Technological Forecasting & Social Change 207 (2024) 123597

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Scraping innovativeness from corporate websites: Empirical evidence on Italian manufacturing SMEs

Carlo Bottai [a,b,*], Lisa Crosato [c,b], Josep Domenech [d,b], Marco Guerzoni [a,e], Caterina Liberati [a,b]

[a] Department of Economics, Management and Statistics; University of Milano-Bicocca, P.zza dell'Ateneo Nuovo, 1, Milano, 20126, Italy
[b] Center for European Studies; University of Milano-Bicocca, P.zza dell'Ateneo Nuovo, 1, Milano, 20126, Italy
[c] Department of Economics; Ca' Foscari University of Venice, Cannaregio, 873, Venezia, 30121, Italy
[d] Department of Economics and Social Sciences, Universitat Politècnica de València, Camí de Vera, s/n, València, 46022, Spain
[e] BETA, Bureau of Theoretical and Applied Economics; University of Strasbourg, Avenue de la Forêt Noire, 61, Strasbourg, 67085, France

ARTICLE INFO

JEL classification:
O3
Q55
Q58

Keywords:
Unconventional data
SMEs
Innovation
Corporate websites
HTML tags

ABSTRACT

Research in innovation studies usually relies on financial statements, surveys, or patents as primary data sources, although these sources of information show some limitations when applied to Small and Medium Enterprises (SMEs). Our paper explores whether the HTML code of a company's website is a further source to better inform innovation policies, under the assumption that how HTML is employed in crafting a corporate website provides insights into the company's innovation capabilities. In particular, we leverage HTML tags and their associations to empirically show that the websites of innovative SMEs are different from non-innovative ones both in terms of their size and coding practices. Our findings, based on a sample of Italian companies, indicate that the features of the HTML code of corporate websites reflect unobservable characteristics related to the skills and creativity present in businesses.

1. Introduction

Assessing the presence and intensity of innovative activity within a firm is a complex task, due in part to the multifaceted nature of innovation and, mainly, to the fact that innovativeness is often hidden within the practices of a firm and the shared knowledge of its employees. That innovation measurement is a major obstacle to understanding the economic role of technological change has been known at least for seventy years now (Kuznets, 1962). Over the past half-century, operational research in innovation studies has made significant progress in developing various methods and indicators to measure this phenomenon. To do so, financial statements, surveys and patents have been used as the main data sources. However, each of these sources has limitations in capturing such latent features, particularly when applied to Small and Medium-sized Enterprises (SMEs) that, instead, can be considered the backbone of national economies (OECD, 2015; OECD and Eurostat, 2018).

Moreover, building innovation measures based on these traditional data sources has important limitations concerning their capacity to be constantly updated. First and foremost, this is because the sources of data on which eventual indicators are built are structurally anchored to the past. For example, financial statements are available only at the end of the year and tend to be released to researchers with some additional delay by the data providers.

As a solution, a growing body of literature proposes to use textual features scraped from corporate websites to measure innovation within firms, as recently reviewed by Rammer and Es-Sadki (2023). This article fits within this ongoing debate by discussing the findings and limitations of the literature, arguing that, instead of solely relying on the textual content of a corporate website, also its HTML structure can provide useful insights into firms' innovativeness. Firstly, we highlight theoretical reasons why the HTML structure of the innovative SMEs' websites should be different and, secondly, we test this hypothesis empirically.

Sampling from the population of Italian manufacturing SMEs active in 2016, we provide several tag-based aggregate statistics that describe the size of the website from various perspectives. In addition, we investigate whether a natural clustering of HTML tags emerges from the data, indicating different coding styles. The empirical exercise leverages the list of innovative SMEs provided by the Italian Ministry of Economic Development. Thus, we can test for differences between

* Corresponding author at: Department of Economics, Management and Statistics; University of Milano-Bicocca, P.zza dell'Ateneo Nuovo, 1, Milano, 20126, Italy.
E-mail addresses: carlo.bottai@unimib.it (C. Bottai), lisa.crosato@unive.it (L. Crosato), jdomenech@upvnet.upv.es (J. Domenech), marco.guerzoni@unimib.it (M. Guerzoni), caterina.liberati@unimib.it (C. Liberati).

https://doi.org/10.1016/j.techfore.2024.123597
Received 7 February 2024; Received in revised form 18 May 2024; Accepted 13 July 2024
Available online 5 August 2024

innovative and non-innovative SMEs concerning the use of specific clusters of tags. A matching procedure allows us to control for possible confounding effects, such as regional or industrial selection bias. The results indicate that innovative SMEs have more up-to-date, larger, richer and better-organized corporate websites compared to their non-innovative counterparts. Hence, the main contribution of this paper is to provide evidence that the HTML structure of a website brings information useful to identify innovative activity within SMEs. Even though scholars mostly overlooked the HTML code of corporate websites so far, future contributions need to consider it more carefully as a useful add-on to the natural language-based indicators proposed in recent years. We also devote particular care to assess the quality of the data used for the analysis, which guarantees its reliability and accuracy.

The paper runs as follows. After recalling some limitations of using conventional data sources for measuring SMEs' innovative activities, we review the ongoing literature on exploiting website textual content as a possible solution and its limits in Section 2. Moving to our contribution, Section 3 proposes theoretical arguments on using HTML code to overcome the limitation in the analysis of the textual content of a web-page. Here, we put forward the hypothesis that the HTML structure of a web-page can be exploited to infer some unobservable characteristics of a firm's innovative capabilities. Section 4 presents the data collection and processing, discussing in detail how we employ the HTML code in the analysis. As the primary contribution of the paper, in Section 5, we show for the first time that there is statistically significant variation in the structure of HTML code between innovative and non-innovative SMEs. In this context, we demonstrate that, as theoretically hypothesized, the HTML code of a corporate website contains an additional information signal, beyond the variables typically used by scholars, that enables to differentiate between innovative and non-innovative SMEs.

2. Measuring innovation with conventional and unconventional data

The data most widely used to trace innovative processes comes from specific surveys, financial data, or patents (Gault, 2013; OECD, 2015; OECD and Eurostat, 2018; Hall et al., 2010; Nagaoka et al., 2010; Mairesse and Mohnen, 2010). Although these sources are characterized by high-quality standards, they suffer from some limitations that cannot be neglected, especially in the case of SMEs.

Surveys are in principle designed to collect any type of information, but responses may be distorted and questions ill-interpreted by the respondent. For example, the EU Community Innovation Survey (CIS) targets a large pool of European firms with a battery of questions dedicated to innovation processes (Arundel and Smith, 2013). However, small businesses are only present on a rotating-sample basis, making it difficult to conduct longitudinal studies of such companies. Furthermore, micro-enterprises with less than ten employees are not surveyed at all. So, the CIS is not well-equipped at capturing micro-firms' innovativeness. Additionally, other surveys designed to explore specific dimensions of innovation are often conducted on smaller sample sizes (Arundel et al., 2013).

Financial statements are another example of official data heavily used in innovation studies, being available for the majority of companies. They report information on innovative activities such as R&D expenses and allow us to infer measures on firms' productivity, profitability and growth. However, for SMEs, many R&D expenses occur informally and are hidden in personnel costs, rather than recorded in financial statements as explicit R&D costs (Santarelli and Sterlacchini, 1990). Other registry data, such as labour contracts, are difficult to interpret, and their availability depends greatly on the data-providers (Kitchin, 2014).

Patents suffer from well-known limitations, such as a different propensity to patent between and within sectors; the existence of non-patentable technological knowledge; and the fact that patents protect inventive steps that do not necessarily represent innovations (Fontana et al., 2013). Additionally, patents can be filed or purchased for strategic intellectual property (IP) purposes, not always tracing new knowledge production (Cohen et al., 2000) and SMEs often lack organizational and economic resources to patent and systematically prefer other forms of IP protection (Holgersson, 2013). Therefore, patents might under-represent SMEs' contribution to the invention process.

Finally, all of these data sources collect information not promptly updated. Surveys require time for construction, collection and quality processing. Financial accounts are usually available for statistical purposes with an 18–24 month delay. Patents are filed at the end of the inventive process and granted after an administrative processing period that can last over 12 months.

These limitations pushed the literature towards the search for alternative or complementary sources of information. Recently, new streams of literature within innovation studies pointed the way towards augmenting these traditional data sources to better understand firms' innovativeness and the innovation process more in general. As extensively reviewed by Antons et al. (2020) and Rammer and Es-Sadki (2023), a sizable innovation studies literature started exploiting, through text data mining techniques, large textual documents where information about the innovative activity of firms can be extracted. In doing so, the web is a priceless source of this kind of document, so scholars have been exploring it as an additional information source. Among the first to embrace this challenge, Nathan and Rosso (2015, 2022) use Reuters and Yahoo! online news sources to capture respectively digital business and product launches as proxy for the innovative performance.

Firms utilize their websites as virtual showcases to display products, disseminate information about their operations, and establish a public image. Accordingly, the content of a corporate website is closely tied to the economic activities of its owner, serving purposes such as spreading information, conducting online transactions, and facilitating customer opinion sharing (Domenech et al., 2012; Blazquez and Domenech, 2018). Consequently, corporate websites emerge as rich sources of information for observing the economic behaviour of firms, prompting a growing body of literature to accept the challenge of mining them for research purposes, including within the field of innovation studies. Unfortunately, the usage of such information is not straightforward.

For instance, Web-scraped data for several UK SMEs were proven by Gök et al. (2015) to offer additional insights compared to traditional sources, such as patents and publications and, closely related, Arora et al. (2013) and Shapira et al. (2016) underlined the advantages of using data scraped from corporate websites to learn innovation behaviour and commercialization strategies of SMEs in emerging technology markets. Guzman and Li (2023) used the text extracted from the corporate websites for a sample of more than 12,000 US startups to proxy their strategic differentiation as the textual distance from similar incumbent public firms.

The occurrence of keywords on corporate websites was used to build a taxonomy of business models employed by small, highly-innovative firms focused on technology commercialization (Libaers et al., 2016), and to develop indicators for different core concepts of the innovation process. Text extracted from corporate websites helped identify product-innovator firms, with reliable, cost-effective results, with a high coverage and spatial granularity (Daas and van der Doef, 2020; Kinne and Lenz, 2021; Axenbeck and Breithaupt, 2021). Similarly, Ashouri et al. (2022) developed a web-scraping platform to identify the presence of product innovation and R&D orientation of firms.

Li et al. (2018) used data based on the corporate websites of various US SMEs applied to Triple Helix framework, while, on the same data, Arora et al. (2020) estimated, through topic modelling techniques, the dynamic capabilities of these firms.

As well, data scraped from corporate websites have been shown useful to investigate the working mechanisms of innovation systems, focusing not only on the textual content of these documents but also on

the hyperlinks between the web-documents collected (Katz and Cothey, 2006; Kinne and Axenbeck, 2020; Abbasiharofteh et al., 2023).

The analysis of web-scraped data proved to be useful in overcoming some limits of conventional data sources, although the usage of these alternative data sources is still circumscribed, primarily due to statistical modelling challenges. As Rammer and Es-Sadki (2023, 2–6) highlight, at least three peculiarities of these kinds of unconventional data sources must be carefully accounted for. First, they are produced by firms for business purposes and the way we measure innovativeness must consider this. Second, the information on innovation is provided in an unstructured way and we must properly analyse the data to extract the information sought. Third, we must employ technical solutions able to deal with big data sources that grow or change over time.

In some cases, the keywords' update invalidated prediction results (Janardan Mehta, 2017) but thus far, the concept drift[1] problem does not have an effective solution for improving text-based models' robustness and stability over time (Daas and van der Doef, 2020).

Also, when dealing with the textual content of a website, other drawbacks must be taken into account. First, text is content- and language-specific, making it challenging to extend results beyond a country's borders and draw meaningful international comparisons.[2] Second, textual content depends on the technology and knowledge base of a sector. Thus, comparisons across sectors are challenging, even between firms in the same supply chain. Third, the textual content also depends on the type of innovation it refers to, since it tends to capture the introduction of new products and their quality rather than process innovations or inventions (Gök et al., 2015). Fourth, textual content is produced potentially with the aim to present information in a way that is the most favourable to the company itself, by omitting or twisting some details, as pointed out by Gök et al. (2015).

While employing HTML code may not resolve all the previously mentioned limitations, it can significantly alleviate them, especially when used in conjunction with textual analysis. For example, HTML transcends language and sector-specific semantics. It is not directly linked to the multifaceted concept of innovation but serves as a proxy for the underlying competencies at the firm level, as will be detailed in the following section. The reliance on HTML is based on technical considerations and avoids the potential bias inherent in corporate representations of the world.

Specifically, we hypothesize that innovative SMEs adopt a distinctive coding style in the development of their websites. Therefore, we argue that the HTML code of a website carries useful information about the innovation activity of a small business.

3. Why innovative SMEs code differently

The use of HTML code, as proposed in this paper, while maintaining the general limitation and opportunities of unconventional web-scraped data, solves the issue of content analysis since it is not specific to any particular language, industry, or technology. Moreover, computational linguistics recently recognized that HTML code improves large language models performance in tasks like named entity recognition, description generation and autonomous web navigation (Ashby and Weir, 2020; Li et al., 2022; Gür et al., 2023). However, it still remains challenging to understand why the HTML code of innovative companies, rather than the textual content, would differ from that of non-innovative ones. Before delving into the data to demonstrate this phenomenon, we provide some theoretical explanations.

First, being oriented towards the commercialization of new products and technologies, we may expect that an innovative SME needs its website to be well-indexed by search engines and social networks since this will increase its visibility and sales. For this reason, commercial practices like e-commerce, customer engagements and user monitoring require to be embedded in the HTML design, as documented since the seminal work of van Duyne and Hong (2003) on the customer-centred web experience.

Second, innovative SMEs tend to employ more highly skilled workers, with advanced technological expertise, than the average small business. Evidence that they facilitate ICT adoption is vast and robust (Haller and Siedschlag, 2011; Giotopoulos et al., 2017, among the many), so innovative SMEs will be more receptive to adopting new technologies and incorporating them into their operations. Therefore, we expect that the presence of high-skilled workers facilitates the integration of cutting-edge technologies into the development process of the website of innovative SMEs. Thus, we expect that outdated or deprecated tags are less likely to be observed for innovative firms.

Third, we surmise that there is a positive relationship between the quality of a company's website and its overall dynamic capabilities. In the rapidly changing internet economies, only companies that excel in website design, functionality, user experience and internet strategy can truly thrive. While a general correlation between dynamic capabilities and innovation capabilities has been established since Lawson and Samson (2001)'s seminal work, there is also compelling and specific evidence within the context of the internet economy. For instance, Daniel and Wilson (2003) demonstrate that firms capable of meeting the demands of the e-business environment are associated with a higher rate of innovation. In this vein, we conjecture that HTML tags linked with functionalities of the internet economies might be more frequent in innovative firms.

Eventually, we believe that the relationship between website style and innovation activity persists even when firms do not develop in-house their website but acquire it, as-a-service, from a third party. In this case, the website can be seen as the design outcome of a user-producer interaction and, as well-established, the innovativeness of the output in user-producer interaction heavily depends, ceteris paribus, on users' input specificity and quality (Lundvall and Johnson, 1994). Indeed, the capability of the users to provide precise information about their needs and their level of knowledge to effectively communicate with the producer have a significant effect on the outcome of the interaction (Guerzoni, 2010). In a nutshell, the kind of HTML code and coding style that is used to create a corporate website reflects the interaction of the company's needs and skills with those of the programmer (Brinck et al., 2001). Hence, we hypothesize that the outcome of this interaction reveals unobservable characteristics related to high skills and creativity that may be indicative of an overall degree of innovativeness of a small business.

To the best of our knowledge, this is the first paper to suggest and test the hypothesis that the HTML code of the corporate website of innovative SMEs is systematically different from that of non-innovative ones.

4. Data collection

This study relies on a sample of innovative and non-innovative Italian SMEs in manufacturing, active in 2016. We specifically focus on the year 2016 as a preliminary step towards conducting a dynamic analysis that will encompass a longer time period up to the present day. We decided not to extend the analysis before 2016, as the number of registered innovative SMEs during that period was extremely small. As well, we choose to restrict the study to the manufacturing sector for two main reasons. Firstly, the manufacturing sector is a crucial driver of economic growth and innovation and encompasses a wide range

[1] In data mining literature, concept drift or drift indicates the unforeseeable changes in the underlying distribution of the data over time (Lu et al., 2019).

[2] Cross-language NLP models might partially solve this problem, as pointed out by the detailed survey by Pikuliak et al. (2021). However, these models may still struggle with nuances, idiomatic expressions, and cultural context (Hershcovich et al., 2022). Additionally, such models require substantial computational resources and may not be as accurate in all languages, especially those less commonly represented in the training data.

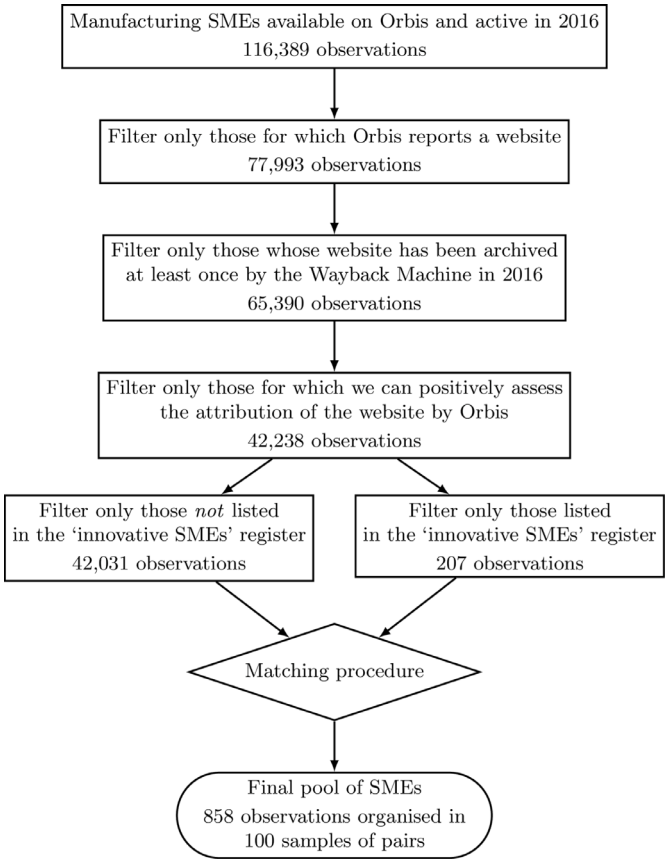

Fig. 1. Sample size over the different steps of the cleaning and matching process.

of industries involved in the production of tangible goods. Secondly, the manufacturing sector typically exhibits unique characteristics compared to other sectors. It often involves complex production processes, technological advancements and research and development activities.

The data collection combined conventional and unconventional data sources. In detail, three information sources have been employed: the Orbis and Aida databases by Bureau van Dijk (BvD) and the Wayback Machine of the Internet Archive (https://web.archive.org/). The two former databases, which are conventional data sources, were utilized for company identification, determination of innovation status, information on size, location and sector, and validation of website data. Website data have been retrieved from the Wayback Machine, an unconventional data source that has archived approximately 682 billion web-pages to date, covering over 25 years of web history. As supported by the existing literature, this resource enables users to track the history and evolution of each web-page over the very time period of the corresponding conventional data. See Arora et al. (2016) and Blazquez et al. (2018) for the advantages and limitations of this approach.[3]

4.1. Creation and validation of the SMEs websites' sample

The list of Italian manufacturing SMEs active in 2016 was obtained via a specific query on the BvD-Orbis platform. For each of the records of the initial screening, we retrieved, among others, the corporate website URL, if provided. Therefore, the process of companies' sample selection first counted 116,389 firms, which dropped to 77,993 observations because of missing URLs (Fig. 1). For the latter companies, we accessed their website, when present, as archived in 2016 by the Wayback Machine. This means that the monitored companies were further reduced to 65,390.

Unfortunately, the websites provided by BvD were not always associated with the right companies, so we performed a verifying procedure to assess the correct matching between URLs and companies. Similar to Barcaroli et al. (2016), the procedure searches for several information pieces on the archived website, such as the firm's tax identification number, address (street name and number) or the postcode of its head office and phone number, keeping the site only if the correspondence is verified. This has improved the data quality, reducing mismatching errors. Indeed, ex-post manual inspection proves an accuracy of 94% in associating the company with the correct website URL, as described in details in Bottai et al. (2022). On the other hand, the verification procedure impacted the final size of the sample, reducing it to 42,238 SMEs.

4.2. Labelling of innovative SMEs and data organization

The list of innovative SMEs provided by BvD-Aida is created by the Italian Chamber of Commerce's Business Register in compliance with the Italian Startup Act (221/2012 law). In 2012, the Italian Startup Act created a specific section in the Italian business register for classifying 'innovative startups' while, a few years later, the Law-Decree DL 3/2015 expanded this possibility to 'innovative SMEs'. Small businesses must meet a set of criteria to enrol on this special section of the business register. Specifically, they must not distribute profits and must develop, produce and commercialize innovative goods or services of high technological value. Additionally, they must meet at least two of the following three criteria: (a) allocate at least 15% of expenses to R&D; (b) employ Ph.D. students or holders, researchers, or Master's degree holders comprising at least one-third or two-thirds of the workforce, respectively; (c) hold, have deposited, or have in license a registered patent or own a legally registered computer programme. Firms are classified as 'innovative startups' if they have been in operation for less than five years and as 'innovative SMEs' otherwise.[4] As underlined by many recent works, the Italian Startup Act provides us with a novel tool to identify innovative SMEs, with at least three clear advantages over previous 'innovativeness' indicators (see Guerzoni et al., 2021; Antonietti and Gambarotto, 2020; Colombelli, 2016, among others). First, it focuses on SMEs, which very likely are not subsidiaries or foreign green-field entrants. Second, the included firms must focus on novel products. Lastly, at least one of the usual (input or output) innovation proxies must be fulfilled by the included firms; but, differently from other measures, it is not *a priori* restricted to just one of them. Table 1 reports the summary statistics for a few indicators available on the collected SMEs. As expected, innovative firms are larger, particularly if we look at the value of their intangible assets (patents and IPRs included) and their R&D expenses level.

The size, industrial sector and location of firms are widely recognized as significant variables that can affect both input and output indicators of innovative activity (Guerzoni et al., 2021). Therefore, any comparison of the difference between innovative and not-innovative firms must account for these confounding factors. In this case, the distributions of the innovative and non-innovative SMEs (Fig. 2) show different patterns, by firm size, industry and geographical location. To perform a thorough analysis of the data we have designed 100 groups of firm pairs. Using an exact matching procedure, we paired each innovative company with a non-innovative company that was

[3] The non-expert reader can find a concise description of the key components of the HTML code in Appendix A. This description will serve as a helpful reference for the upcoming analysis.

[4] For the complete list of criteria, see Decree 179/2012 Art. 25 and Decree 3/2015 Art. 25 and its modifications.

Table 1
Descriptive statistics for the 42,238 manufacturing SMEs active in 2016.

	Innovative			
	min.	mean	max.	st. dev.
Employees (num.)	1.00	23.36	211.00	33.34
Tot. Assets (th EUR)	28.98	5,020.92	51,810.47	7,977.80
Fixed assets (th EUR)	0.40	1,793.17	19,935.46	3,197.78
Fixed intangible assets (th EUR)	0.00	547.73	8,432.78	1,158.98
Patents and IPRs (th EUR)	0.00	88.78	2,394.97	324.03
R&D expenses (th EUR)	0.00	381.22	8,201.22	1,056.20
	Non-innovative			
	min.	mean	max.	st. dev.
Employees (num.)	0.00	21.16	244.00	26.15
Tot. Assets (th EUR)	9.83	4,986.70	883,522.84	9,821.78
Fixed assets (th EUR)	0.00	1,586.91	596,630.82	5,334.69
Fixed intangible assets (th EUR)	0.00	148.59	119,758.07	1,138.42
Patents and IPRs (th EUR)	0.00	14.26	15,382.77	159.51
R&D expenses (th EUR)	0.00	26.79	7,068.00	169.02

(i) of a similar size; (ii) located in the same geographical region (NUTS 2); and (iii) operating in the same industry (NACE, 3-digit). After excluding innovative SMEs without any potential matching, we maximized the number of couples and the number of linked firms of both kinds preserved with the Hopcroft–Karp algorithm (Hopcroft and Karp, 1973). The matching procedure has been repeated 100 times, using each time a different pool of non-innovative SMEs to have several groups to compare.[5]

Thus, we have obtained 178 innovative SMEs and 680 non-innovative matches, organized in 100 paired-firms samples.

4.3. Building web-based data

Web-based data were collected using the method described in Blazquez et al. (2018) and Crosato et al. (2021, 2023). We accessed the homepage of each company's website and downloaded its HTML code. From this code, we identified and recorded the different HTML tags and how frequently they appeared on each website. To ensure the quality of the data, we filtered out web-pages with less than ten HTML tags, as these are likely to be low-quality snapshots by the Wayback Machine. The resulting dataset is in the form of a document-term matrix, where rows represent websites, columns represent HTML tags and each cell contains the absolute frequency of a tag on a specific website. Overall, we identified a total of 1,326 distinct HTML tags, some of which are very common, while others are much rarer across the analysed websites. A well-known practice in text mining studies is to remove the most common and rare words, in the conviction that they rarely convey important information (Feldman and Sanger, 2006; Piantadosi, 2014; Zipf, 1949). Similarly and based on Zhang (2008)'s work, which extends certain natural language properties to programming languages, we decided to retain only HTML tags that appeared at least twice in the dataset, either on the same web-page or different ones. We also excluded mandatory tags like `<html>`, `<head>` and `<body>` from the analysis. Additionally, we removed all HTML tags used by less than 1% or more than 99% of the web-pages in the sample. This cleaning process left us with a final set of 71 HTML tags for the analysis, whose frequency distribution is compared with the overall frequency distribution in Fig. 3.[6]

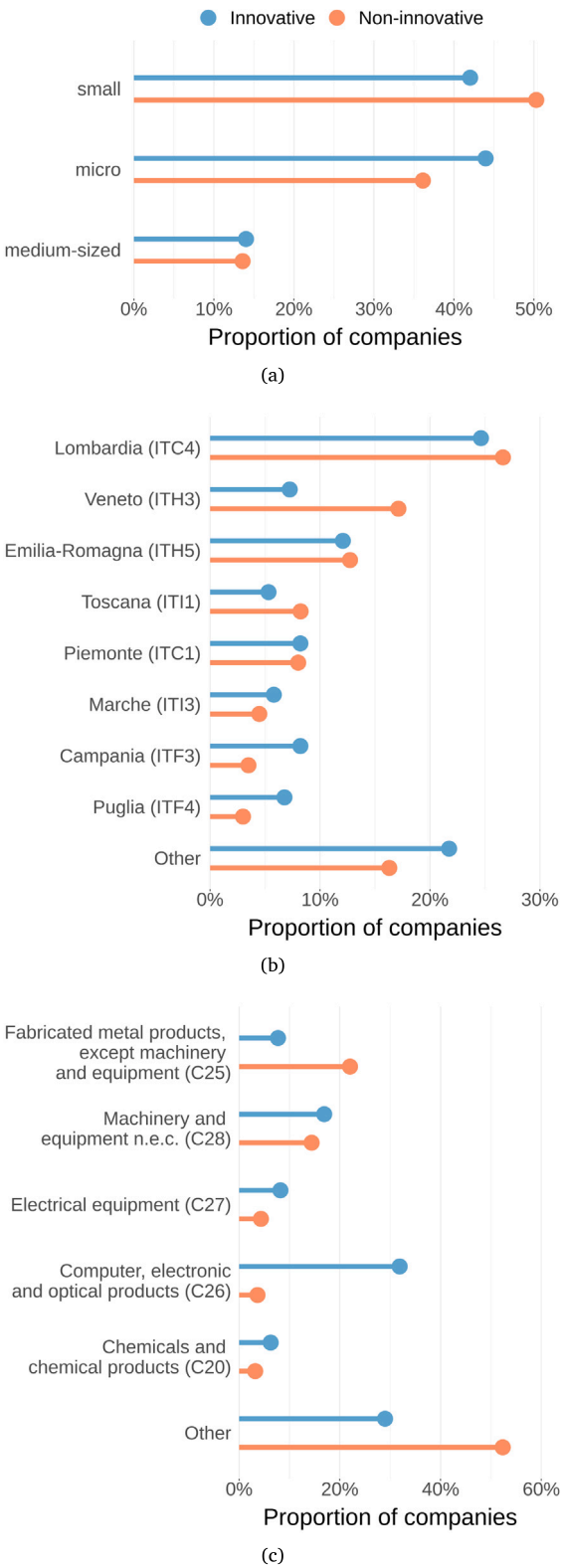

Fig. 2. Distribution of the 42,238 manufacturing SMEs active in 2016 by **(a)** firm size; **(b)** geographical location; and **(c)** industrial sector.

[5] To prevent losing too many of the innovative SMEs from the matches, we reset the list of already matched controls once that less than 10% of the innovative SMEs found a match among the potential controls left to match.

[6] The 71 tags monitored in the analysis covers 50%–65% of the estimated number of the available and valid HTML tags (see https://meiert.com/en/indices/html-elements/). They also include all the 28 most commonly used tags, as estimated by Rosu (2020) based on more than eleven million web-pages.

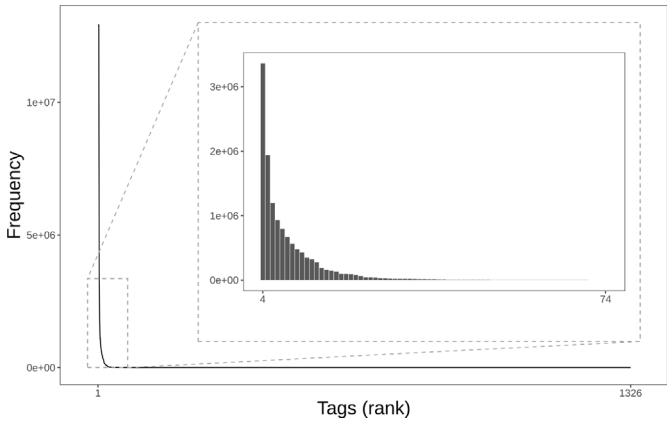

Fig. 3. Frequency of the HTML tags included in the data, before the cleaning procedure described in the text. In the insert, a zoom on the 71 HTML tags that are preserved at the end of the cleaning process.

Table 2
Descriptive statistics on the aggregated *size* variables.

size variable	min	median	mean	max
`html_size`	1,121	44,103	62,321.0	488,281
`text_size`	11	4,780	7,214.0	107,512
`images`	0	20	27.5	240
`hyperlinks`	0	62	79.3	950
`stylesheets`	0	16	23.0	116

Moreover, we calculated five aggregate statistics that measure different *size* aspects of a corporate website. The first two, `html_size` and `text_size`, measure (in bytes) the size of the HTML code for each web-page and the length of the text included within the web-page, respectively. `images` is an indicator of the number of images included in the document, while `hyperlinks` counts the number of user-clickable hyperlinks present in the HTML document. Lastly, `stylesheets` synthesizes the number of external resources, primarily CSS files, used in the web-page. Table 2 shows selected descriptive statistics of the *size* variables.

5. Empirical analysis: Innovative vs non-innovative SMEs

The empirical analysis proceeds in two stages. As a first investigation, we test whether innovative and non-innovative SMEs differ according to the aggregate statistics we built to capture different aspects of a corporate website size. Secondly, we delve deeper into the data to check if certain HTML tags naturally group together, indicating distinct coding styles between innovative and non-innovative firms. For this purpose, we utilize patterns of tag co-occurrence within web-pages and conduct an elaborate cluster analysis on the 71 HTML tags. Subsequently, we compare the usage of these tags between innovative and non-innovative firms at the cluster level and, if the usage of tags within a cluster is found to be different, at the individual tag level too.

5.1. Are the websites of the innovative SMEs bigger?

We compare the size of innovative and non-innovative SMEs websites by analysing differences in the 100 samples of paired firms.

Specifically, for each pair i composed of an innovative firm (inno) and its non-innovative match (ninno), let us define $\delta_i = x_{i,\text{inno}} - x_{i,\text{ninno}}$. Say that x is the variable `html_size`. If $x_{i,\text{inno}} = 112,345$ and $x_{i,\text{ninno}} = 11{,}345$ are the values for a generic pair i, then the difference is $\delta_i = 101,000$. Consequently, let us define $\Delta^j = (\delta_1, \delta_2, \ldots, \delta_n)$ as the vector of the n pairs differences composing a generic sample j. The density distributions of these vectors for all the 100 samples and for each *size variable* are depicted in Fig. 4.

Visual inspection of these differences reveals dissimilar size tendencies of innovative and non-innovative firms. Should the mean of the distributions be zero, there might be no differences between innovative and non-innovative SMEs. At the same time, if the mass of a given distribution is skewed to the right, that specific variable should be systematically larger for the innovative firms than for the otherwise comparable, non-innovative ones. In Fig. 4, the ribbon represents the area between the minimum and maximum density for each x-value among the 100 samples while the five shades of blue that fill the ribbon roughly represent subsequent difference fifths, delimited by the median of the estimated quintiles. These blue shades elicit therefore the extent to which the differences are positive, negative or evenly distributed around zero. For instance, in Fig. 4a about 60% of the differences are positive, suggesting that `html_size` is larger for the innovative firms than for the others.

Since the visual inspection suggests that the difference between innovative and non-innovative firms for any of the *size variables* is larger than zero, to confirm this we run a battery of three tests.

We first test in each sample (j) whether the mean of the differences (μ_{Δ^j}) is zero with a standard paired t-test as in

$$H_0 : \mu_{\Delta^j} = 0, \text{ for } j = 1, \ldots, 100 \tag{1}$$

Moreover, since Fig. 4 suggests a skewed and fat-tailed distribution of these variables, we also run two additional robust tests. We inspect whether the distribution of the differences is symmetrical around zero with the Wilcoxon signed-rank test. This test evaluates the symmetry of a distribution with respect to a hypothesized value, zero in this case. It assesses deviations from the median, considering both their sign and magnitude. The null hypothesis is that the sums of positive and negative ranks are equivalent. That is,

$$H_0 : \text{Med}_{\Delta^j} = 0, \text{ for } j = 1, \ldots, 100 \tag{2}$$

where Med_{Δ^j} is the median of the distribution of the differences of the paired observations in the sample j.

As a second robust test, we check for discrepancies in the tails of the distribution of the differences, applying a test for the symmetry of the deciles (see method D in Wilcox and Erceg-Hurn, 2012). In this last case, the null hypothesis is

$$H_0 : \theta_{q,\Delta^j} + \theta_{1-q,\Delta^j} = 0, \text{ with } q = 0.1, 0.2, 0.3, 0.4;\ j = 1, \ldots, 100 \tag{3}$$

For instance, when $q = 0.1$ we are comparing the first decile of the difference of a variable ($\theta_q = \theta_{0.1}$) with its ninth decile ($\theta_{1-q} = \theta_{0.9}$) and test under the null hypothesis that their sum is zero. In other words, we are testing whether the two deciles are equal in magnitude, but with opposite signs. If this hypothesis is not rejected, it means that the compared deciles are symmetrical around zero and those parts of the difference distribution are not dissimilar. However, even if there is no difference at the extremes of the distribution, there could be an imbalance at the 20% or 40% level, still indicating more positive (or negative) differences.

For all of the three tests above, we opted for a conservative bilateral testing strategy, deriving the sign of the difference, if any, from the visual inspection of Fig. 4.

The output of the analysis consists in the median of the p-values referring to the tests on the 100 samples (Table 3). It is worth noticing that the results evenly suggest that the corporate websites of the innovative SMEs tend to be bigger both in terms of HTML code underlying their homepage, the text it contains, the number of hyperlinks and CSS style-sheets used. Indeed, all the median p-values are under the 5% significance level for all of the *size variables* but, in two cases, for the first decile. As for the number of images, we can see that the difference distributions are not symmetrical in their central part, but it is at the extremes (first and ninth decile). This probably causes the t-test to reject the hypothesis of zero mean at the 10% level of significance only. However, the two couples of deciles dissimilar at the 5% significance level and the Wilcoxon signed-rank test indicate a mostly positive difference in the number of images, pointing to more images on the innovative firms' websites.

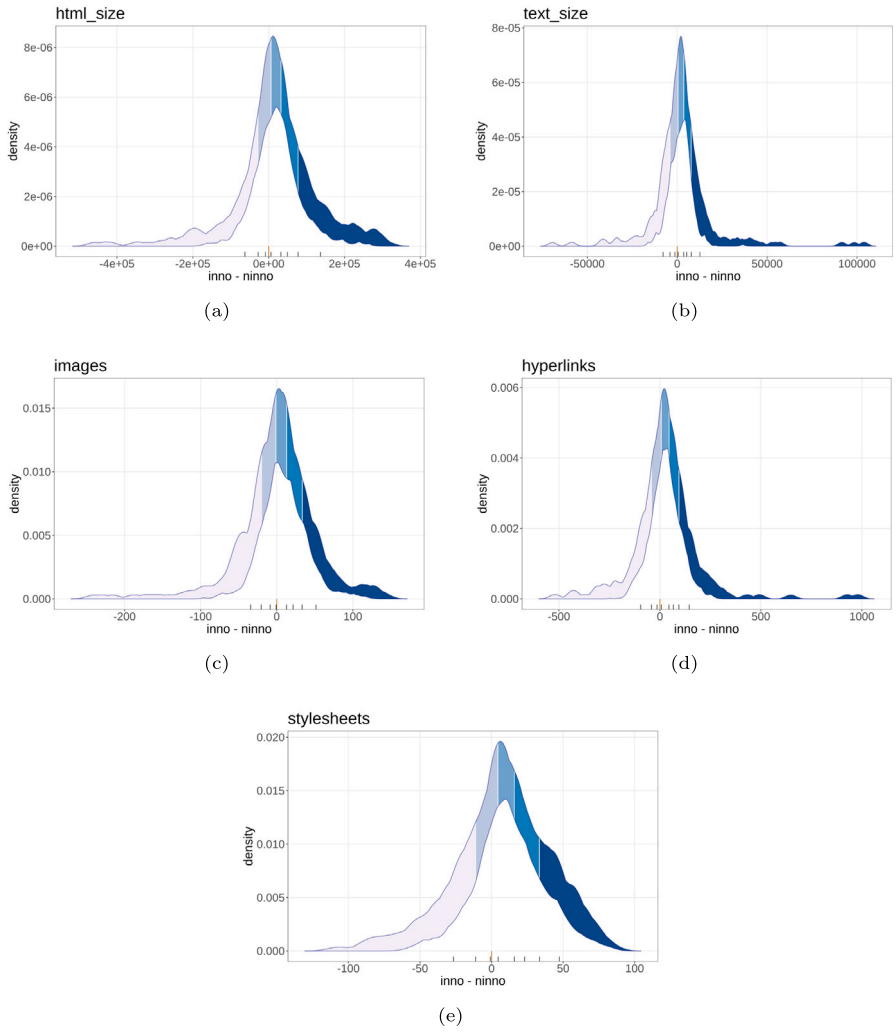

Fig. 4. Density plot of the difference in the *size variables* between innovative and non-innovative SMEs. The ribbon represents the area between the minimum and maximum density for each x-value among the one hundred samples. The filling colour depends on the (estimated) median quintiles of the one hundred samples. The small ticks on the x-axis represent the (estimated) median deciles of the one hundred samples.

Table 3
Median p-value of Paired t-test, Wilcoxon signed-rank test, Quantile test (*method D*) on the differences of size variables between innovative SMEs and paired firms. Visual Differences (V.D.) report the sign observed in Fig. 4.

Size variable	Paired t-test	Wilcoxon signed-rank[a]	Quantile test				V. D.
			10%	20%	30%	40%	
`html_size`	0.001	0.000	0.006	0.002	0.000	0.000	+
`text_size`	0.012	0.003	0.061	0.017	0.002	0.002	+
`images`	0.082	0.041	0.121	0.036	0.052	0.048	+
`hyperlinks`	0.008	0.002	0.064	0.006	0.000	0.000	+
`stylesheets`	0.000	0.000	0.002	0.000	0.000	0.000	+

[a] The signed-rank p-values are calculated using the Pratt ties correction.

5.2. Do innovative firms code differently?

To study the coding style of the sample of SMEs, we apply a hierarchical cluster analysis leading to the identification of seven clusters, which are represented through a Silhouette plot in Fig. 5 (Rousseeuw, 1987). For the interested reader, Appendix B explains the detailed and robust process to cluster tags based on their pairwise similarity and why seven appears as the most reasonable number of clusters based on five different evaluation metrics.

Cluster description. As shown by Fig. 5, *cluster 1* (C1) is composed of two tags – `<form>` and `<input>` – whose aim is to collect user inputs through online forms. *Cluster 2* (C2) groups three tags – `<table>`, `<tr>` and `<td>` – useful to include tables in a website. The third cluster (C3) includes new semantic elements introduced by HTML5, the latest HTML Recommendation at the time to which the data refer. Four tags – `<header>`, `<footer>`, `<section>` and `<nav>` – are useful to communicate to the browser the role taken by different sections of a web-page, and are among the most typical of this 'new paradigm'. While the fifth tag in C3, `<i>`, was already part of the HTML, since its very beginning, to signal portions of text in italic. Over time, it has been replaced by analogous portions of CSS language—a companion of HTML used for describing the presentation of a web document. However, `<i>` found a new aim, recently, as programmers started using

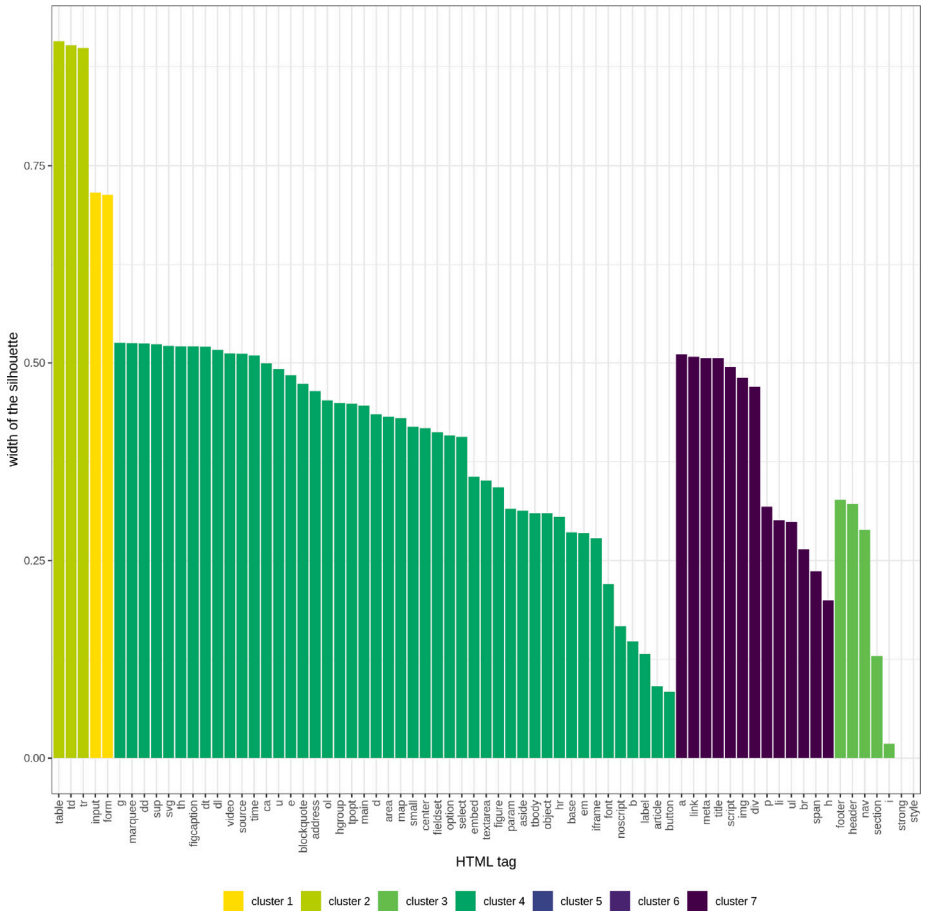

Fig. 5. Silhouette plot of the seven-cluster solution.

it to include icons – like a thumb up – in web-pages.[7] *Cluster 4* (C4) is the largest group, mainly composed of tags that are used by only a few of the web-pages included in the sample. *Cluster 5* (C5) and *cluster 6* (C6) are singletons, each composed of only one tag—as such, they will not be analysed as clusters in the following.

Lastly, *cluster 7* (C7) contains a relatively large group of commonly used tags (see Rosu, 2020) and can be read in opposition to C4. For instance, among tags for inline text semantics, `<a>`, `<span>` and `<br>`, in C7, are used very frequently, while `<small>` and `<u>`, in C4, quite rarely. The same can be said about tags used to structure the text content – `<div>`, `<p>`, `<ul>` and `<li>` in C7 versus `<main>` in C4 – as well as about tags regarding the inclusion of images and multimedia in a website – `<img>` in C7 versus `<figure>`, `<area>` and `<map>` in C4.

We should note that some of the HTML tags are often used in combination with others. For instance, to add a table to your web-page, you must have a `<table>` environment in which you can add rows and columns using tags `<tr>` and `<td>`, respectively. Likewise, the adoption of a given coding style in a part of a web-page will probably result in the same choice in the rest of the page: e.g., the use of `<nav>` to include a menu in a web-page calls for a `<footer>` section. Therefore, we can interpret a larger or smaller use of tags belonging to a particular cluster as indicative of the kind of technology and coding style characterizing a web-page, and, in turn, of the kind of capabilities that an SME put into practice while interacting with the web-programmers and designers that coded it. Thus, we interpreted each cluster as the expression of a coding style.

Table 4
Median p-value of Paired t-test, Wilcoxon signed-rank test, Quantile test (*method D*) on the differences of size variables between innovative SMEs and paired firms. Visual Differences (V.D.) report the sign observed in Fig. 6. Clusters C5 and C6 are omitted since composed of one tag only.

Cluster	Paired t-test	Wilcoxon signed-rank[a]	Quantile test		V. D.
			20%	40%	
C1	0.399	0.391	0.412	0.379	None
C2	0.001	0.001	0.000	0.000	–
C3	0.003	0.003	0.008	0.003	+
C4	0.663	0.603	0.663	0.524	None
C7	0.001	0.003	0.000	0.009	+

[a] The signed-rank p-values are calculated using the Pratt ties correction.

Difference evaluation. To quantify the degree to which a firm follows the coding style represented by a specific cluster, we introduce a measure of adherence defined by the proportion of that cluster's tags detected on a firm web-page.
Formally, if t_k is the number of tags composing cluster k and $t_{h,k}$ is the number of tags of cluster k detected on the website of the hth firm, the adherence a is defined as $a_{h,k} = t_{h,k}/t_k$. For example, if among the five tags in C3 only `<footer>` is used in the web-page of the hth firm, the adherence of that firm to cluster 3 is $a_{h,3} = 1/5$.

In order to evaluate whether innovative and non-innovative firms exhibit a different adherence to each cluster, we proceed similarly to Section 5.1. We compute the difference in adherence between paired firms, plot their distributions for all the 100 samples for visual inspection and proceed to test the null hypotheses as in Eqs. (1), (2) and

[7] For example, this is how popular web-development tools like Bootstrap (https://getbootstrap.com/) and Font Awesome (https://fontawesome.com/) advise users to include icons into a web-page.

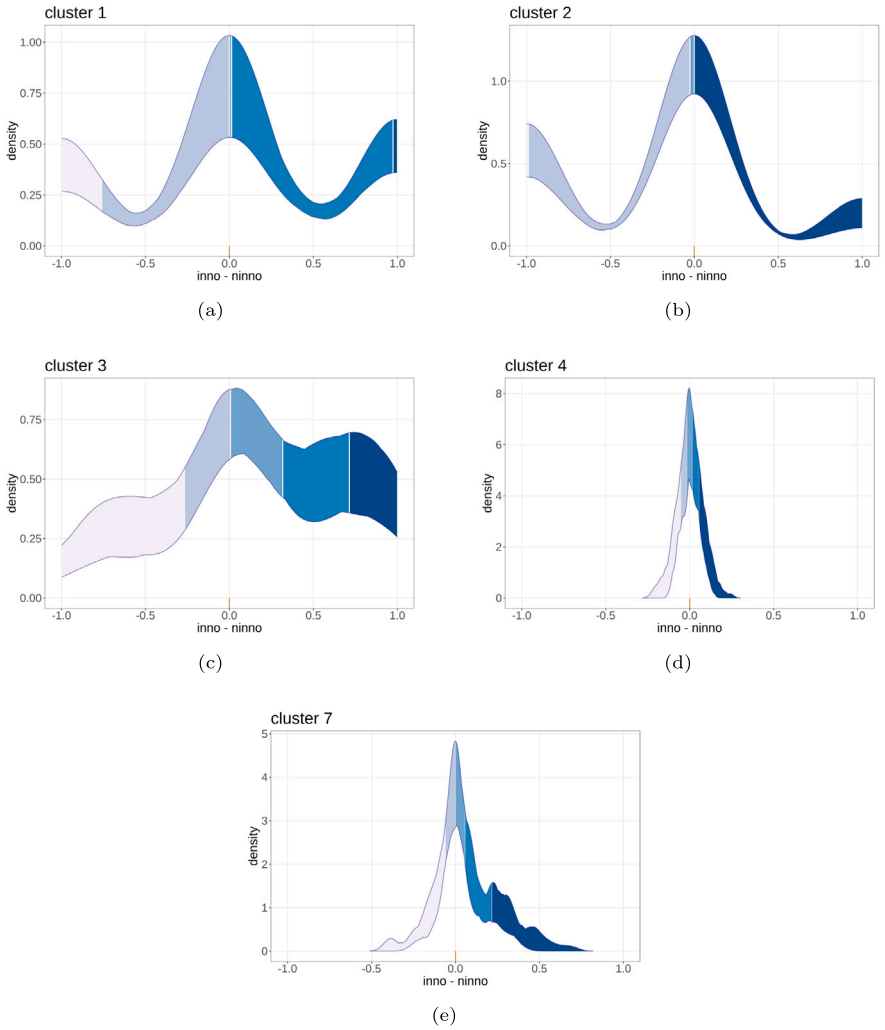

Fig. 6. Density plot of the difference in the adherence to each cluster between innovative and non-innovative SMEs. The ribbon represents the area between the minimum and maximum density for each x-value among the one hundred samples. The filling colour depends on the (estimated) median quintiles of the one hundred samples.

(3). However, to account for a narrower variation range in the data, we decided to test for the differences in cluster adherence in Eq. (3) using the distribution quintiles, preventing potential overlapping between the deciles. The density distributions of the differences in Fig. 6, together with the blue bands representing the differences' fifths, allow for a visual understanding of the difference sign. The tests' outcomes, summarized as the median p-value over the 100 samples, along with the visually-determined sign of the differences, if any, are reported in Table 4.

Innovative and non-innovative SMEs do not adhere equally to clusters C2, C3 and C7 since all the three tests' median p-values are below the 1% level of significance (Table 4). In particular, Fig. 6 shows that innovative SMEs tend to use a smaller proportion of the tags in C2 than their non-innovative counterparts, while the opposite holds for the tags in C3 and C7.

As a robustness check and to enrich the interpretation framework, for the clusters showing unalike adherence between innovative and non-innovative firms, we also test for differences across individual tags (Table 5).[8]

On the one hand, the HTML tags composing C2–`<table>`, `<td>` and `<tr>`–are less used by innovative SMEs than by other comparable firms (Table 5). These tags were used to give a structure to a web-page, putting the content of the page in different cells of a table with transparent borders. This stylistic solution, commonly used in the past, is today discouraged by standard practices because it is unfit for responsive pages—i.e., able to adapt to any device, from a large desktop PC to a small smartphone.[9] Therefore, the fact that innovative SMEs tend to use these tags less than their non-innovative counterparts is in line with the idea that the first group of firms is better equipped at capturing novel technological trends.

On the other hand, the HTML tags typical of HTML5, mostly clustered in C3 – like `<footer>` or `<header>` – are more often used within the web-pages of innovative firms (Table 5). As explained by Tabarés (2021), the radical novelty introduced into the web technology by HTML5 was pushed not only by technical limitations of the previously existing standard but also by the new needs of the so-called 'platform economy'. Said differently, the HTML was suited for rendering text content, but not that much for embedding videos or

[8] In commenting, we consider as significant those differences where the majority of tests provide p-values under the 5% level of significance. We ignore those cases where none of the three tests described above point to significant differences.

[9] About the extensive use of the `<table>` tag for web design purposes done in the second half of the 1990s, see Siegel (1996, ch. 4). About more modern approaches that overcome tables in favour of more flexible HTML structures and a more extensive use of CSS, see the ideas of 'tableless' and 'responsive' web design, among others.

Table 5
Median *p*-value of Paired *t*-test, Wilcoxon signed-rank test, Quantile test (*method D*) on the differences of size variables between innovative SMEs and paired firms. Visual Differences (V.D.) report the sign derived visually from figures in the Supplementary Material. Please, notice that the only tags included in the table with *p*-values above 5% for all the tests are
 and <title>.

HTML tag	Clu.	Paired *t*-test	Wilcoxon signed-rank[a]	Quantile test				V. D.
				10%	20%	30%	40%	
<table>	C2	0.175	0.005	0.199	0.007	0.002	0.002	–
<td>	C2	0.184	0.006	0.145	0.012	0.004	0.002	–
<tr>	C2	0.158	0.005	0.118	0.009	0.002	0.002	–
<footer>	C3	0.000	0.001	0.008	0.000	0.000	0.000	+
<header>	C3	0.051	0.011	0.028	0.011	0.012	0.014	+
<i>	C3	0.022	0.002	0.009	0.002	0.002	0.002	+
<nav>	C3	0.015	0.019	0.029	0.022	0.027	0.028	+
<section>	C3	0.004	0.009	0.010	0.008	0.015	0.015	+
<a>	C7	0.008	0.002	0.064	0.006	0.000	0.000	+
 	C7	0.468	0.154	0.277	0.223	0.219	0.182	None
	C7	0.001	0.000	0.026	0.002	0.000	0.000	+
<h>	C7	0.009	0.002	0.156	0.004	0.000	0.000	+
<img>	C7	0.082	0.041	0.121	0.036	0.052	0.048	+
	C7	0.037	0.004	0.235	0.029	0.004	0.000	+
<link>	C7	0.000	0.000	0.002	0.000	0.000	0.000	+
<meta>	C7	0.007	0.002	0.006	0.006	0.002	0.004	+
<p>	C7	0.065	0.005	0.175	0.008	0.002	0.004	+
<script>	C7	0.001	0.000	0.014	0.002	0.000	0.000	+
	C7	0.013	0.001	0.039	0.002	0.000	0.000	+
<title>	C7	0.171	0.054	0.129	0.051	0.054	0.059	None
	C7	0.011	0.002	0.073	0.004	0.000	0.000	+

[a] The signed-rank *p*-values are calculated using the Pratt ties correction.

letting the user interact with the website. The HTML5 paradigm helps web developers to design flexible web-pages – more suited for any kind of device –, embed multimedia content into them and facilitate the interaction of casual users with the website. These technical advances are key for value propositions related to digital platforms. For example, having a responsive and mobile-first website is key for marketing and economic reasons. First, the page will look nicer to the customer who approaches it from any device, and it will be easier for her to interact with its contents. Moreover, search engines tend to prioritize, in the results ranking, these mobile-first pages. Therefore, in line with the working hypothesis, it is not surprising that the innovative SMEs employ this cluster of tags more than the similar, non-innovative ones, and invest more in alternative stylistic and technological solutions.

Consistently with a larger adherence to C7 by the innovative firms, similar conclusions hold also when we move to the analysis of the individual tags (Table 4). The tags in C7 are used to enrich the structure of a website, whether with stylistic choices – like for and – or with links to hypertexts and style-sheets (<a> and <link>, respectively), or to notify to search engines and social media platforms useful metadata about the web-page (<meta>), by structuring lists or menus – like <ul> and <li> –, and finally embellishing it with more images (<img>). Lastly, more <script> tags in the innovative SMEs websites are quite indicative of more advanced technology: this tag is used to embed (client-side) executable code written in JavaScript.

Additional evidence regarding the companies that adhere most to clusters 2, 3, and 7 is provided in the Supplementary Material.

6. Conclusion

In this paper, we proposed and explored the use of HTML code, retrieved from corporate websites, as an alternative information source for identifying innovative SMEs. To accomplish this, we collected the website URLs for all active Italian manufacturing SMEs present in BvD-Orbis. Subsequently, we thoroughly checked the firm-website matching accuracy. This novel assessment procedure markedly decreased the number of false positives in the data, even though significantly reducing the sample size. We identified those small businesses that enrolled in the special section of the Italian Business Register defined by the Italian Startup Act as 'innovative firms'. With this data, we built a robust framework comprising one hundred samples of both innovative and non-innovative SMEs, carefully matched based on relevant confounding factors.

The homepage of each corporate website was then scraped, from the Internet Archive, to obtain a selection of HTML tags and a few size measures of the websites, which were used as a basis for comparing innovative and non-innovative small businesses.

The results point to larger websites for innovative SMEs, whether measured by the HTML code underlying their homepage, by the size of the enclosed text or by other alternative measures of a web-page size. Additionally, seven clusters of HTML tags – based on the frequency with which they are co-used on the web pages – emerged. We showed that three of these clusters are employed to a different degree by the two groups of businesses. Innovative SMEs' websites are richer, more up-to-date, and more complex compared to those of non-innovative ones. These results indicate that the stylistic and technological features of the HTML code of corporate websites reflect unobservable characteristics related to high skills and creativity, indicative of a small and medium enterprise's capacity to undertake innovative activities.

Therefore, our study represents the first step in transforming the HTML code of corporate websites into data for identifying innovative SMEs and potentially deriving innovation indicators.

However, we recognize that the present study has specific limitations. Firstly, the focus on manufacturing firms circumscribes the conclusions. In the future, it will be important to extend the approach to other sectors for which websites represent a unique showcase, such as services. Secondly, we did not analyse tags other than those found on website homepages. Further research should include additional features that could validate the conclusions.

More in general, we should point that working with unconventional data, such as web-scraped data, is not straightforward due to their unique features. Collecting, merging, storing, and analysing unstructured data, such as the HTML code of corporate websites, require adequate skills and expertise. By developing this project and with reference to the literature, we identified the following challenges, which we partly addressed.

Accessing unconventional data, particularly those collected from web platforms, is increasingly challenging. The information generated by online ventures is often deemed a strategic asset by companies, leading to restricted access or questioning the legitimacy of data scraping (Scassa, 2019).

Unconventional data sources have limitations that can impact the quality of derived statistics (Cebrián and Domenech, 2023). Among others, incompleteness, inconsistency and inaccuracy are commonly encountered (Karr et al., 2006). For instance, Kinne and Axenbeck (2020) discovered under-representation of the agricultural sector, among firms with fewer than five employees and recently established firms. On the contrary, patent assignees were found to be over-represented in web-based data. These drawbacks can be solved by integrating unconventional with conventional data, as demonstrated in previous studies involving firms' websites (Daas and van der Doef, 2020; Crosato et al., 2021, 2023). On the other hand the integration can present challenges too, because conventional data is tidily structured in relational databases, whereas unconventional data rarely follows this format. For example, in this paper, we discuss the creative yet arduous process of identifying keys to merge website data with financial statements, as well as the subsequent evaluation of data quality.

Finally, the dynamic nature of corporate websites allows for the creation of panels to analyse firm online behaviours, but it also implies the need for constant forward monitoring. On the other hand, the use of the Wayback Machine of the Internet Archive, or similar tools, makes backward analysis of website's past structure possible (Arora et al., 2016; Blazquez et al., 2018).

Concluding, in recent years the use of corporate websites as information sources for measuring the innovative activity of SMEs has shown its potential. Until now, scholars mostly exploited the natural language content of these websites to build useful innovation indicators. However, we have shown that the HTML structure of these corporate websites too must be acknowledged alongside their textual features as informative of the innovativeness of a business. Therefore, we hope that future research does not overlook this information piece as an additional feature to build web-based firm-level innovation indicators.

CRediT authorship contribution statement

Carlo Bottai: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Lisa Crosato:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Josep Domenech:** Data curation. **Marco Guerzoni:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Caterina Liberati:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data are freely available on the Internet.

Acknowledgments

We thank the Italian Ministry of University and Research (MUR) for sponsoring this work under the 'Departments of Excellence 2018–2022' funding schema. We greatly acknowledge the DEMS Data Science Lab of the University of Milano–Bicocca for supporting this work by providing computational resources. This study was carried out within the project 'Data Driven Innovation. Measuring its Effects on Industries, Firms and Business Models' prot. nr. 2022JHZL37 and received funding from the European Union Next-GenerationEU - National Recovery and Resilience Plan (NRRP) – MISSION 4 COMPONENT 2, INVESTIMENT 1.1 Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN) – CUP N. H53D23002440006; H53D23002450006. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

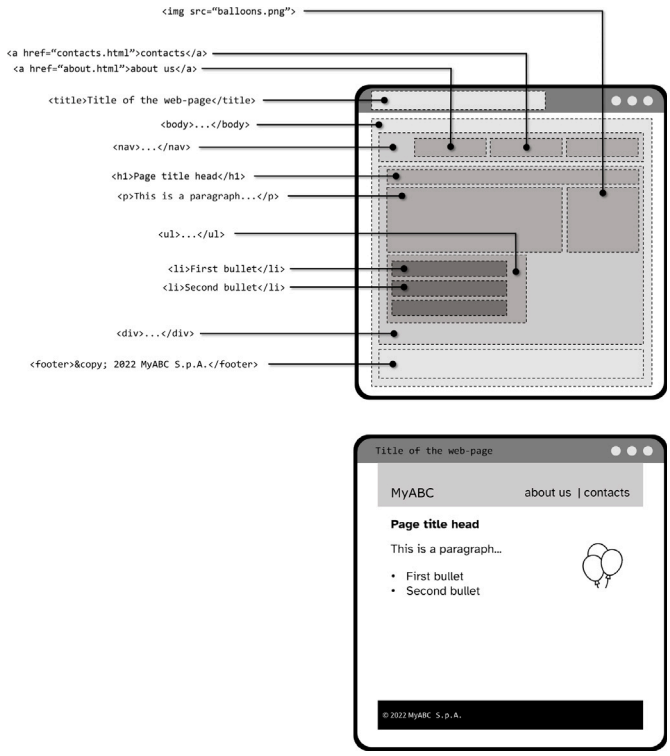

Fig. A.7. A hypothetical HTML code describing the structure of a web-page (upper part) and how this translates into what we would see through a web-browser (bottom part).

Appendix A. HTML code

The HyperText Markup Language (HTML) is a markup language designed to help a web-browser display the structure and content of a web document. Together with Cascading Style Sheets (CSS) and JavaScript (JS), it is the main technological component of the World Wide Web (WWW). Specifically, the HTML describes, semantically, the structure of a web-page by surrounding portions of the text with *tags* to describe its function. E.g., `<a href="https://www.dolor_sit_am.et">Lorem ipsum</a>` informs the browser that 'Lorem ipsum' is the *anchor* of a link pointing to https://www.dolor_sit_am.et; as specified by the hypertext reference (`href`) *attribute* of the `<a>` tag. Being informed about this special function of this portion of text, the browser will provide you with the possibility to click on it to be redirected to the specified URL. Instead, other *tags*, like `<img src=''nameplate.png''>` directly introduce content into the page; in this case the browser will display `nameplate.png` as an image within the web-page. Fig. A.7 illustrates how a hypothetical HTML code, describing the structure of a web-page, translates into what we would see through a web-browser. In the upper portion of the figure, several dashed-bordered blocks represent some elements that constitute this imaginary web-page. Each block corresponds to an HTML tag. For example, the `<title>` tag contains meta-data informing the browser about the title of the web-page, that in general will be displayed in the upper part of the browser's window. The tag `<nav>` contains that menu of the web-page, and within it, you could find a list of `<a>` tags. Each of these `<a>` tags, points to a sub-page of the web-site: in the example, by clicking on the first you would open a web-page with the story, vision and mission of the company; while by clicking on the second you would see a web-page with a map representing where the company is located. Then, we have a `<div>` tag, that is a generic box containing other tags. Within it, we find a `<h1>` tag, with the main title of the

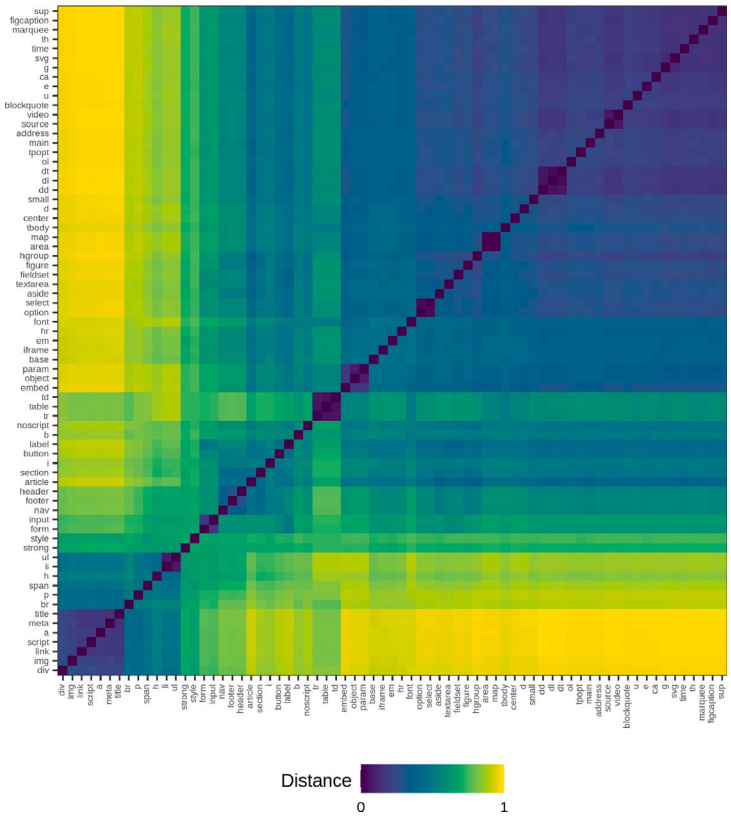

Fig. B.8. Distance matrix **D** between HTML tags. A dark blue colour indicates an association between two tags, while light yellow suggests its absence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

web-page, followed by a `<p>` tag with a portion of text. The `<img>` tag includes an image, taken from the 'balloons.png' file, in the web-page. The `<ul>` tag identifies a block of bullet points, each of which is declared by a `<li>` tag. Lastly, the `<footer>` is a box semantically describing a portion of text to be rendered at the end of the web-page. The bottom portion of Fig. A.7 displays the rendered page, based on the HTML code just described.

Appendix B. Cluster analysis

We cluster HTML tags based on their pairwise similarity, where two tags are more *similar* the more they tend to be used together on the same web-page. Specifically, we follow a three-step approach (Sulc et al., 2022): (i) we create a website-tag binary matrix; (ii) we calculate a distance $d_{tt'}$ between tags; (iii) we group the tags in clusters so that two similar tags will tend to be part of the same cluster.

More precisely, the first step consists of transforming the website-tag contingency table $\mathbf{M} = [m_{wt}]$ into a presence-absence one, $\mathbf{A} = [a_{wt}]$, such that $a_{wt} = 1$ if the HTML tag t is used at least once in the code of web-page w, and 0 otherwise.

In the second step, to calculate the pairwise similarity between tags, we use the 'simple matching coefficient' introduced by Sokal and Michener (1958). Supposing that the relationship between two tags is defined by the following table:

t \ t'	Absent	Present
Absent	α	β
Present	γ	δ

the similarity between tags t and t' is defined as $s_{tt'} \equiv (\alpha + \delta)/(\alpha + \beta + \gamma + \delta)$, where δ is the number of web-pages that use both tags, and so on. Unlike the widely used Jaccard coefficient, the 'simple matching' takes into account the agreement between two HTML tags, both in their mutual presence and absence on a given web-page, comparing it to the number of sampled web-pages. In a series of tests not reported in the paper, we observed that, compared to alternative coefficients, this preserves important information about the similarity between tags that was effectively exploited by the clustering algorithm to bring out coherent groups of tags. As shown by Gower and Legendre (1986), from the similarity matrix we can get a Euclidean distance matrix $\mathbf{D} = [d_{tt'}]$, by defining the pairwise distance between the HTML tags t and t' as $d_{tt'} = \sqrt{1 - s_{tt'}}$. The obtained distance matrix is represented in Fig. B.8, where we can spot a number of block-diagonal darker areas suggesting the existence of groups of tags that tend to be co-used (and not simply used) by several web-pages.

In the third step we apply a hierarchical agglomerative clustering technique to the distance matrix, calculating the distance between the groups with the unweighted pair group method with arithmetic mean, known as UPGMA (Sokal and Michener, 1958). The iterative process starts by considering each HTML tag as a trivial cluster; then, at each step, it blends the two most similar clusters, until only one giant cluster is left (see, e.g., Everitt et al., 2011). A major advantage of agglomerative clustering is to allow the researcher to suggest the number of clusters (k) ex-post based on the output of the process. Here the results are depicted through a Silhouette plot in Fig. 5, where a seven-clusters solution is highlighted.

The number of clusters was set to seven, based on four internal evaluation criteria for hierarchical clustering of binary data (Šulc et al., 2018), reported in Fig. B.9. We adopt two likelihood-based evaluation criteria for which lower values indicate a better number of clusters: the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), properly defined for categorical data (Bacher et al., 2004). Furthermore, we compute the 'Best K' index (BK), a variability-based metric defined as the second-order difference of the incremental entropy of the dataset with k clusters (Chen and Liu, 2009): the highest

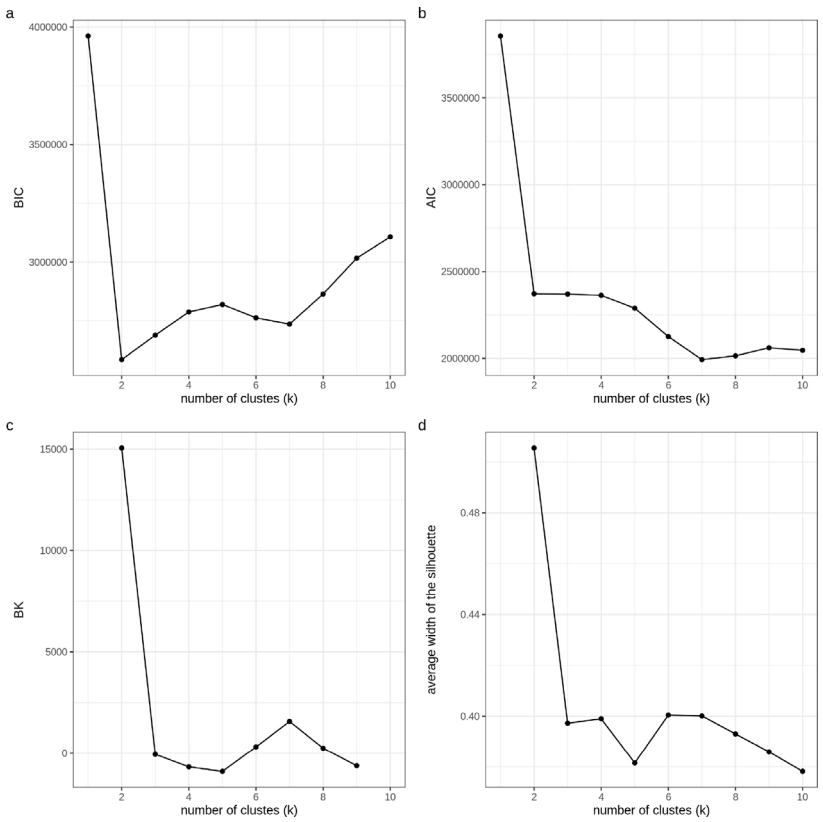

Fig. B.9. Internal evaluation criteria for hierarchical clustering of binary data. **(a)** Bayesian Information Criterion (BIC), **(b)** Akaike Information Criterion (AIC), **(c)** 'Best K' index (BK) and **(d)** Silhouette profile.

this metric is, the more appropriate the number of clusters. The previous three criteria prefer *compact* clusters; i.e., small clusters that group together similar objects. The last considered metric – the so-called *silhouette* (Rousseeuw, 1987) – is a distance-based metric that tries to balance *compactness* with *separation*; i.e., to find clusters not only with a high within-cluster similarity but also with a high between-clusters distance. A silhouette s_t close to one (minus one) implies that the average distance of tag t to the other members of its cluster is short (long), while s_t is about zero when t could have been assigned about at random to the current cluster or its second-best option. Rousseeuw proposes to use the average, over all the HTML tags, of s_i to detect the number of clusters (k) that better represents the data under scrutiny.

Since the four adopted criteria suggest a seven-cluster ($k = 7$) either as the first- or second-best solution, we decided to adopt it, also considering that the alternative, two-clusters solution, does not offer advantages in terms of results interpretability.

Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.techfore.2024.123597.

References

Abbasiharofteh, M., Kinne, J., Krüger, M., 2023. Leveraging the digital layer: the strength of weak and strong ties in bridging geographic and cognitive distances. J. Econ. Geogr. http://dx.doi.org/10.1093/jeg/lbad037, lbad037.

Antonietti, R., Gambarotto, F., 2020. The role of industry variety in the creation of innovative start-ups in Italy. Small Bus. Econ. 54, 561–573. http://dx.doi.org/10.1007/s11187-018-0034-4.

Antons, D., Grünwald, E., Cichy, P., Salge, T.O., 2020. The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. R & D Manage. 50, 329–351. http://dx.doi.org/10.1111/radm.12408.

Arora, S.K., Li, Y., Youtie, J., Shapira, P., 2016. Using the Wayback Machine to mine websites in the social sciences: A methodological resource. J. Assoc. Inf. Sci. Technol. 67, 1904–1915. http://dx.doi.org/10.1002/asi.23503.

Arora, S.K., Li, Y., Youtie, J., Shapira, P., 2020. Measuring dynamic capabilities in new ventures: Exploring strategic change in US green goods manufacturing using website data. J. Technol. Transfer 45, 1451–1480. http://dx.doi.org/10.1007/s10961-019-09751-y.

Arora, S.K., Youtie, J., Shapira, P., Gao, L., Ma, T., 2013. Entry strategies in an emerging technology: A pilot web-based study of graphene firms. Scientometrics 95, 1189–1207. http://dx.doi.org/10.1007/s11192-013-0950-7.

Arundel, A., O'Brien, K., Torugsa, A., 2013. How firm managers understand innovation: implications for the design of innovation surveys. In: Gault, F. (Ed.), Handbook of Innovation Indicators and Measurement. Edward Elgar Publishing, pp. 88–108. http://dx.doi.org/10.4337/9780857933652.00012, chapter 4.

Arundel, A., Smith, K., 2013. History of the community innovation survey. In: Gault, F. (Ed.), Handbook of Innovation Indicators and Measurement. Edward Elgar Publishing, pp. 60–87. http://dx.doi.org/10.4337/9780857933652.00011, chapter 3.

Ashby, C., Weir, D., 2020. Leveraging HTML in free text web named entity recognition. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, pp. 407–413. http://dx.doi.org/10.18653/v1/2020.coling-main.36.

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., Cunningham, S., 2022. Indicators on firm level innovation activities from web scraped data. Data Brief 42, 108246. http://dx.doi.org/10.1016/j.dib.2022.108246.

Axenbeck, J., Breithaupt, P., 2021. Innovation indicators based on firm websites—which website characteristics predict firm-level innovation activity?. PloS One 16, 1–23. http://dx.doi.org/10.1371/journal.pone.0249583.

Bacher, J., Wenzig, K., Vogler, M., 2004. SPSS TwoStep Cluster – a first evaluation. Discussion Paper 2004-2. Wirtschafts- und Sozialwissenschaftliche Fakultät, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Barcaroli, G., Scannapieco, M., Summa, D., 2016. On the use of Internet as a data source for official statistics: a strategy for identifying enterprises on the Web. Rivista Italiana Econ. Demogr. Stat. 70, 25–41.

Blazquez, D., Domenech, J., 2018. Big data sources and methods for social and economic analyses. Technol. Forecast. Soc. Change 130, 99–113. http://dx.doi.org/10.1016/j.techfore.2017.07.027.

Blazquez, D., Domenech, J., Debón, A., 2018. Do corporate websites' changes reflect firms' survival? Online Inf. Rev. 42, 956–970. http://dx.doi.org/10.1108/OIR-11-2016-0321.

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., Liberati, C., 2022. Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In: Proceedings of the 2022 ACM Conference on Information Technology for Social Good. Association for Computing Machinery, New York, NY, USA, pp. 338–344. http://dx.doi.org/10.1145/3524458.3547246.

Brinck, T., Gergle, D., Wood, S.D., 2001. Usability for the Web: Designing Web Sites that Work. Morgan Kaufmann Publishers.

Cebrián, E., Domenech, J., 2023. Is Google Trends a quality data source? Appl. Econ. Lett. 30, 811–815. http://dx.doi.org/10.1080/13504851.2021.2023088.

Chen, K., Liu, L., 2009. Best K: critical clustering structures in categorical datasets. Knowl. Inf. Syst. 20, 1–33. http://dx.doi.org/10.1007/s10115-008-0159-x.

Cohen, W.M., Nelson, R.R., Walsh, J.P., 2000. Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent (or not). http://dx.doi.org/10.3386/w7552, Working Paper 7552. National Bureau of Economic Research.

Colombelli, A., 2016. The impact of local knowledge bases on the creation of innovative start-ups in Italy. Small Bus. Econ. 47, 383–396. http://dx.doi.org/10.1007/s11187-016-9722-0.

Crosato, L., Domenech, J., Liberati, C., 2021. Predicting SME's default: Are their websites informative? Econom. Lett. 204, 109888. http://dx.doi.org/10.1016/j.econlet.2021.109888.

Crosato, L., Domenech, J., Liberati, C., 2023. Websites' data: a new asset for enhancing credit risk modeling. Ann. Oper. Res. 1–16. http://dx.doi.org/10.1007/s10479-023-05306-5.

Daas, P.J.H., van der Doef, S., 2020. Detecting innovative companies via their website. Stat. J. IAOS 36, 1239–1251. http://dx.doi.org/10.3233/SJI-200627.

Daniel, E.M., Wilson, H.N., 2003. The role of dynamic capabilities in e-business transformation. Eur. J. Inf. Syst. 12, 282–296. http://dx.doi.org/10.1057/palgrave.ejis.3000478.

Domenech, J., de la Ossa, B., Pont, A., Gil, J.A., Martinez, M., Rubio, A., 2012. An intelligent system for retrieving economic information from corporate websites. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, Washington, DC, USA, pp. 573–578. http://dx.doi.org/10.1109/WI-IAT.2012.92.

Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Hierarchical Clustering, fifth ed. John Wiley & Sons, pp. 71–110. http://dx.doi.org/10.1002/9780470977811.ch4, chapter 4.

Feldman, R., Sanger, J., 2006. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

Fontana, R., Nuvolari, A., Shimizu, H., Vezzulli, A., 2013. Reassessing patent propensity: Evidence from a dataset of R & D awards, 1977–2004. Res. Policy 42, 1780–1792. http://dx.doi.org/10.1016/j.respol.2012.05.014.

Gault, F. (Ed.), 2013. Handbook of Innovation Indicators and Measurement. Edward Elgar Publishing, http://dx.doi.org/10.4337/9780857933652.

Giotopoulos, I., Kontolaimou, A., Korra, E., Tsakanikas, A., 2017. What drives ICT adoption by SMEs? evidence from a large-scale survey in Greece. J. Bus. Res. 81, 60–69. http://dx.doi.org/10.1016/j.jbusres.2017.08.007.

Gök, A., Waterworth, A., Shapira, P., 2015. Use of web mining in studying innovation. Scientometrics 102, 653–671. http://dx.doi.org/10.1007/s11192-014-1434-0.

Gower, J.C., Legendre, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classification 3, 5–48. http://dx.doi.org/10.1007/BF01896809.

Guerzoni, M., 2010. The impact of market size and users' sophistication on innovation: The patterns of demand. Econ. Innov. New Technol. 19, 113–126. http://dx.doi.org/10.1080/10438590903016526.

Guerzoni, M., Nava, C.R., Nuccio, M., 2021. Start-ups survival through a crisis. Combining machine learning with econometrics to measure innovation. Econ. Innov. New Technol. 30, 468–493. http://dx.doi.org/10.1080/10438599.2020.1769810.

Gür, I., Nachum, O., Miao, Y., Safdari, M., Huang, A., Chowdhery, A., Narang, S., Fiedel, N., Faust, A., 2023. Understanding HTML with large language models. http://dx.doi.org/10.48550/arXiv.2210.03945, Mimeo 2210.03945. arXiv.

Guzman, J., Li, A., 2023. Measuring founding strategy. Manage. Sci. 69, 101–118. http://dx.doi.org/10.1287/mnsc.2022.4369.

Hall, B.H., Mairesse, J., Mohnen, P., 2010. Measuring the returns to R & D. In: Hall, B.H., Rosenberg, N. (Eds.), Handbook of the Economics of Innovation, Vol. 2. North-Holland, pp. 1033–1082. http://dx.doi.org/10.1016/S0169-7218(10)02008-3, chapter 24.

Haller, S.A., Siedschlag, I., 2011. Determinants of ICT adoption: Evidence from firm-level data. Appl. Econ. 43, 3775–3788. http://dx.doi.org/10.1080/00036841003724411.

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Piqueras, L.C., Chalkidis, I., Cui, R., et al., 2022. Challenges and strategies in cross-cultural nlp. arXiv preprint arXiv:2203.10020.

Holgersson, M., 2013. Patent management in entrepreneurial SMEs: a literature review and an empirical study of innovation appropriation, patent propensity, and motives. R & D Manage. 43, 21–36. http://dx.doi.org/10.1111/j.1467-9310.2012.00700.x.

Hopcroft, J.E., Karp, R.M., 1973. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. SIAM J. Comput. 2, 225–231. http://dx.doi.org/10.1137/0202019.

Janardan Mehta, S., 2017. Concept drift in streaming data classification: Algorithms, platforms and issues. Procedia Comput. Sci. 122, 804–811. http://dx.doi.org/10.1016/j.procs.2017.11.440.

Karr, A.F., Sanil, A.P., Banks, D.L., 2006. Data quality: A statistical perspective. Stat. Methodol. 3, 137–173. http://dx.doi.org/10.1016/j.stamet.2005.08.005.

Katz, J.S., Cothey, V., 2006. Web indicators for complex innovation systems. Res. Eval. 15, 85–95. http://dx.doi.org/10.3152/147154406781775922.

Kinne, J., Axenbeck, J., 2020. Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. Scientometrics 125, 2011–2041. http://dx.doi.org/10.1007/s11192-020-03726-9.

Kinne, J., Lenz, D., 2021. Predicting innovative firms using web mining and deep learning. PloS One 16, 1–18. http://dx.doi.org/10.1371/journal.pone.0249071.

Kitchin, R., 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures & their Consequences. SAGE Publications, http://dx.doi.org/10.4135/9781473909472.

Kuznets, S., 1962. Inventive activity: Problems of definition and measurement. In: of Economic Research, National Bureau (Ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton University Press, pp. 19–52. http://dx.doi.org/10.1515/9781400879762-002.

Lawson, B., Samson, D., 2001. Developing innovation capability in organisations: a dynamic capabilities approach. Int. J. Innov. Manag. 5, 377–400. http://dx.doi.org/10.1142/S1363919601000427.

Li, Y., Arora, S.K., Youtie, J., Shapira, P., 2018. Using web mining to explore Triple Helix influences on growth in small and mid-size firms. Technovation 76–77, 3–14. http://dx.doi.org/10.1016/j.technovation.2016.01.002.

Li, J., Xu, Y., Cui, L., Wei, F., 2022. MarkupLM: Pre-training of text and markup language for visually rich document understanding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 6078–6087. http://dx.doi.org/10.18653/v1/2022.acl-long.420.

Libaers, D., Hicks, D., Porter, A.L., 2016. A taxonomy of small firm technology commercialization. Ind. Corp. Change 25, 371–405. http://dx.doi.org/10.1093/icc/dtq039.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G., 2019. Learning under concept drift: A review. IEEE Trans. Knowl. Data Eng. 31, 2346–2363. http://dx.doi.org/10.1109/TKDE.2018.2876857.

Lundvall, B.Ä., Johnson, B., 1994. The learning economy. J. Ind. Stud. 1, 23–42. http://dx.doi.org/10.1080/13662719400000002.

Mairesse, J., Mohnen, P., 2010. Using innovation surveys for econometric analysis. In: Hall, B.H., Rosenberg, N. (Eds.), Handbook of the Economics of Innovation, Vol. 2. North-Holland, pp. 1129–1155. http://dx.doi.org/10.1016/S0169-7218(10)02010-1, chapter 26.

Nagaoka, S., Motohashi, K., Goto, A., 2010. Patent statistics as an innovation indicator. In: Hall, B.H., Rosenberg, N. (Eds.), Handbook of the Economics of Innovation, Vol. 2. North-Holland, pp. 1083–1127. http://dx.doi.org/10.1016/S0169-7218(10)02009-5, chapter 25.

Nathan, M., Rosso, A., 2015. Mapping digital businesses with big data: Some early findings from the UK. Res. Policy 44, 1714–1733. http://dx.doi.org/10.1016/j.respol.2015.01.008.

Nathan, M., Rosso, A., 2022. Innovative events: product launches, innovation and firm performance. Res. Policy 51, 104373. http://dx.doi.org/10.1016/j.respol.2021.104373.

OECD, 2015. Frascati manual 2015: Guidelines for collecting and reporting data on research and experimental development. In: The Measurement of Scientific, Technological and Innovation Activities. OECD Publishing, http://dx.doi.org/10.1787/9789264239012-en, URL https://www.oecd-ilibrary.org/content/publication/9789264239012-en, 1st ed. 1963 entitled The Proposed Standard Practice for Surveys of Research and Experimental Development.

OECD, Eurostat, 2018. Oslo Manual 2018: Guidelines for Collecting, Reporting and using Data on Innovation, the Measurement of Scientific, Technological and Innovation Activities. OECD Publishing and Eurostat, http://dx.doi.org/10.1787/9789264304604-en, URL https://www.oecd-ilibrary.org/content/publication/9789264304604-en, 1st ed. 1992 entitled OECD Proposed Guidelines for Collecting and Interpreting Technological Innovation Data.

Piantadosi, S.T., 2014. Zipf's word frequency law in natural language: A critical review and future directions. Psychon. Bull. & Rev. 21, 1112–1130. http://dx.doi.org/10.3758/s13423-014-0585-6.

Pikuliak, M., Šimko, M., Bieliková, M., 2021. Cross-lingual learning for text processing: A survey. Expert Syst. Appl. 165, 113765.

Rammer, C., Es-Sadki, N., 2023. Using big data for generating firm-level innovation indicators - a literature review. Technol. Forecast. Soc. Change 197, 122874. http://dx.doi.org/10.1016/j.techfore.2023.122874.

Rosu, C., 2020. HTML study. URL https://www.advancedwebranking.com/seo/html-study/. (Last Accessed 19 July 2023).

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Santarelli, E., Sterlacchini, A., 1990. Innovation, formal vs. informal R & D, and firm size: Some evidence from Italian manufacturing firms. Small Bus. Econ. 2, 223–228. http://dx.doi.org/10.1007/BF00389530.

Scassa, T., 2019. Ownership and control over publicly accessible platform data. Online Inf. Rev. 43, 986–1002. http://dx.doi.org/10.1108/OIR-02-2018-0053.
Shapira, P., Gök, A., Salehi, F., 2016. Graphene enterprise: Mapping innovation and business development in a strategic emerging technology. J. Nanoparticle Res. 18, 269. http://dx.doi.org/10.1007/s11051-016-3572-1.
Siegel, D., 1996. Creating Killer Web Sites: The Art of Third-Generation Site Design. Hayden Books.
Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Scient. Bull. 38, 1409–1438.
Šulc, Z., Cibulková, J., Procházka, J., Řezanková, H., 2018. Internal evaluation criteria for categorical data in hierarchical clustering: Optimal number of clusters determination. Adv. Methodol. Stat. (Metodološki Zvezki) 15, 1–20. http://dx.doi.org/10.51936/lxut1974.
Sulc, Z., Cibulkova, J., Rezankova, H., 2022. Nomclust 2.0: an R package for hierarchical clustering of objects characterized by nominal variables. Comput. Statist. 37, 2161–2184. http://dx.doi.org/10.1007/s00180-022-01209-4.
Tabarés, R., 2021. HTML5 and the evolution of HTML; tracing the origins of digital platforms. Technol. Soc. 65, 101529. http://dx.doi.org/10.1016/j.techsoc.2021.101529.
van Duyne, J.A., Hong, J.I., 2003. The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience. Addison-Wesley.
Wilcox, R.R., Erceg-Hurn, D.M., 2012. Comparing two dependent groups via quantiles. J. Appl. Stat. 39, 2655–2664. http://dx.doi.org/10.1080/02664763.2012.724665.
Zhang, H., 2008. Exploring regularity in source code: Software science and Zipf's law. In: 2008 15th Working Conference on Reverse Engineering. IEEE Computer Society, pp. 101–110. http://dx.doi.org/10.1109/WCRE.2008.37.
Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley.

Carlo Bottai is a Postdoctoral Research Fellow in Economic Statistics at the Department of Economics, Management and Statistics, University of Milano–Bicocca. He holds a Ph.D. in Economics and Complexity from the University of Torino and Collegio Carlo Alberto. Previously, he worked at the Eindhoven University of Technology on projects co-funded by the EU's MSCA-COFUND action and the EPO Academic Research Programme. His main research interests are in the Economics and Geography of Innovation, Technology Policy, and Webaugmented Data for Economics.

Lisa Crosato is an Associate Professor at Ca' Foscari University of Venice (Italy). Her research interests are in business and economic statistics, with particular reference to unsupervised and supervised learning, robustness, and the use of website data in the context of SMEs.

Josep Domenech is a professor of Applied Economics at the Universitat Politecnica de Valencia with a multidisciplinary background in Business, Economics, and Computer Engineering. As Principal Investigator of twelve research projects, his work predominantly involves the combination of economics and internet technologies. His research has been published in various journals, including Technological Forecasting and Social Change, Expert Systems with Applications, Knowledge and Information Systems, Economics Letters, Online Information Review, and Performance Evaluation. He frequently collaborates with three Joint Research Centre of the European Commission and the Valencian Institute of Economic Research, focusing on new methodologies for tracking business dynamics.

Marco Guerzoni is an Associate Professor of Applied Economics at DEMS, Bicocca University, and a Research Fellow at BETA, Strasbourg University. He co-founded Despina, the Big Data Lab at the University of Turin. His research spans the management and economics of innovation, technology policy, and advanced data science. Recently, he has focused on the methodological implications of big data and machine learning for business and social sciences. He has published articles in international journals such as Research Policy, the Cambridge Journal of Economics, the Italian Journal of Applied Statistics, and the Journal of Evolutionary Economics.

Caterina Liberati is an Associate Professor in Economic Statistics at the University of Milano-Bicocca. Actually she is a fellow of the Center for European Studies (CefES-DEMS), leading two projects about the usage of unconventional data for monitoring SMEs activity. Her research interests range between Supervised/Unsupervised Learning, Indicators and Internet data.