates of the business demography, which would eventually turn into more informed decisions.

Among the limitations of this study, first, it is worth noting that only the homepage of the website was analyzed; that is, no changes in inner sections were taken into account. Second, the sample only includes companies based in Spain, so generalizing the results to different countries must be done cautiously. Finally, we must emphasize that we describe how website status correlates with a company's activity status, without causal analysis. Although this is useful for monitoring purposes, our results do not indicate that managers should continuously change corporate websites to increase company survival.

References

Agarwal, R. and Audretsch, D. B. (2001). "Does entry size matter? The impact of the life cycle and technology on firm survival". *Journal of Industrial Economics*, Vol. 49, pp. 21 – 43. DOI: 10.1111/1467-6451.00136.

Albino, V., Balice, A., and Dangelico, R. M. (2009). "Environmental strategies and green product development: An overview on sustainability-driven companies". *Business Strategy and the Environment*, Vol. 18, pp. 83 – 96. DOI: 10.1002/bse.638.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., and Cochran, J. J. (2014). *Statistics for Business & Economics*. Cengage Learning, 12th edition.

Arora, S. K., Li, Y., Youtie, J., and Shapira, P. (2016). "Using the wayback machine to mine websites in the social sciences: A methodological resource". *Journal of the Association for Information Science and Technology*, Vol. 67, pp. 1904 – 1915. DOI: 10.1002/asi.23503.

Arora, S. K., Youtie, J., Shapira, P., Gao, L., and Ma, T. (2013). "Entry strategies in an emerging technology: A pilot web-based study of graphene firms". *Scientometrics*, Vol. 95, pp. 1189 – 1207. DOI: 10.1007/s11192-013-0950-7.

Astebro, T. and Winter, J. (2012). "More than a dummy: The probability of failure, survival and acquisition of firms in financial distress". *European Management Review*, Vol. 9, pp. 1 – 17. DOI: 10.1111/j.1740-4762.2011.01024.x.

Audretsch, D. B. (1991). "New-firm survival and the technological regime". *The Review of Economics and Statistics*, Vol. 73, pp. 441 – 450. DOI: 10.2307/2109568.

Audretsch, D. B., Houweling, P., and Thurik, A. R. (2000). "Firm survival in the Netherlands". *Review of Industrial Organization*, Vol. 16, pp. 1 – 11. DOI: 10.1023/A:1007824501527.

Basile, R., Pittiglio, R., and Reganati, F. (2017). "Do agglomeration externalities affect firm survival?". *Regional Studies*, Vol. 51, pp. 548–562. DOI: 10.1080/00343404.2015.1114175.

Blazquez, D. and Domenech, J. (2014). "Inferring export orientation from corporate websites". *Applied Economics Letters*, Vol. 21, pp. 509 – 512. DOI: 10.1080/13504851.2013.872752.

Blazquez, D. and Domenech, J. (2017). "Web data mining for monitoring business export orientation". *Technological and Economic Development of Economy.* DOI: 10.3846/20294913.2016.1213193.

Bridges, S. and Guariglia, A. (2008). "Financial constraints, global engagement, and firm survival in the United Kingdom: Evidence from micro data". *Scottish Journal of Political Economy*, Vol. 55, pp. 444 – 464. DOI: 10.1111/j.1467-9485.2008.00461.x.

Bureau van Dijk (2010). "SABI: Sistema de análisis de balances ibéricos.". CD-ROM (Version 36.1).

Cefis, E. and Marsili, O. (2005). "A matter of life and death: Innovation and firm survival". *Industrial and Corporate Change*, Vol. 14, pp. 1167 – 1192. DOI: 10.1093/icc/dth081.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge, 3rd edition.

Delmar, F., McKelvie, A., and Wennberg, K. (2013). "Untangling the relationships among growth, profitability and survival in new firms". *Technovation*, Vol. 33, pp. 276 – 291. DOI: 10.1016/j.technovation.2013.02.003.

DGIPYME (2017). "Estadísticas PYME: Evolución e indicadores". Available at: http://www.ipyme.org/Publicaciones/Estadisticas-PYME-2016.pdf (accessed March 30, 2017).

Domenech, J., de la Ossa, B., Pont, A., Gil, J. A., Martinez, M., and Rubio, A. (2012). "An intelligent system for retrieving economic information from corporate websites". In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 573 – 578, Macau, China.

Esteve-Pérez, S. and Mañez-Castillejo, J. A. (2008). "The resource-based theory of the firm and firm survival". *Small Business Economics*, Vol. 30, pp. 231 – 249. DOI: 10.1007/s11187-006-9011-4.

Eurostat (2008). *NACE Rev. 2 Statistical classification of economic activities in the European Communities*. EUROSTAT Methodologies and Working papers. Office for Official Publications of the European Communities, Luxembourg.

Eurostat (2014). *High-tech aggregation by NACE Rev. 2*. Eurostat indicators of High-tech industry and knowledge - intensive services. Office for Official Publications of the European Communities, Luxembourg.

Eurostat and OECD (2007). *Eurostat-OECD Manual on Business Demography Statistics.* Office for Official Publications of the European Communities, Luxembourg.

Evans, D. S. (1987). "Tests of alternative theories of firm growth". *Journal of Political Economy*, Vol. 95, pp. 657–674.

Gallego Álvarez, I., María García Sánchez, I., and Rodríguez Domínguez, L. (2008). "Voluntary and compulsory information disclosed online. the effect of industry concentration and other explanatory factors.". *Online Information Review*, Vol. 32, pp. 596 – 622. DOI: 10.1108/14684520810913990.

Gémar, G., Moniche, L., and Morales, A. J. (2016). "Survival analysis of the Spanish hotel industry". *Tourism Management*, Vol. 54, pp. 428 – 438. DOI: 10.1016/j.tourman.2015.12.012.

Geroski, P. (1995). "What do we know about entry?". *International Journal of Industrial Organization*, Vol. 13, pp. 421 – 440. DOI: 10.1016/0167-7187(95)00498-X.

Geroski, P., Mata, J., and Portugal, P. (2010). "Founding conditions and the survival of new firms". *Strategic Management Journal*, Vol. 31, pp. 510 – 529. DOI: 10.1002/smj.823.

Giovannetti, G., Ricchiuti, G., and Velucchi, M. (2011). "Size, innovation and internationalization: A survival analysis of italian firms". *Applied Economics*, Vol. 43, pp. 1511 – 1520. DOI: 10.1080/00036840802600566.

Gök, A., Waterworth, A., and Shapira, P. (2015). "Use of web mining in studying innovation". *Scientometrics*, Vol. 102, pp. 653 – 671. DOI: 10.1007/s11192-014-1434-0.

Görg, H. and Spaliara, M.-E. (2014). "Financial health, exports and firm survival: Evidence from UK and French firms". *Economica*, Vol. 81, pp. 419 – 444. DOI: 10.1111/ecca.12080.

Guariglia, A., Spaliara, M.-E., and Tsoukas, S. (2016). "To what extent does the interest burden affect firm survival? Evidence from a panel of UK firms during the recent financial crisis". *Oxford Bulletin of Economics and Statistics*, Vol. 78, pp. 576 – 594. DOI: 10.1111/obes.12120.

INE (2016). "Encuesta de uso de TIC y Comercio Electrónico en las empresas 2015-2016". Available at: http://ine.es/dynt3/inebase/?path=/t09/e02/a2015-2016 (accessed October 10, 2016).

Jacobson, T. and von Schedvin, E. (2015). "Trade credit and the propagation of corporate failure: An empirical analysis". *Econometrica*, Vol. 83, pp. 1315 – 1371. DOI: 10.3982/ECTA12148.

Jenkins, S. P. (1995). "Easy estimation methods for discrete-time duration models". *Oxford Bulletin of Economics and Statistics*, Vol. 57, pp. 129 – 136. DOI: 10.1111/j.1468-0084.1995.tb00031.x.

Kahle, B. and Gilliat, B. (2016). "Wayback machine". Available at: http://archive.org/web/ (accessed March 25, 2016).

Lancaster, T. (1990). *The econometric analysis of transition data.* Cambridge University Press.

Li, Y., Arora, S., Youtie, J., and Shapira, P. (In press). “Using web mining to explore Triple Helix influences on growth in small and mid-size firms”. *Technovation.* DOI: 10.1016/j.technovation.2016.01.002.

Libaers, D., Hicks, D., and Porter, A. L. (2010). “A taxonomy of small firm technology commercialization”. *Industrial and Corporate Change*, Vol. 25, pp. 371 – 405. DOI: 10.1093/icc/dtq039.

Manjón-Antolín, M. C. and Arauzo-Carod, J.-M. (2008). “Firm survival: methods and evidence”. *Empirica*, Vol. 35, pp. 1 – 24. DOI: 10.1007/s10663-007-9048-x.

Mata, J. and Portugal, P. (1994). “Life duration of new firms”. *The Journal of Industrial Economics*, Vol. 42, pp. 227 – 245. DOI: 10.2307/2950567.

Meroño-Cerdan, A. L. and Soto-Acosta, P. (2007). “External web content and its influence on organizational performance”. *European Journal of Information Systems*, Vol. 16, pp. 66 – 80. DOI: 10.1057/palgrave.ejis.3000656.

Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* John Wiley & Sons, Ltd, Chichester, UK.

Overbeeke, M. and Snizek, W. E. (2005). “Web sites and corporate culture: A research note”. *Business & Society*, Vol. 44, pp. 346 – 356. DOI: 10.1177/0007650305275748.

Tagesson, T., Blank, V., Broberg, P., and Collin, S.-O. (2009). “What explains the extent and content of social and environmental disclosures on corporate websites: A study of social and environmental reporting in Swedish listed corporations". *Corporate Social Responsibility and Environmental Management*, Vol. 16, pp. 352 – 364. DOI: 10.1002/csr.194.

Tang, L., Gallagher, C. C., and Bie, B. (2015). "Corporate social responsibility communication through corporate websites: A comparison of leading corporations in the United States and China". *International Journal of Business Communication*, Vol. 52, pp. 205 – 227. DOI: 10.1177/2329488414525443.

Tsoukas, S. (2011). "Firm survival and financial development: Evidence from a panel of emerging asian economies". *Journal of Banking & Finance*, Vol. 35, pp. 1736 – 1752. DOI: 10.1016/j.jbankfin.2010.12.008.

Vaughan, L. (2014). "Discovering business information from search engine query data". *Online Information Review*, Vol. 38, pp. 562–574. DOI: 10.1108/OIR-08-2013-0190.

Youtie, J., Hicks, D., Shapira, P., and Horsley, T. (2012). "Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies". *Technology Analysis & Strategic Management*, Vol. 24, pp. 981 – 995. DOI: 10.1080/09537325.2012.724163.

Notes

[1] From the total sample of 780 firms, 92% were small and medium-sized (SMEs), in line with the prevailing productive structure in Spain (DGIPYME, 2017).

[2] Statistical Classification of Economic Activities in the European Community (Eurostat, 2008)

[3] BORME, from their initials in Spanish, 'Boletín Oficial del Registro Mercantil'.

[4] We considered inactive the following firm status: in extinction; in dissolution; in liquidation; in a finished receivership where dissolution or liquidation has been ordered, but is not yet done; or in receivership in progress (if it is the most recent status and no additional information is available), except when the firm had been taken over or had merged (Eurostat and OECD, 2007).

Table I: Global descriptive statistics and correlation matrix

Variable	Mean	SD	1	2	3	4	5	6	7
1. *Active*	0.969	0.174							
2. *Web_status*	2.618	0.868	0.272						
3. *Size*	3.855	1.241	0.046	0.149					
4. *Age*	24.823	14.318	0.007	0.022	0.258				
5. *Debt*	60.470	31.608	−0.127	−0.009	0.016	−0.140			
6. *Productivity*	0.270	4.071	0.007	0.021	−0.129	−0.042	0.046		
7. *Profitability*	1.099	22.526	0.112	0.046	−0.005	0.005	−0.291	0.044	
8. *High_tech*	0.199	0.399	0.076	0.143	0.207	−0.062	−0.048	−0.021	−0.005

Procedures employed: Pearson's r coefficient for pairs of continuous variables; Point-biserial coefficient for pairs of a continuous and a binary variable; Phi coefficient for pairs of binary variables; and Eta for pairs of a continuous and a categorical variable with more than two levels (Cohen et al., 2002)

Table II: Descriptive statistics of qualitative variables and group comparisons

	All (N=3254)	*Active*(0) (N=102)	*Active*(1) (N=3152)	Chi-squared (*p*-value)
Active(0)	3.1%			
Active(1)	96.9%			
High_tech(0)	80.1%	97.1%	79.5%	
High_tech(1)	19.9%	2.9%	20.5%	0.000
Web_status(1)	8.8%	50.0%	7.5%	
Web_status(2)	37.7%	40.2%	37.6%	
Web_status(3)	36.4%	7.8%	37.3%	
Web_status(4)	17.1%	2.0%	17.6%	0.000

Notes: *Web_status*(1): Down; *Web_status*(2): Unchanged; *Web_status*(3): Minor change; *Web_status*(4): Major change.

Table III: Descriptive statistics of quantitative variables and group comparisons

	$Active(0)$	$Active(1)$	Mann-Whitney U
	(N=102)	(N=3152)	(p-value)
$Size$	3.761	3.401	0.007
Age	21.501	22.815	0.875
$Debt$	88.990	60.810	0.000
$Productivity$	30.947	48.787	0.000
$Profitability$	-3.775	2.000	0.000

Table IV: Multi-period logistic regression with web status. Dependent variable: $Active$

Variables	β	OR	SE	z-value	p-value
(Intercept)	3.483	32.557	0.598	5.827	0.000
$Web\_status(2)$	1.628	5.094	0.227	7.158	0.000
$Web\_status(3)$	3.423	30.661	0.390	8.786	0.000
$Web\_status(4)$	3.970	52.985	0.727	5.460	0.000
Observations	$3,254$				
Log-likelihood	-349.349				

Notes: $Web\_status(2)$: Unchanged; $Web\_status(3)$: Minor change; $Web\_status(4)$: Major change. Time dummies were included.

Table V: Multi-period logistic regression with web and structural variables. Dependent variable: *Active*

Variables	β	OR	SE	z-value	p-value
(Intercept)	3.917	50.249	0.974	4.023	0.000
$Web\_status(2)$	1.579	4.850	0.395	4.001	0.000
$Web\_status(3)$	2.490	12.061	0.522	4.768	0.000
$Web\_status(4)$	3.242	25.585	0.951	3.410	0.001
$Size$	0.206	1.229	0.169	1.218	0.223
$Debt$	-0.018	0.982	0.006	-3.053	0.002
$Productivity$	0.034	0.193	1.034	0.283	0.777
$Profitability$	0.005	1.005	0.011	0.443	0.658
$High\_tech$	10.811	$49,563.01$	960.4	0.019	0.985
Observations	$3,034$				
Log-likelihood	-154.116				

Notes: $Web\_status(2)$: Unchanged; $Web\_status(3)$: Minor change; $Web\_status(4)$: Major change. Time dummies were included.

Table VI: Discrete-time duration model. Dependent variable: $1 - \mathit{Active}$

Variables	β	HR	SE	z-value	p-value
(Intercept)	-4.942	0.007	1.029	-4.802	0.000
$Web\_status(2)$	-1.202	0.301	0.348	-3.454	0.001
$Web\_status(3)$	-2.480	0.084	0.484	-5.128	0.000
$Web\_status(4)$	-2.764	0.063	0.753	-3.668	0.000
Observations	$3,194$				
Log-likelihood	-195.262				

Notes: $Web\_status(2)$: Unchanged; $Web\_status(3)$: Minor change; $Web\_status(4)$: Major change. Time dummies were included.