Mazzoni et al. *EPJ Data Science* (2024) 13:51
https://doi.org/10.1140/epjds/s13688-024-00491-0

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Measuring corporate digital divide through websites: insights from Italian firms

Leonardo Mazzoni[1], Fabio Pinelli[2*] and Massimo Riccaboni[2,3]

*Correspondence:
fabio.pinelli@imtlucca.it
[2]IMT School of Advanced Studies Lucca, Lucca, Italy
Full list of author information is available at the end of the article

Abstract

With the increasing pervasiveness of Information and Communication Technology (ICT) in the fabric of economic activities, the corporate digital divide has become a crucial issue for the assessment of Information Technology (IT) competencies and the digital gap between firms and territories. With little granular data available to measure the phenomenon, most studies have used survey data. To address this empirical gap, we scanned the homepages of 182,705 Italian companies and extracted ten characteristics related to their digital footprint to develop a new index for the corporate digital assessment. Our results show a significant digital divide between Italian companies according to size, sector and geographical location, opening new perspectives for monitoring and data-driven analysis.

Keywords: Digital divide; Website data; Web mining; Corporate big data; Digital business capabilities; Digital transformation

1 Introduction

Digital transformation has proven to be a compelling force capable of shaping the strategic orientation of organisations, driving growth and fostering innovation through the integration of digital technologies and related capabilities [1–3]. With the increasing pervasiveness of Information and Communication Technologies (ICTs) in the fabric of economic activities [4–6], a heterogeneous response of individuals, companies and institutions has taken place, reflected in different rates of adoption and mastery of digital tools. This phenomenon has been analysed under the general umbrella of the digital divide, a concept originally coined to describe the gap in access to IT infrastructure in terms of connection quality and device availability [7–9]. Afterwards, the capillary spread of ICT also contributed to widening the definition of the digital divide beyond infrastructural endowments by including specialised human resources and digital market providers [10, 11]. This shift from simple adoption rate to digital skills and further to related performance has triggered an academic debate on the digital divide at multiple levels such as access, usage and performance [9]. All these layers reflect different and progressive modalities of digital engagement. These modalities have significantly increased the relevance of the digital divide to numerous everyday activities in the business and socio-economic spheres that rely on digital artifacts and/or digital skills [6, 12, 13].

Springer

In this respect, the literature on the digital divide has not developed a homogeneous corpus of analysis. New forms are rapidly emerging, whose specific aspects for industries, companies and territorial levels are receiving increasing attention [14–17]. Within this literature, the corporate digital divide has emerged as a central and burgeoning theme in the management literature, as it has implications for the competencies development during the current digital transformation [15]. The corporate digital divide literature analyses the reasons for differences in ICT adoption across organisations, including the role of geographical and competitive factors [18]. Similar to individuals, organisations that are unable to keep up with continuous innovation cycles often struggle to adapt to new technologies. This phenomenon leads to a digital divide across multiple dimensions, including firm size, sector, location and performance, which in turn help shape the digital divide itself. The corporate digital divide is not primarily due to the adoption of some basic digital technologies that are widely used in companies, but to the level of knowledge, in other words, the link between technology and skills [15]. As yet, the literature on the digital divide has remained relatively silent on how the digital divide works at the corporate level, and there are very few studies that address this specific topic [16]. This can be attributed to the fact that the corporate digital divide is still difficult to observe because of the lack of accounting metrics that can provide information about investments in information technology (IT) or digital capabilities [19, 20].

The analysis of corporate websites [21] offers a possible way out of the scarcity of data on the digital behaviour of companies. Accordingly, corporate websites are the "digital footprint" of organisations and part of the new codified knowledge that is becoming increasingly accessible to researchers and analysts to study corporate behaviour and performance in addition to/alongside more traditional data sources [22–24]. This is because websites are the self-presentation of strategic information to external stakeholders: the marketed products/services, delivery modalities, mission and vision, internal competencies, relationships with other companies and universities, research activities, their location and facilities [22, 25–28].

This makes websites, especially corporate websites, an indispensable open data source, not only to measure the visibility and reputation of the company, but also to investigate the digital competencies that go beyond this [29]. While a company may hire software houses or external IT consultants to develop websites, assessing the quality of these efforts requires specialised knowledge and skills [30, 31]. Accordingly, some studies have exploited the characteristics of websites as broader signals of companies' digital awareness [32, 33]. Previous research, however, has identified a shortcoming of studies analysing the digital behaviour of companies with large website samples [16]. Thanks to the recent development of web scraping techniques [34–36], we aim to contribute to fill this gap by using information from the company website to investigate the corporate digital divide at a large scale, also taking into account different company characteristics (e.g. size, industry, age and geographical context). Specifically, we scraped the websites of 182,705 Italian firms in 2021 and extracted ten characteristics related to technical libraries, performance, security level, speed, links and social media. We then analysed the characteristics of the company websites in combination with the corresponding company balance sheet data. Instead of examining the content of the websites, we focused on a list of "objective" available IT features, following the emerging stream of research linking new IT technologies

to business-related capabilities [37, 38]. We then summarise the IT characteristics of corporate websites in a new Corporate Digital Assessment Index (CoDAI).

We analyse the Italian case because of the well-known sharp socio-economic disparities between northern and southern regions [39]. Moreover, Italy is a unique case in Europe of an industrialised country lagging behind other EU countries in terms of digital readiness.[1]

Our findings highlight a significant corporate digital divide in terms of company characteristics such as size, sector and age, as well as territorial characteristics of company location, and open up new perspectives for monitoring and data-driven analysis. Even when controlling for the effects of the wide band spread, our results persist and pave the way for further empirical research. Considering the low correlation between the IT characteristics of websites, the digital-related variables we used seem to capture specific capabilities and confirm the need to study the different aspects of this multifaceted phenomenon. To address the need for synthetic indicators and improve comparability among firms (and territories), we propose an aggregation of ten features. These features are interpreted and categorised according to a theoretical building process based on the company's digital space(s): technical capabilities, internal organisation, external stakeholder engagement, and digital culture.

The integration of these four sub-indicators results in the Corporate Digital Assessment Index (CoDAI). To show the reliability of our index, we use maps, regressions and a comparative analysis of the top 1% of low and high-performing companies in terms of their CoDAI score.

The article is organised as follows. Section 2 reviews the literature on the economic and digital divide and the role of websites in measuring digital capabilities. Section 3 describes the data collection process and the methodology used. Section 4 presents the results of the empirical investigation of the digital divide phenomenon. Section 5 discusses the findings, highlighting the limitations of this work and providing concluding remarks.

2 Literature review

2.1 The digital divide notion: a corporate perspective

The notion of the digital divide was initially coined to refer to the different levels of ICT penetration among individuals and households [8, 11, 19]. Then, the massive spread of the Internet drew attention beyond the simple adoption rate to another level of ICT use [10]. The transition from the industrial society to the information economy [4] and the recent conceptualisation of "onlife" societies [40] with the new role of ICT as "reality shapers" [6] has increased the relevance of the digital divide as an expression of the socio-economic gap between individuals, companies and territories. Accordingly, the digital divide has evolved into a multifaceted and more sophisticated theme that also encompasses the skills and abilities developed to use technological tools [8, 12, 13]. This has made it possible to shed more light on the competitive factors of the digital economy, as ICTs are firmly embedded in the fabric of socio-economic systems [5, 18]. Moreover, with the pervasiveness of digital transformation over the last decade, the digital divide concept has gained further importance as it can serve as a measure of economic performance [15]. This is particularly relevant when considering the increasing penetration of ICT and the complementarity between physical and digital technologies in the production and consumption of goods and

[1] See the results of the Digital Economy Society Index, available at: https://digital-strategy.ec.europa.eu/en/policies/desi.

services (e.g., cloud computing, artificial intelligence) [41]. Although there is a growing awareness of the digital performance of countries and regions, the concept of the digital divide has rarely been applied to companies as units of analysis.

As recent contributions have shown, digital transformation has a remarkable impact on the structure of companies, as it is a strategic transformation of the organisation and core capabilities of businesses enabled by digital technologies [3]. The rapid pace of technological change over the past decade has challenged the status quo of firms across industries, leading to gaps in digital awareness among managers and employees and an uneven accumulation of digital skills [1, 15]. Internal routines and relationships with customers and suppliers have changed radically in various aspects such as responsiveness, control of production and digital marketing [42–44]. In the realm of IT, digital transformation stands out as a unique development due to the incorporation of advanced technologies such as cloud computing and artificial intelligence (AI), as well as the availability of big data resources. Thanks to their generative and combinatorial properties, these technological advances have allowed companies to open their boundaries to external contributions (e.g. sharing code on GitHub). It is essential to recognise the importance of linking digital transformation to resources, processes and organisational solutions that may be completely new to the firm or require some adaptation of existing ones [42, 45].

Accordingly, the development of new capabilities to favour business model adaptation to the new techno-economic scenario requires a digital sensing activity by the firm [42]. The ubiquity of ICTs requires not only the ownership of specific resources but also the creation of specialised human resources to frame the new possibilities opened by digital affordance property, that is, the creation of endless reconfiguration by the use of the same inputs or a creative (re)combination of them [41, 44]. All in all, while we have quite extensive theoretical evidence that digital transformation has impacted the structure and strategic approach of companies, we still struggle to collect detailed and fine-grained measurements at the firm level [20]. In other words, there is still a lack of data to assess organisations' response to the adoption of digital technologies and whether they have developed adequate digital capabilities. Therefore, the literature is relatively silent on the corporate digital divide between different types of companies operating in different industries and regions.

There is a growing body of literature recognising the complex nature of the paradoxical situation in which there appears to be an abundance of digital technologies without sufficient evidence of this in the data [20, 46, 47]. This can be attributed to various factors, including the typical delay between the introduction of a disruptive technology such as AI and the subsequent increase in productivity. The monopolies of large technology companies in these key technologies can also partly explain this trend. Finally, the lack of suitable statistical measures, such as those of human capital in the IT sector, may contribute to masking the phenomenon.

Against this background, the digital footprints of organisations and individuals have recently become available for empirical analysis thanks to the proliferation of methods such as web scraping [17, 22, 26, 35, 48]. In other words, looking at information published on the Internet from the perspective of digital signalling theory [32, 33] makes it possible to relate the digital artefacts produced (such as the characteristics of a company website) to a set of underlying digital and managerial capabilities of firms [49]. Recent studies have exploited this source of information, investigating different aspects related to internal strategic fac-

tors, innovative behaviour and relational aspects of the firm [22, 26, 36, 48, 50–54]. Further details on this recent trend are presented in Sect. 2.3.

2.2 The use of websites to measure the digital footprint of firms

The widespread use of websites by companies, including small and medium-sized enterprises (SMEs), represents an interesting potential source of information to bridge the gap between the need to assess the digital performance of firms and the lack of detailed indicators. This is possible due to the strategic role that websites play in the information and knowledge economy.

A website is a digital tool capable of reducing information asymmetry between two parties and facilitating corporate operations (delivery, customer service, internationalisation) [55]. Its maintenance, use, and development imply some extra costs for the firms. This makes a website a *"near-costless measure of marketing"* [17] and, more generally, an effective proxy for capturing the digital footprint of economic actors [21, 22, 48, 56].

The massive use of websites offers some advantages in data collection compared to traditional methods due to their (i) "unobtrusiveness", (ii) accessibility, (iii) temporal frequency, (iv) granularity and (v) coverage [21, 22, 24, 26, 57–59].

Unobtrusiveness derives from the possibility of collecting the reported information directly, without the need to directly involve a company or a set of companies (e.g. in surveys). This saves time in collecting information and is less variable than a traditional survey. Accessibility results from the open-access nature of websites as sources of information and the reproducibility of the analysis. Websites, especially corporate websites, are updated much more frequently for business reasons than traditional information sources such as surveys (temporal frequency).

In addition, the information reported on a website has a much more customisable degree of granularity than traditional collection methods (albeit with increasing post-processing work). Finally, the massive analysis of websites has the potential to extend coverage to almost the entire company population (without non-response), overcoming the limitations of traditional survey methods based on the selection of a representative sample, increasing the granularity of industrial classification [60]. This makes it possible to agnostically recognise common and recurring characteristics and identify hidden peculiarities that remain under the tip of the iceberg. A first key signal is the identification of dead companies by checking homepage URLs and capturing "zombie firms" via very few working hyperlinks [60].

However, the available company websites are a convenience sample. In addition, extracting data directly from websites may pose some problems regarding the reliability of the information, even though companies are more likely to receive negative feedback from the customers and/or business partners they interact with if they exaggerate or mystify reality. Therefore, the general trend in the literature is to consider this information as truthful or at least verifiable [28]. Regardless of the availability of websites, there are some doubts about the typology of companies that can be found on the internet. For example, market-oriented companies (e.g., Business to Client – B2C) are most likely to be found in this group.

However, we have less evidence of subcontractors or intermediaries (Business to Business – B2B) whose activities may remain in the shadows to respect the will of their end

customers (e.g. preserving trade secrets or preventing competition) [28]. In this case, encouraging signals come from the demand for certificates (such as the International Organisation for Standardisation – ISO) by value chain leaders, who generally urge subcontractors and suppliers to show them as digital "business cards". Nevertheless, the available company websites are a non-representative sample to study company dynamics, and an appropriate process can be put in place to rebalance the available information across geographical areas and sectors.

The difficulties in processing website data are related to its reliability and data mining techniques. There are still some technical problems with the homogeneous processing of website data, which by its nature is organised in different formats, has different dimensionality and consists of non-textual parts (e.g. graphics and images) [22, 50, 56].

2.3 The use of webscraping at the corporate level

Initial scholarly works dealing with corporate web scraping focused primarily on analysing relatively small samples, with an emphasis on textual content. These efforts were mainly aimed at assessing the reliability of websites as alternative sources of information in relation to other innovation indicators (such as the research and development rate), for the cases of green technologies [22] and nanotechnologies [50]. Both studies underline the complementarity between traditional sources of information and information from company websites, the latter being considered extremely useful for assessing the commercialization phase of the innovation process. Further research has been conducted to analyse the company's external collaborations, specifically evaluating different types of relationships, considering university, industry and government as potential partners, by counting keywords for three consecutive years and identifying statistical evidence of the impact on sales growth. [26]. Along the same line, corporate websites have also been used to operationalise management concepts such as "dynamic capabilities" at the micro level. Through the application of topic modeling, it has been demonstrated that changes to a website can reflect a strategic change within the company [36].

Despite the increasing popularity, research on these new data sources is very fragmented, with many exploratory analyses and elements that may not be applicable in all situations, as previously emphasised by [61].

However, things seem to have changed with the recent advent of massive corporate web scraping in social science, which allows researchers to access huge datasets that are not limited to textual content but also include the structures of websites, especially hyperlinks. This approach facilitates comprehensive comparisons between information from websites and balance sheets or national business registers [48, 51–54, 62]. Early studies of this kind have examined the distribution of URLs on a large scale and looked at the potential websites that serve as complementary sources of information to the Big Data of statistical offices or the repositories of private companies [48]. Various topics were investigated, such as the emergence of inter-firm relationships and the proximity of companies in terms of organisational forms and strategies, and these data were integrated with georeferenced company addresses using machine learning algorithms to validate potential classifications of company profiles [51]. Knowledge flows between companies were also studied on a large scale, looking at the overlapping characteristics of social media and website hyperlinks and uncovering different structural networks between the two types [62]. The use of hyperlinks on a large scale deepened the theoretical understanding of the

connections that the geography of innovation approaches might draw between the digital spaces and the real business connections, building on a framework of digital layers of firms, analysing different types of proximity (cognitive, geographical and organisational) [52, 53] thus confirming the idea originally proposed by [51]. More recently, improvements have been proposed to refine the methods for identifying the most relevant information to be extracted from websites. This includes the use of enterprise-level AI in three European countries, combining website texts, hyperlinks and company metadata with machine learning techniques to improve the accuracy of information extraction [54].

3 Data collection and methodology

3.1 Data collection

We use the Advanced Industrial Document Automator (AIDA) database to collect Italian company data. The AIDA database contains the balance sheet information of companies provided by Bureau van Dijk.[2] On a total of 2,180,820 Italian firms available in the dataset, we found 450,348 firms with URL website information available, 21% of the total.

We implemented in `python` our crawler script using a combination of different well-known libraries: `requests`,[3] `scrapy`,[4] `beautifulsoup`,[5] `trafilatura`,[6] `builtwith`.[7] Moreover, we collect results for each website querying Google Lighthouse framework[8] and external services to assess the security of the websites.Google Lighthouse is an automated open-source tool for improving the quality of websites. You can apply it to any website. It evaluates performance, accessibility, progressive web apps, search engine optimisation (SEO), etc. For more details, see Sect. 3.2. The framework used to evaluate the security of websites is Security Header.[9] This is a cybersecurity company that focuses on identifying vulnerabilities in web applications and Application Programming Interfaces (APIs) with its Dynamic Application Security Testing (DAST) tools. The service offers an evaluation system that analyses HTTP response headers.

The crawler tool has been executed in a multiprocess server. Each website request is performed by a single process executed in a dedicated CPU. The server bandwidth was up to 1 Gigabit.[10]

The crawler extracts the content and the corresponding features only from the homepage of the respective company website. This choice may mean that some information found on other subpages, such as dedicated social media pages, is not taken into account, which could be a limitation of our work. However, our approach is in line with the current literature on corporate web mining, which considers the information on homepages as the priorities chosen by companies [63]. Moreover, this eases the computational, time and space requirements for obtaining the data needed for our purposes, which focus on technical features found mainly on homepages.

[2] https://www.bvdinfo.com/it-it/le-nostre-soluzioni/dati/nazionali/aida.

[3] docs.python-requests.org.

[4] https://scrapy.org/.

[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[6] https://trafilatura.readthedocs.io/en/latest/.

[7] https://pypi.org/project/builtwith/.

[8] https://developers.google.com/web/tools/lighthouse.

[9] https://securityheaders.com/about/.

[10] https://www.garr.it/en/garr-en.

Next, we stored the results of the crawling activities on a MongoDB[11] instance, a document-oriented NOSQL database that gave us the flexibility we needed to store and update the collected information. The characteristics of each company were stored in a single MongoDB document.

The crawling activity was performed from January 2021 to March 2021, with a further update in September 2021. As a result, we obtained valid content for 347,010 companies. For the other URLs, the crawler encountered two possible issues: timeout errors and an HTTP code not equal to 200. Therefore, we obtained a valid set of features for 77% of the original sample. Together with the homepage URL, we included information on the company characteristics, such as geographical localisation at different levels: regional (NUTS-2[12]), provincial (NUTS-3) and municipal level. Further characteristics of the firm are the main sector of activity (NACE REV.2 digit[13]), the age and the size (micro, small, medium-sized or large firm according to the number of employees). Data on company characteristics are not available on AIDA for all companies in our sample. Therefore, when we merged the website information with the firm characteristics, we obtained a final sample of 182,705 observations (i.e. firms).

3.2 Methodology

We interpreted the economic meaning of each extracted feature based on the existing literature [48, 52, 57, 64–69]. Our review of the literature reveals a gap in the analysis of the digital strategies of companies using web mining. While the business and management literature has mainly focused on social media presence and hyperlinks as indicators of external relationships, the information systems literature has mainly analysed companies' technical frameworks and security levels. To fully grasp firms' digital strategies, we believe it is necessary to integrate both perspectives. As we will discuss further in Sect. 4.2, this interdisciplinary approach enables a more nuanced understanding of digital transformation at the micro level. For the present analysis, we considered the directly usable information from homepages and disregarded text, images and code analysis.[14]

We considered ten features as follows:

- *the length of the URL*. Short URLs are easier to remember and represent a sign of cleanliness and user-centricity. Moreover, short URLs (also including strategic keywords) are more likely to be discovered by SEO optimisation, and thus factors wisely strategised by firms [67, 70]. This variable considers the number of characters included in the URL without the HTTP:// prefix;
- *Social media presence: Facebook, LinkedIn and Instagram*. We consider three popular social media (Facebook, LinkedIn and Instagram), checking for the absence/presence of links to those social media [66, 71, 72]. Indeed, social media are valuable tools for the implementation of digital marketing strategies. These are dummy variables that count the presence of the respective social media links;
- *The quality of internal links*. The presence of unique inner links can be interpreted as a sign of good navigability to increase the likelihood that users will stay longer on the

[11] https://www.mongodb.com/.

[12] https://ec.europa.eu/eurostat/web/nuts/background.

[13] https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-ra-07-015.

[14] This choice is motivated by the goal of focusing on features that do not require advanced post-processing.

websites as they are correctly directed to content that matches their preferences [68]. This feature is computed considering the number of HREF HTML tags with the same domain reported in the URL or with a relative subpath;

- *The quality of external links*. The higher presence of unique outer links can be interpreted as a strong sign of stakeholder engagement and embeddedness in the digital business ecosystem. Accordingly, recent literature has used external hyperlinks to measure the engagement of companies in business relationships [52, 69]. For the external links, we count the number of HREF elements in the HTML file with a domain other than the one specified in the URL;
- *Quality of the technical frameworks*. Modern web development standards provide a better user experience and reflect more technical competencies. Google Lighthouse collects this metric. Although it is still little used in research projects, recent work has emphasized the usefulness of Google Lighthouse for evaluating the quality of websites, also considering that it is an open source tool [73, 74]. For this indicator, Google Lighthouse provides a continuous value ranging from 0 to 1;
- *Request access time*. A webpage with a short loading time implies good speed and a good level of usability. Fast loading times can also indicate the use of cutting-edge technological stack development [69, 75]. This variable specifies the number of seconds it takes to load the web page;
- *Website's age*. We estimated the age of each website by checking the first year of its presence in the Wayback Machine Archive.[15] Older websites are interpreted as indicative of a consolidated digital history of the firm [69], also in terms of go-to-market strategies for export [65]. In this case, we count the number of years between the first appearance in the Wayback Machine Archive and 2021, the year of analysis;
- *Website security*. A high-level security header represents a proxy of the awareness of the risks of cyberattacks [76–80]. With the increasing number of threats in web-based systems, we need to enforce security policies from server to client. This is mainly about enabling security options for the HTTP responses to prevent man-in-the-middle attacks or code injections. Nowadays, this is a crucial strategic aspect for companies. We used the Security Headers platform to assess the security of each website. The service used returns a discrete rating (e.g. A, B, etc.) of the website and we assign them corresponding numbers.

In the following, we conduct an exploratory analysis of the extracted characteristics for our sample of 182,705 companies.

Figure 1 describes how our final dataset is distributed along three firm characteristics: industry, size, and age. Above we show the number of enterprises for the two-digit code of NACE Rev. 2. In our sample, most firms belong to the categories manufacturing (C) and wholesale and retail trade (G). Note that not all industries were included in this list: we only selected the 10 most important industrial categories. The second plot shows the size distribution of companies. Since the size distribution of Italian firms is skewed to the left, as expected, most of the firms in our dataset are micro and small enterprises, i.e. firms with less than 50 employees. At the bottom of Fig. 1 we show the age of the firms in 5-year

[15] https://archive.org/web/.

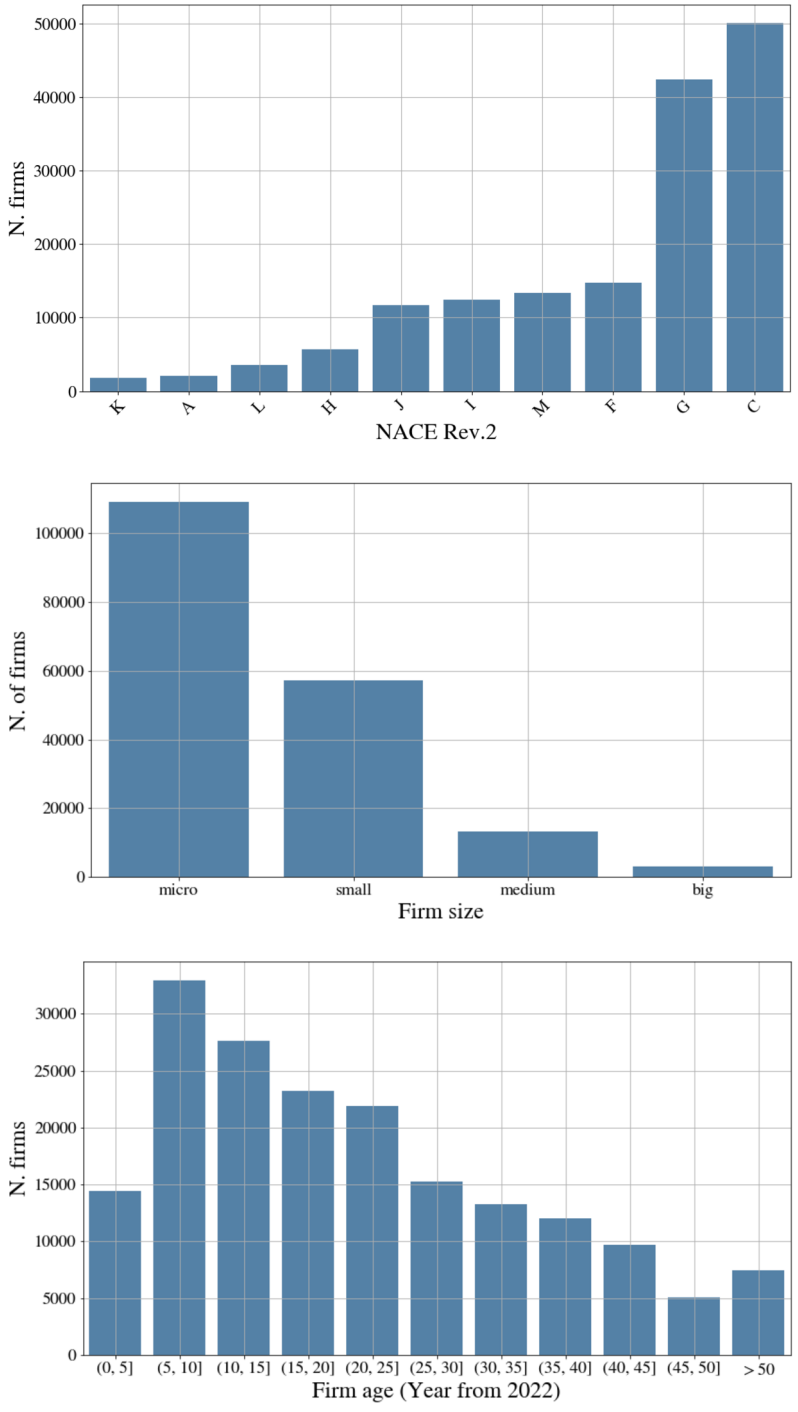

Figure 1 Key characteristics of selected Italian companies with a known physical and digital presence. Top: Number of companies analysed, grouped by NACE 2 Digit code (sector A: Agriculture, forestry and fishing; sector C: Manufacturing; sector F: Construction; sector G: Wholesale and retail trade; sector H: Transportation; sector I: Accommodation and food service; sector J: Information and communication; sector K: Financial and insurance activities; sector L: Real estate activities; sector M: Professional, scientific & technical activities). Middle: Size distribution. Bottom: Age distribution

increments. The results show a higher concentration of relatively young companies aged 5 to 25 years.

To assess the representativeness of the sample, we compared the extracted data with the general composition of our original sample of Italian companies in AIDA. To ensure a meaningful comparison, we filtered our original database to include only active companies with at least one employee. This process ensures consistency with our final dataset, which only includes companies with active and functioning websites. This excludes inactive companies and sole proprietorships that rarely use websites for business purposes. As far as industry is concerned, the representativeness of our sample varied between 15% in NACE sector A (agriculture) and 40% in sector C (manufacturing). We also found good values for sector J (information and communication) (35%) (see Table 9) - column 2 in the appendix). We also assess the representativeness by firm size. Our sample covers captured 73% of the total number of large companies, with decreasing coverage for smaller companies: 54% of medium firms, 37% of small firms and only 21% of micro firms (see Table 10 - column 2 in the Appendix).

To triangulate our data, we have also compared our statistics with those of the Italian statistical office (ISTAT), which provides information on the distribution of companies by sector and size.

In terms of industry, we obtained a representativeness ranging from 5% of NACE sector I (food and accommodation) to 28% of sector J (information and communication), with a good level (22%) also for sector C (manufacturing) (ISTAT, 2020).[16]

In terms of size, we were able to capture 75% of the total number of large firms, with decreasing coverage for smaller firms: 56% of medium-sized firms, 31% of small firms and only 8% of micro firms (ISTAT, 2020) (similar values to our original database).

Since our original database from AIDA did not contain information on the age of the companies, we rely on ORBIS.[17] After filtering for active companies in Italy, our sample represents 7% of the total ORBIS sample[18] and becomes more representative the further back in time we go. The largest percentage (33%) of companies in our sample have been in operation for over 50 years. The percentages decrease from (27%) to (23%) for firms that are 40 years old, (18%) to (15%) for firms that are 30 years old, (11%) to (12%) for firms that are 20 years old, (10%) to (8%) for firms that are 10 years old, and (6%) to (2%) for firms that are less than 10 years old. This gradual decline can be attributed to the fact that younger may be waiting longer to invest in websites as they prioritise building a larger market share given the financial constraints many startups face in their first years of existence (ORBIS, 2021).

Figure 2 maps the features extracted by means of the crawling activity at NUTS-3 regions (descriptive statistics are available in Table 1).

We report values corresponding to the three quantiles, i.e. 33%, 66% and 100%, where a darker colour indicates higher values. Seven out of the ten extracted features, namely URL length, LinkedIn, quality of the internal links, quality of the technical frameworks, request

[16] The percentage deviations compared to our primary database can be attributed firstly to the broader range of companies covered by ISTAT compared to the AIDA database. Secondly, not all firm types are required by law to disclose their financial statements and are therefore covered by private platforms such as AIDA.

[17] ORBIS is another product of the Bureau van Dijk information database, but with global coverage.

[18] Variations in the percentage can be attributed to the broader coverage of ORBIS compared to AIDA. Accordingly, our crawled sample achieves a representativeness of 26 compared to the active companies in AIDA.

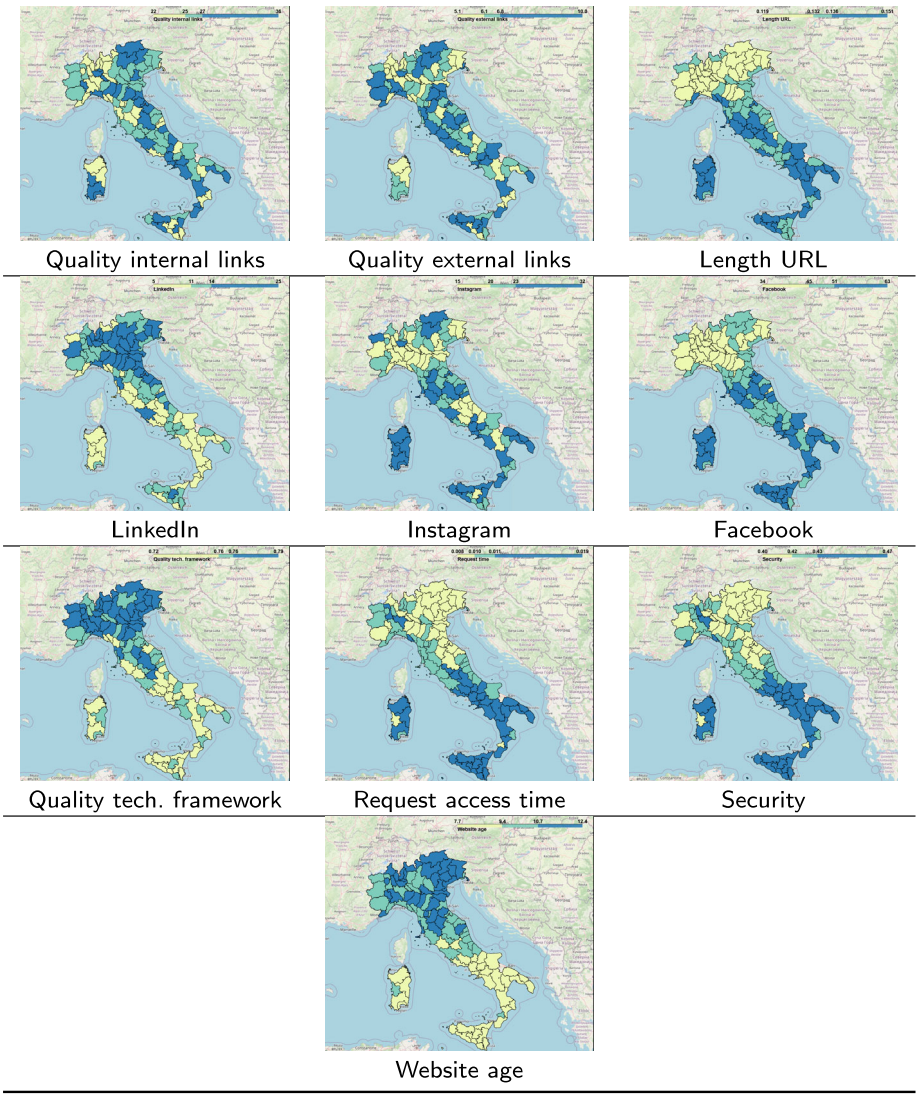

Figure 2 The spatial distribution of corporate website features in Italy, at the NUTS-3 regions level. Features are classified in three quantiles: from low (yellow) to high (dark blue) levels

Table 1 Descriptive statistics of the features of the corporate websites of interest

	min	max	mean	std
Quality internal links	0.000	4885.000	27.473	48.921
Quality external links	0.000	15,814.000	6.942	42.228
Quality tech. framework	0.380	1.000	0.837	0.103
Length URL	4.000	105.000	19.214	5.436
Facebook	0.000	1.000	0.461	0.498
Instagram	0.000	1.000	0.219	0.414
LinkedIn	0.000	1.000	0.159	0.366
Website age	0.000	25.000	10.308	7.455
Request time	0.127	502.262	5.265	9.418
Security	0.000	15.000	6.268	1.941

access time, website age and website security) exhibit clear spatial differences across Italian NUTS-3 regions. More scattered territorial patterns are reported for external links. This spatial distribution of websites might correspond to wider adoption of digital tools for Northern enterprises w.r.t. firms located in the South, thus indicating a more rooted digital mindset. When analysing the use of social media, we notice some differences between LinkedIn, Facebook and Instagram. Firms in northern Italy are more likely to use LinkedIn than the other two social media (Facebook and Instagram), which are more popular in the central and southern regions. Instagram shows greater granularity than Facebook, which could depend on the sectoral composition of local production systems. For example, companies in tourist-oriented areas (such as the coastal regions, Tuscany, Rome and the main islands) use Instagram more intensively than the other social media included in the analysis. Social media are therefore of particular interest due to their different levels of penetration, which suggests a specific interpretation of their use. For example, the presence of Facebook can be understood as the spread of a broader marketing culture characterised by a generalist approach. On the other hand, Instagram can be understood as the dissemination of a specialised marketing culture, mainly used by certain B2C sectors that use visual content, including tourism, gastronomy and the cultural sector. LinkedIn can be interpreted as propagating a culture that focuses on human resources and serves as an indicator of mature labour markets where the use of professional networks can significantly support concrete business efforts.

In general, Table 2 shows a low level of correlation between the variables, with the only exceptions being the correlation between Security and Request time (0.769) and the correlation between Facebook and Instagram (0.524). While the reason for the positive correlation between social media is intuitive, the correlation between security and request time might depend on the fact that low-quality websites tend to underperform both in terms of both vulnerability and access time. This result supports our approach of collecting different indicators to capture a multifaceted phenomenon, such as the corporate digital divide.

In general, our explorative data analysis allows us to validate some characteristics of the enterprises included in our dataset built through the crawler. As a result, we rely on 182,705 firms out of the initial sample (composed of 347,010 firms) because a consistent number of firms showed missing values for employment, a useful indicator to categorise the size of each firm in different groups.

To analyse the relationship between firms' characteristics and corporate websites, we run a set of simple OLS (Ordinary Least Squared) regressions with the following specifications:

$$y = \alpha + \beta_1 x_f + \beta_2 x_t + \epsilon \tag{1}$$

where y represents one of the ten digital dimensions (i.e., one of the relevant features of the corporate websites), and x_f is the vector that represents the considered characteristics of the firm: age, industry, and size. In each regression, we include a set of controls for the area in which the firm is located, represented by x_t. The set of control variables includes age, location (regional and urban areas) and presence of highspeed broadband. This is because, beyond the specific patterns detectable by the traditional characteristics of the firm, the notion of the corporate digital divide can be influenced (and captured) also by the experience of the firm and the place where the firm is located (e.g. a common available

Table 2 Correlation matrix between the ten corporate website features

	Quality internal links	Quality external links	Quality tech. framework	Instagram	LinkedIn	Facebook	Website age	Security	Request time	Length URL
Quality internal links	1.000***	0.018***	−0.042***	0.151***	0.104***	0.179***	0.078***	−0.053***	−0.009***	−0.023***
Quality external links	0.018***	1.000***	−0.005*	0.073***	0.062***	0.087***	0.022***	−0.013***	0.001	−0.003
Quality tech. framework	−0.042***	−0.005*	1.000***	0.038***	0.055***	0.008***	−0.038***	−0.055***	−0.011***	−0.026***
Instagram	0.151***	0.073***	0.038***	1.000***	0.178***	0.524***	0.007**	−0.032***	0.006**	−0.004
LinkedIn	0.104***	0.062***	0.055***	0.178***	1.000***	0.293***	0.080***	−0.074***	−0.040***	−0.096***
Facebook	0.179***	0.087***	0.008***	0.524***	0.293***	1.000***	0.002	−0.033***	0.001	0.023***
Website age	0.078***	0.022***	−0.038***	0.007**	0.080***	0.002	1.000***	−0.056***	−0.060***	−0.167***
Security	−0.053***	−0.013***	−0.055***	−0.032***	−0.074***	−0.033***	−0.056***	1.000***	0.769***	0.072***
Request time	−0.009***	0.001	−0.011***	0.006**	−0.040***	0.001	−0.060***	0.769***	1.000***	0.075***
Length URL	−0.023***	−0.003	−0.026***	−0.004	−0.096***	0.023***	−0.167***	0.072***	0.075***	1.000***

IT infrastructure or the diffusion of digital means to run a business activity varies between urban and rural areas and from the North to the South). In this regard, to avoid possible misspecification of the model, we control for the impact of the wide band. This is because some features of corporate websites, such as the request access time, might depend on the wide band coverage of the area where the firm is located. The wide band data are obtained as open data from the Italian Authority for Communications Guarantees.[19] The data represent, for each Italian census area, the ratio of households reached by broadband with speed greater than 30mbs over the total number of households in that area.

4 Results

4.1 Econometric analysis

In this section, we discuss the results of the regression analysis for the ten digital-related characteristics of company websites. Table 3 summarises the results for the analysis of the digital divide across different dimensions of Italian firms: size, sector, age and location.

Micro firms show a negative and significant effect for unique internal links, while medium and large companies pay more attention to internal navigability. In particular, external links are not significant for micro companies, which underlines an unequal weighting of links and accessibility for internal stakeholders. In contrast, medium and large companies show a positive and statistically significant effect in this regard, indicating a conscious effort to improve connectivity and accessibility for external stakeholders. When examining the quality of the technical framework, a positive and statistically significant effect can only be seen in medium-sized companies, while no influence can be found in larger companies. This unexpected result suggests a nuanced interpretation that indicates a decreasing marginal benefit of state-of-the-art libraries in large companies.

The social media variables have a significant effect across all dimensions, with a negative sign for microenterprises and a progressive value of the magnitude effect, with the highest value for big firms.

Age, which is representative of digital experience, is of varying importance depending on the size of the company. Particularly noteworthy is the increased relevance in large companies, suggesting that they recognised the importance of website development many years earlier on average (in line with the results of [81]).

Security, speed (request time) and URL length should be interpreted in the opposite direction, as lower request times and shorter URLs improve the accessibility of corporate websites. Lower access times and shorter URLs significantly improve the accessibility of corporate websites and emphasise the critical role of efficiency and conciseness in driving online engagement.

We also analyse the impact of the corporate digital divide across NUTS-2 digit sectors. We take as a sector of reference in terms of quality, Information and Communication (J) (in line with results of [63]), for its direct involvement in the coding process or the marked attention to the quality of websites, as a fundamental tool to communicate with customers. A positive and significant effect is found in the wholesale and retail trade (G) for the unique internal and external links, underscoring the criticality of both internal navigability and external connections, particularly in the context of business-to-consumer (B2C) activities.

[19] https://maps.agcom.it/.

Table 3 Regression results considering as dependent variables website feature extracted. Notice that for the features Instagram, Linkedin and Facebook, we use the Logit model, and consequently, we report the relative Pseudo R-squared

	Quality internal links	Quality external links	Quality tech. framework	Instagram	LinkedIn	Facebook	Website age	Security	Request time	Length URL
Constant	0.0054*** (0.0000)	0.0004*** (0.0000)	0.7708*** (0.0020)	−0.9649*** (0.0270)	−1.8102*** (0.0320)	0.3756*** (0.0220)	0.4058*** (0.0030)	0.4150*** (0.0010)	0.0097*** (0.0000)	0.1433*** (0.0000)
Micro firm	−0.0010*** (0.0001)	−0.0001*** (0.0000)	−0.0046*** (0.0010)	−0.2365*** (0.0130)	−0.5756*** (0.0160)	−0.1358*** (0.0110)	−0.0934*** (0.0020)	0.0124*** (0.0010)	0.0014*** (0.0000)	0.0016*** (0.0000)
Medium firm	0.0019*** (0.0001)	0.0002*** (0.0000)	0.0061*** (0.0010)	0.3223*** (0.0230)	0.6536*** (0.0230)	0.1959*** (0.0200)	0.0737*** (0.0030)	−0.0132*** (0.0010)	−0.0007*** (0.0000)	−0.0022*** (0.0000)
Big firm	0.0048*** (0.0000)	0.0003*** (0.0000)	0.0033 (0.0030)	0.5051*** (0.0420)	1.1129*** (0.0400)	0.3389*** (0.0380)	0.1092*** (0.0050)	−0.0293*** (0.0020)	−0.0006 (0.0000)	−0.0024** (0.0010)
Agriculture, forestry and fishing (A)	−0.0006** (0.0000)	−0.0000 (0.0001)	0.0010 (0.0030)	0.6944*** (0.0490)	−0.5749*** (0.0900)	0.2595*** (0.0470)	0.0312*** (0.0070)	−0.0071* (0.0030)	−0.0004 (0.0000)	−0.0033** (0.0010)
Manufacturing (C)	−0.0004*** (0.0001)	−0.0001** (0.0000)	−0.0010 (0.0010)	−0.1339*** (0.0200)	0.2718*** (0.0230)	−0.4942*** (0.0160)	0.0985*** (0.0020)	−0.0023* (0.0010)	−0.0004** (0.0000)	−0.0151*** (0.0000)
Construction (F)	−0.0013*** (0.0000)	−0.0001*** (0.0000)	−0.0021 (0.0020)	−0.8037*** (0.0310)	−0.1939*** (0.0330)	−0.6532*** (0.0220)	−0.0117*** (0.0030)	0.0103*** (0.0010)	0.0012*** (0.0000)	−0.0035*** (0.0000)
Wholesale and retail trade (G)	0.0028*** (0.0001)	0.0001*** (0.0000)	−0.0055*** (0.0010)	0.2998*** (0.0190)	−0.0021 (0.0240)	0.0325* (0.0160)	0.0610*** (0.0020)	0.0013 (0.0010)	0.0009*** (0.0000)	−0.0101*** (0.0000)
Transportation (H)	−0.0017*** (0.0000)	−0.0001 (0.0000)	−0.0095*** (0.0020)	−0.7632*** (0.0440)	−0.2870*** (0.0460)	−0.7353*** (0.0310)	0.0225*** (0.0040)	0.0066*** (0.0020)	0.0004 (0.0000)	−0.0028*** (0.0010)
Accomodation and food service (I)	−0.0007*** (0.0000)	0.0000 (0.0000)	−0.0146*** (0.0020)	0.9041*** (0.0240)	−1.4107*** (0.0520)	0.7412*** (0.0230)	0.0455*** (0.0030)	−0.0024 (0.0010)	−0.0008*** (0.0000)	0.0086*** (0.0010)
Information and communication (J)	0.0007*** (0.0000)	0.0002*** (0.0000)	0.0052** (0.0020)	−0.2616*** (0.0290)	1.2750*** (0.0280)	−0.1325*** (0.0230)	0.0700*** (0.0030)	−0.0188*** (0.0010)	−0.0023*** (0.0000)	−0.0268*** (0.0010)
Financial and insurance activities (K)	−0.0005* (0.0000)	−0.0000 (0.0001)	0.0034 (0.0040)	−0.9819*** (0.0800)	0.7227*** (0.0590)	−0.7089*** (0.0510)	0.0171* (0.0070)	−0.0172*** (0.0030)	−0.0025*** (0.0000)	−0.0086*** (0.0010)

Table 3 (*Continued*)

	Quality internal links	Quality external links	Quality tech. framework	Instagram	LinkedIn	Facebook	Website age	Security	Request time	Length URL
Real estate activities (L)	0.0006 (0.0000)	0.0001* (0.0000)	–0.0110*** (0.0030)	0.1682*** (0.0430)	0.2634*** (0.0520)	0.0364 (0.0360)	0.0844*** (0.0050)	–0.0021 (0.0020)	–0.0011 (0.0000)	0.0010 (0.0010)
Professional, scientific & technical activities (M)	–0.0000 (0.0000)	0.0000 (0.0000)	0.0080*** (0.0020)	–0.2119*** (0.0270)	1.1374*** (0.0270)	–0.3845*** (0.0220)	0.0394*** (0.0030)	–0.0136*** (0.0010)	–0.0017*** (0.0000)	–0.0179*** (0.0000)
Urban area	0.0003*** (0.0001)	0.0000** (0.0000)	0.0006 (0.0010)	0.1302*** (0.0130)	0.1695*** (0.0150)	0.0521*** (0.0110)	0.0027 (0.0020)	–0.0023 (0.0010)	–0.0001 (0.0001)	–0.0022*** (0.0000)
North	0.0001 (0.0001)	–0.0000 (0.0000)	0.0106*** (0.0010)	–0.1020*** (0.0150)	0.2878*** (0.0180)	–0.2006*** (0.0120)	0.0190*** (0.0020)	–0.0082*** (0.0010)	–0.0009*** (0.0000)	–0.0053*** (0.0000)
South	0.0003*** (0.0001)	–0.0000 (0.0000)	–0.0079*** (0.0010)	0.0674*** (0.0180)	–0.1089*** (0.0230)	0.1618*** (0.0150)	–0.0454*** (0.0020)	0.0137*** (0.0010)	0.0020*** (0.0000)	0.0010** (0.0000)
Firm age	–0.0000*** (0.0000)	0.0000 (0.0000)	–0.0004*** (0.0000)	–0.0125*** (0.0000)	–0.0091*** (0.0000)	–0.0103*** (0.0000)		0.0001** (0.0000)	–0.0000 (0.0000)	0.0000*** (0.0000)
Wide band	0.0002** (0.0001)	0.0000 (0.0000)	0.0012 (0.0010)	0.0431* (0.0170)	0.1311*** (0.0200)	0.0146 (0.0140)	–0.0039* (0.0020)	0.0016 (0.0010)	0.0002 (0.0000)	–0.0003 (0.0000)
R-squared: Adj. R-squared:	0.028 0.027	0.002 0.002	0.005 0.005	0.037	0.078	0.035	0.071 0.071	0.013 0.013	0.009 0.009	0.041 0.041
N. of Observations	182,705	182,705	182,705	182,705	182,705	182,705	182,705	182,705	182,705	182,705

Concerning best practice, a positive and significant effect is detected in the sectors Professional, Scientific and Technical activities (M) and Information and Communication (J), supporting the idea that, on average, firms in those sectors are better informed about the importance of high-quality technical libraries. In line with their business functions, Social Media seems to have different behaviours, confirming the exploratory mapping reported in Sect. 3. Accordingly, Instagram and Facebook seem to play an important role in Accommodation and Food Service Activities (sector I), as free channels to engage with tourism. LinkedIn, more oriented to attracting highly skilled profiles, seems to have a bigger impact on sectors with highly specialised knowledge, such as the sector of Professional, scientific & technical activities (M).

Similarly, in this context, the interpretations of security and request time should be reversed. Notably, Information and Communication (J) and Professional, Scientific and Technical activities (M) exhibit the highest level of performance in our analysis.

The geography of the corporate digital divide also presents sharp evidence that urban contexts offer a higher marginal utility in the use of digital tools for firms operating in dynamic and densely populated economic systems [53]. In such environments, firms have the advantageous opportunity to leverage the network effect [82]. Geographical location emerges as a significant determinant, shedding light on how the digital divide represents a dual facet of territorial economic development. Disparities are evident through the distinct behaviours exhibited by North and South dummy variables. This supports the idea that more industrialised regions, prevalently situated in the north, are conducive environments for adopting digital strategies and the relative adoption of skills and investment, particularly when compared to their counterparts in the south. Although beyond this paper's goal, spatial correlation cannot be neglected when examining geographical level. This means that spillover effects may occur between firms connected through physical and digital relationships [53]. Additionally, it is important to note that external providers may create websites, potentially leading to a concentration of skills in specific areas where the same provider may have a wide portfolio of clients.

In all the regressions, we control for the effect of the wide band, operating under the assumption that a robust web infrastructure might overshadow certain digital web proxies. Disentangling the infrastructural effect is not a trivial task. Accordingly, websites can rely on external hosts or local servers. In the former scenario, the reactivity of the different web pages can be related to the quality of the website. Conversely, in the latter case, we may observe extensive broadband coverage coexisting with websites of subpar quality, or on the other side limited coverage within high-quality websites. Overall, our analysis reveals only four instances of the statistically significant impact of the wide band out of ten regressions conducted. This outcome assumes great significance as it affirms the efficacy of our approach in capturing the digital footprint of firms, transcending the influence of shared web infrastructure quality, through the meticulous collection of diverse features of corporate websites.

4.2 Corporate digital assessment index

Our findings show a multifaceted composite picture that reveals how the digital divide is present in different dimensions of business activity. One of the main challenges in applying such a data-driven approach is that there is no clearly defined theoretical framework for categorising proxies into distinct digital capabilities.

Evaluating the quality of websites is a complex undertaking because there are numerous unstructured data elements that need to be contextualised within a logical theoretical framework to understand their potential economic significance. Furthermore, given the high degree of customisation, it is not trivial to develop synthetic measures to capture the diverse attributes of websites, making it difficult to establish accountable criteria [68].

Within this literature, [57] provide one of the first single index, called Web Assessment Index (WAI).[20]

Although the WAI paved the road to build a composite index for websites, it only focused on universities' websites. Therefore, the literature still lacks specific composite indexes for firms. After reviewing the recent literature on firm-level behaviour on the web [48, 51, 52, 65–69], we have developed a theoretical framework that integrates both managerial and information systems literature to analyse the digital readiness of firms. Specifically, we borrow and adapt from the well-known Technology, Organisation, Environment (TOE) framework (for a similar approach applied to firms' websites, see [85]). To this consolidated framework, we add a fourth layer that focuses on digital culture to the model.

Elaborating on the TOE framework, we interpreted and classified the extracted features into four different aspects related to the digital strategy of the firm:

1. stakeholder engagement
2. technical capabilities
3. internal organisation
4. digital culture

Stakeholder engagement includes the firm's capabilities to establish meaningful links and connections with external stakeholders, including customers and suppliers. This indicator tries to capture firms' embeddedness in the competitive external environment. Technical capabilities encompass the firm's proficiency in building a state-of-the-art digital framework focusing on elements such as libraries, security and speed. With this indicator, we aimed to measure the firm's technology stack. The internal organisation represents the firm's aptitude to effectively and purposefully orchestrate internal information flows. With this indicator, we aimed to proxy Organisational structure. Digital Culture stands for the historical trajectory in which the firm is inserted, reflecting its awareness of the strategic significance of the web.

Aggregating these four indicators, we propose a synthetic index to capture the corporate digital divide with a single dimension.

To build our index, we normalise the variables with a *MinMax* technique, inverting the scale of the values where necessary (i.e., the length of the URL, Facebook, request time, and security header) for interpretability. Then, we calculated the Corporate Digital Assessment Index (CoDAI) as the weighted sum of each element, as follows:

$$\begin{aligned} CoDAI = &(StakeholderEngagement/4) \\ &+ (TechnicalCapabilities/3) \end{aligned}$$

[20]The authors built their index on universities' websites, considering four dimensions of analysis: site content, speed, accessibility and navigability. While the first two are quite straightforward, the last two need to be explicitly explained. As also underlined by [83], accessibility has been proxied respectively by search engine indexes and popularity (the number of external links can imply more traffic). Navigability has been mainly measured as the number of steps (clicks) to access relevant information for the user. Other measures have been introduced as usability, which can be seen as an extension of navigability, more oriented to the convenience of users to navigate the website for several reasons, such as easiness, responsivity, and aesthetic reasons, which have become crucial, especially in the last years to retain and acquire users [57, 68, 84].

Table 4 Corporate Digital Assessment Index (CoDAI): indicators and dimensions

Indicator	Dimension
Quality external links	Stakeholder Engagement
Facebook	
Instagram	
Linkedin	
Quality tech. framework	Technical Capabilities
Security	
Request time	
Length URL	Internal Organisation
Quality internal links	
Website age	Digital Culture

Table 5 Corporate Digital Assessment Index (CoDAI): descriptive statistics

Dimension	min	max	mean	std
Stakeholder Engagement	0.000	0.756	0.230	0.133
Technical Capabilities	0.289	1.000	0.779	0.071
Internal Organisation	0.073	0.875	0.437	0.024
Digital Culture	0.000	1.000	0.412	0.298
CoDAI	0.830	3.148	1.858	0.350

$$+ (InternalOrganisation/2)$$

$$+ (DigitalCulture)$$

Table 4 shows the dimensions and indicators we used to build the CoDAI. Notice that a high CoDAI indicates a more pronounced digital readiness. We report in Table 5 the descriptive statistics for the CoDAI and its sub-dimensions. It is possible to see that each CoDAI sub-dimension varies between 0 and 1, even if some of them vary in a narrower range. We left the CoDAI theoretically ranging from 0 to 4 to better monitor each sub-indicator's contribution.

Subsequently, in Table 6, we provide a comprehensive summary of the regression outcomes concerning the determinants of the CoDAI and its respective four dimensions. Our proposed index highlights the expected significance levels for firm dimensions, underscoring the prominence of large firms relative to medium and micro firms. We find interesting results also across sectors, wide band and firm's age. For instance, the case of wholesale and retail trade (G) shows a positive and significant effect on stakeholder engagement but a negative and significant effect on the technical capabilities subdimension. The wide band is significant only for stakeholder engagement. This could be explained by the fact that most of the proxies employed to realise such a category rely on adopting external platforms. Unsurprisingly, we ascertain a negative effect concerning the firm's age for three out of four CoDAI sub-dimensions. This could be interpreted as a more reactive digital behaviour of newborn digital firms in comparison to more experienced ones.

Furthermore, the relevance of our proposed framework is evident if we compare the results of our CoDAI with the simple sum of the ten indicators (see Table 7). Employing a simple sum reveals that certain aspects of the digital divide, previously highlighted, tend to become obscured and challenging to interpret. For instance, the industrial prominence of Accommodation and Food Service (I) on the ICT sector (J) contrasts markedly with previous results reported in Sect. 4.1. In that section, the Information and Communication

Table 6 Regression results considering the CoDAI and its components as dependent variables. For a legend of NACE industries, see Table 3

	Stakeholder eng.	Technical Capabilities	Digital Culture	Internal Organization	CoDAI
Constant	0.827***	2.346***	0.252***	0.862***	1.672***
	(0.006)	(0.002)	(0.003)	(0.001)	(0.003)
Micro firms	−0.076***	−0.018***	−0.049***	−0.003***	−0.075***
	(0.003)	(0.001)	(0.001)	(0.000)	(0.002)
Mid-sized firms	0.119***	0.020***	0.040***	0.004***	0.079***
	(0.005)	(0.002)	(0.003)	(0.000)	(0.003)
Large firms	0.221***	0.033***	0.058***	0.007***	0.128***
	(0.010)	(0.004)	(0.005)	(0.001)	(0.006)
NACE sector A	0.029*	0.009	−0.028***	0.003*	−0.017*
	(0.012)	(0.005)	(0.006)	(0.001)	(0.007)
NACE sector C	0.132***	0.002	0.064***	0.015***	0.105***
	(0.004)	(0.002)	(0.002)	(0.000)	(0.003)
NACE sector F	0.033***	−0.014***	−0.022***	0.002***	−0.018***
	(0.005)	(0.002)	(0.003)	(0.000)	(0.003)
NACE sector G	0.047***	−0.008***	0.037***	0.013***	0.053***
	(0.004)	(0.002)	(0.002)	(0.000)	(0.003)
NACE sector H	0.039***	−0.017***	−0.001	0.001	0.004
	(0.008)	(0.003)	(0.004)	(0.001)	(0.005)
NACE sector I	−0.074***	−0.011***	0.051***	−0.009***	0.024***
	(0.006)	(0.002)	(0.003)	(0.001)	(0.004)
NACE sector J	0.201***	0.026***	0.079***	0.028***	0.152***
	(0.006)	(0.002)	(0.003)	(0.001)	(0.004)
NACE sector K	0.149***	0.023***	0.011	0.008***	0.060***
	(0.013)	(0.005)	(0.007)	(0.001)	(0.008)
NACE sector L	0.050***	−0.008*	0.048***	−0.000	0.058***
	(0.009)	(0.004)	(0.005)	(0.001)	(0.006)
NACE sector M	0.238***	0.023***	0.048***	0.018***	0.124***
	(0.006)	(0.002)	(0.003)	(0.001)	(0.003)
Urban area	0.030***	0.003**	0.004**	0.003***	0.014***
	(0.003)	(0.001)	(0.001)	(0.000)	(0.002)
North	0.067***	0.020***	0.007***	0.005***	0.033***
	(0.003)	(0.001)	(0.002)	(0.000)	(0.002)
South	−0.038***	−0.024***	−0.030***	−0.001*	−0.048***
	(0.004)	(0.002)	(0.002)	(0.000)	(0.002)
Firm age	−0.001***	−0.000***	0.007***	−0.000***	0.006***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Wide band	0.018***	−0.001	−0.001	0.000	0.003
	(0.004)	(0.001)	(0.002)	(0.000)	(0.002)
R-squared:	0.049	0.014	0.17	0.043	0.168
Adj. R-squared:	0.048	0.014	0.17	0.043	0.168
N. of observations	182,705	182,705	182,705	182,705	182,705

sector was identified as one of the highest performers. In addition, the signs of the South and North geographical areas are inverted, and the wide band variable becomes positive and significant. This again reverts findings reported in Sect. 4.1 and confirmed by the CoDAI in Table 6.

To illustrate the pronounced digital divide between companies in Italy, we map the results of our CoDAI in the NUTS 3 regions. Figure 3 shows three different geographical clusters corresponding to the three main regions of Italy: North, Centre and South (see also Fig. 4 for the map of municipalities). This finding underlines the suitability of digi-

Table 7 OLS regression results. The dependent variable is the sum of the values of the ten indicators of the corporate websites

	sum of the ten indicators
Constant	2.601***
	(0.011)
Micro firm	–0.178***
	(0.006)
Mid-sized firm	0.244***
	(0.010)
Large firm	0.416***
	(0.019)
Agriculture, forestry and fishing (NACE sector A)	0.115***
	(0.023)
Manufacturing (NACE sector C)	–0.064***
	(0.008)
Construction (NACE sector F)	–0.299***
	(0.011)
Wholesale and retail trade (NACE sector G)	0.088***
	(0.008)
Transportation (NACE sector H)	–0.321***
	(0.015)
Accommodation and food service (NACE sector I)	0.311***
	(0.011)
Information and communication (NACE sector J)	0.172***
	(0.011)
Financial and insurance activities (NACE sector K)	–0.205***
	(0.024)
Real estate activities (NACE sector L)	0.101***
	(0.018)
Professional, scientific & technical activities (NACE sector M)	0.071***
	(0.011)
Urban area	0.055***
	(0.005)
North	–0.026***
	(0.006)
South	0.019*
	(0.008)
Firm age	0.001***
	(0.000)
Wide band	0.027***
	(0.007)
R-squared:	0.038
Adj. R-squared:	0.038
N. of Observations	182,705

tal signals from websites as good candidates to reflect the existing economic disparities already demonstrated by traditional data sources, especially in terms of business productivity and differences in human capital endowments [86, 87].

Furthermore, we conduct a comparative analysis between companies with high and low CoDAI scores in terms of their size, macro sector and age. For this purpose, we selected two different subsamples from the whole dataset: one comprising 1827 companies belonging to the top 1% with the highest CoDAI scores and the other consisting of the 1% of companies with the lowest CoDAI scores. In Fig. 5 we visualise the distributions of these

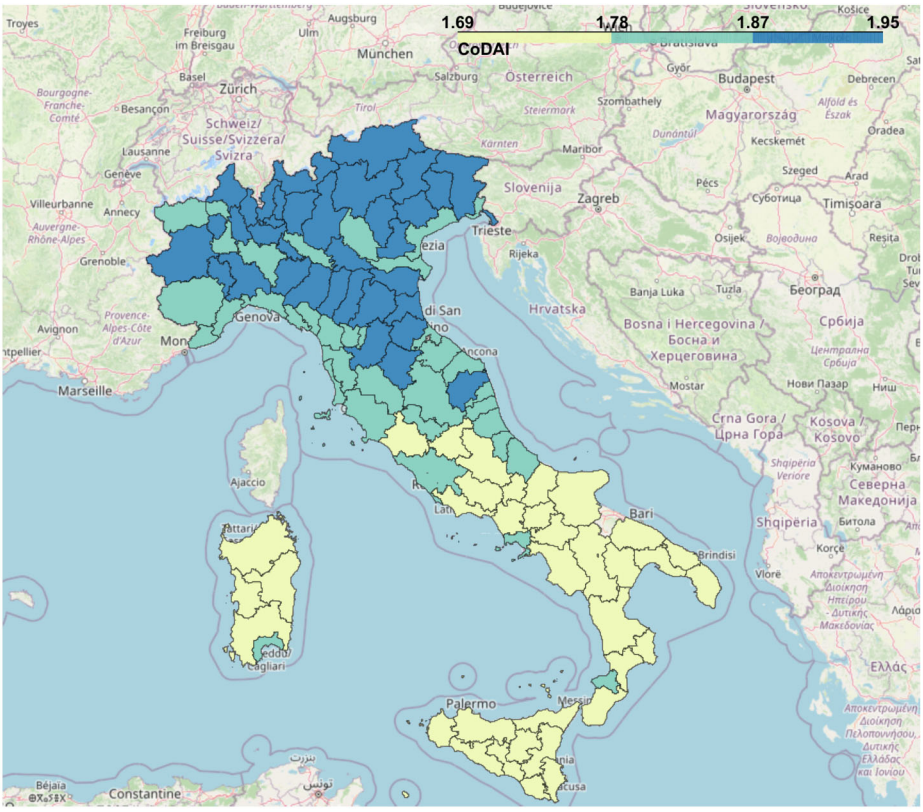

Figure 3 The distribution of the CoDAI across Italian regions (NUTS-3)

two samples in terms of company size, macro sector and age. Companies with a low CoDAI are mainly micro-enterprises. Companies with a high CoDAI value, on the other hand, are larger in size, with big firms almost exclusively represented in this category. This can be interpreted as a more significant role of digital readiness in more structured organisations. The distribution according to the age of the companies provides additional information. Older companies tend to have higher CoDAI values than younger ones. This supports the idea that startups in the early stages of their development may have a lower level of digital readiness than established companies. In terms of macro sectors, it is interesting to note that companies with a high CoDAI level predominate in Sector J (Information and communcation), which used to serve as a benchmark for the other sectors, while firms with a lower CoDAI level are more likely to be found in sector G (Wholesale and retail trade), where the use of digital means to communicate with customers is an important business infrastructure. To corroborate our findings with external data, we look at the sales performance of the highest and lowest 1% of companies in relation to their CoDAI level. We obtain information on the performance of the 90% of them, for the years 2018, 2019, 2020 and 2021 (51% are represented by the top 1% of the CoDAI distribution). Unsurprisingly, we find that companies with a higher CoDAI have higher average sales than companies with a lower CoDAI, which increases the reliability of our composite index, see Fig. 6. This is also evident when we report the average annual sales taking into account the company size (Fig. 7).

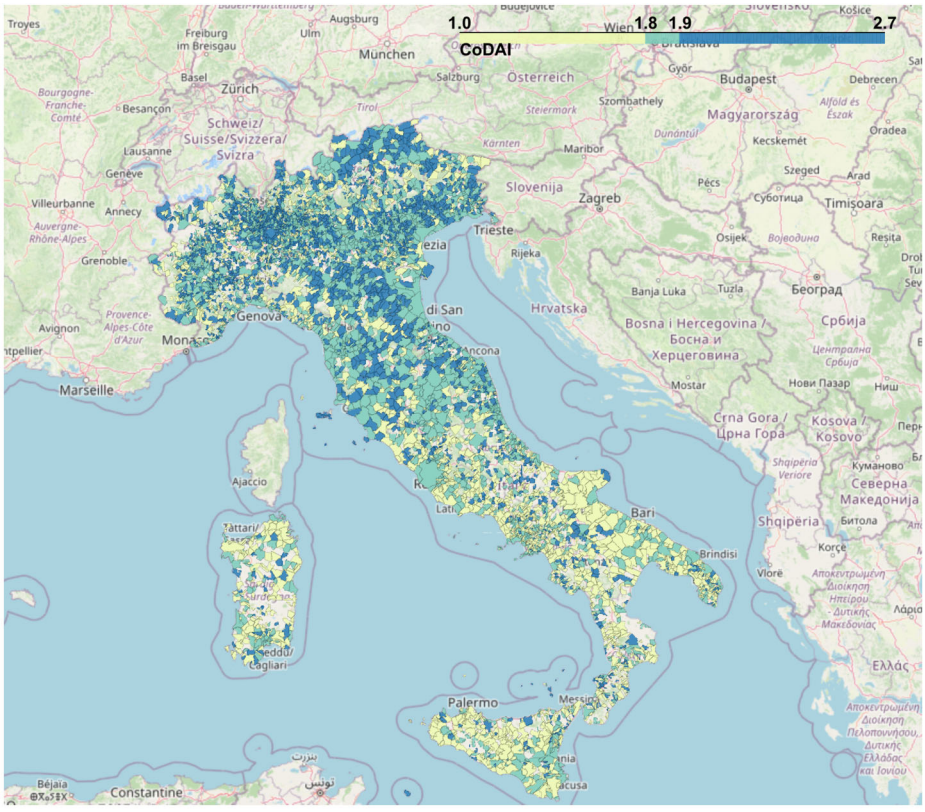

Figure 4 The distribution of the CoDAI across Italian municipalities

5 Discussion and conclusions

Today, the astonishing spread of ICT in production and consumption processes has revolutionised the way companies operate. This technological change has inevitably led to an apparent corporate digital divide, characterised by differences in the timeliness and quality of digital adoption between digitally competent and traditional "brick-and-mortar" enterprises.

With few exceptions, the extant literature has mapped the digital divide phenomenon using data on broadband access, often overlooking the role of digital tools and the relative competencies built upon them. In this paper, we advance research on the digital divide by using web-scraping techniques to identify and understand digital economy phenomena [17, 22, 23, 35, 48, 51, 56].

In particular, this exploratory study introduces a novel approach to mapping and assessing the corporate digital divide at a granular level, going beyond the traditional country-level perspective. Leveraging the potential of web scraping techniques, we delve into the landscape of the corporate digital divide by extracting, storing and analysing a comprehensive set of website characteristics related to the digital footprints of 182,705 Italian firms. We deliberately focus on technical features, as these are less prone to manipulation compared to content and user experience, and also have a lower degree of variability. The results of our regressions and the mapping of the extracted characteristics reveal different aspects of digital readiness with profound differences between companies by size, industry, location and age. The results also hold when controlling for broadband distribution.

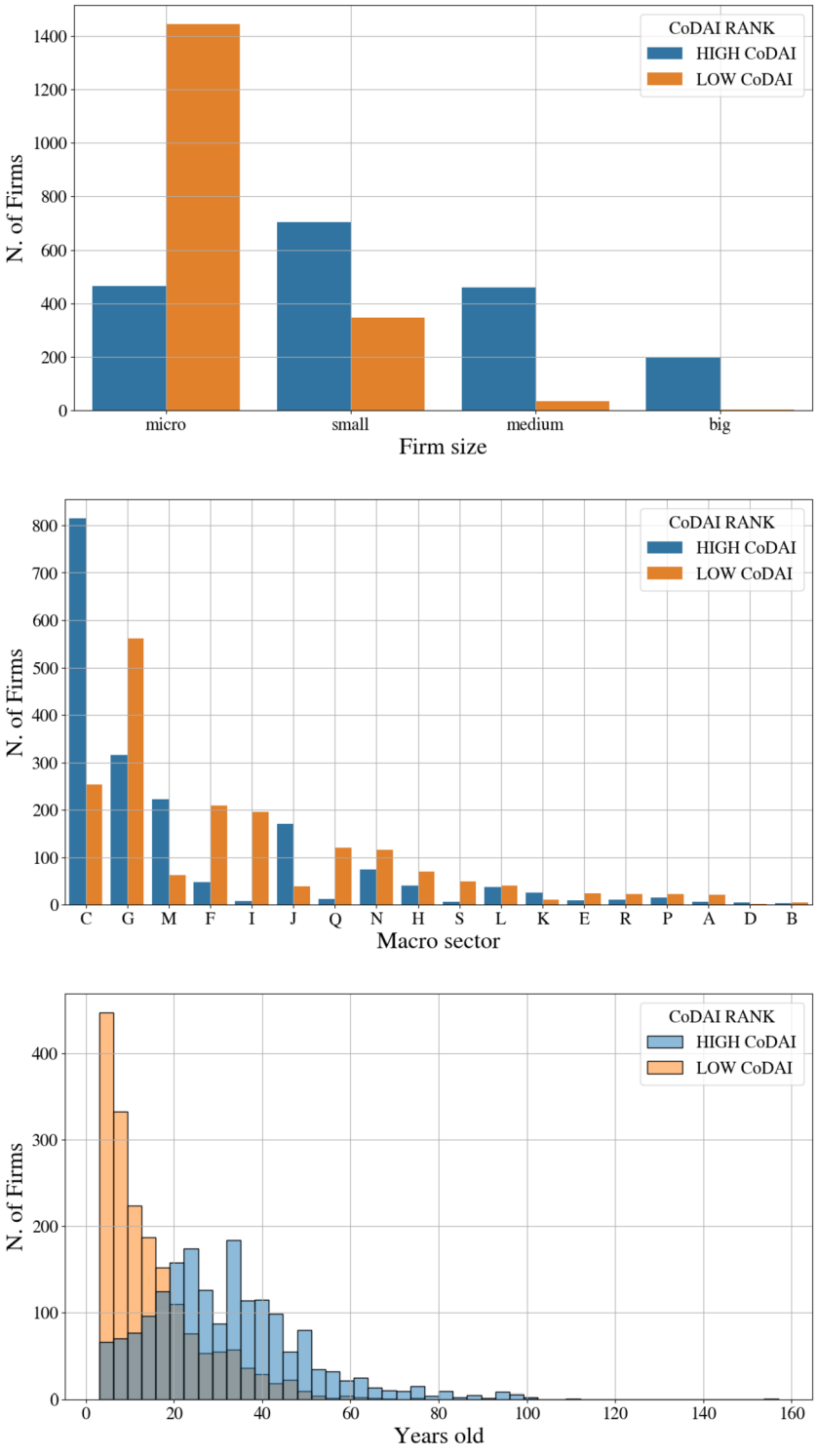

Figure 5 A comparison between the top 1% firms with the highest CoDAI and the lowest 1%

To improve the comparability of our results across firms, we propose the Corporate Digital Assessment Index (CoDAI), which facilitates the interpretation of results based on four different dimensions that encompass the company's digital strategy. The CoDAI and

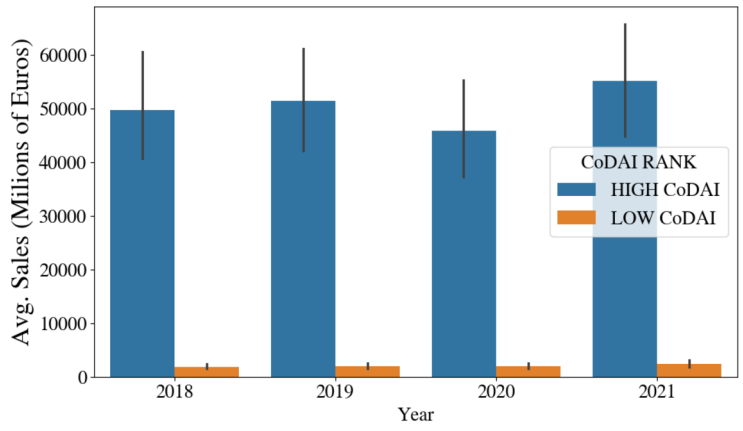

Figure 6 The average sales for 1% of companies with low and high CoDAI for the years 2018, 2019, 2020, and 2021

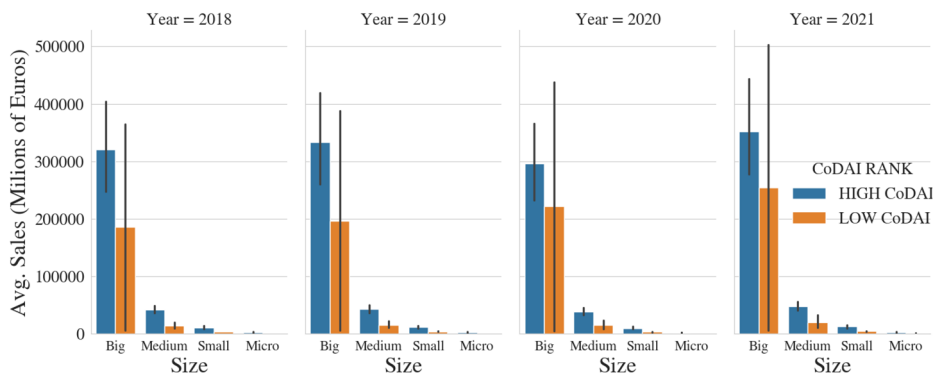

Figure 7 The average sales for 1% of companies with low and high CoDAI for 2018, 2019, 2020, and 2021, divided by their size

its four dimensions confirm the main drivers of the corporate digital divide highlighted in the existing literature, especially the prevalence of larger companies operating in ICT-related fields and located in more industrialised urban contexts. To increase the reliability of the index, we performed a comparison of the sales performance of companies within the top and bottom 1% of the CoDAI ranking, which was in favour of the highest 1%. In addition, we mapped the CoDAI to NUTS 3 regions and municipalities, confirming the traditional economic divide between northern and southern Italy.

The main contribution of our work is the development of a new methodology for capturing aspects of the digital strategies of firms using a novel set of unobtrusive web-related measures. The use of such a framework has important implications for the business and economics literature and provides access to information that would otherwise only be available through ad hoc surveys or interviews. The advantage also lies in the ability to collect these measures dynamically and on a large scale, which is often impossible with the above approaches.

Notwithstanding the relevance of our contribution to the analysis of the digital divide between companies, this work is not without its limitations. The literature on the economic interpretation of websites and digital resources is still in its infancy. Therefore, the

iterative dialog between empirical testing and theoretical development provides food for thought for multidisciplinary research between economics, management and data science.

First, future research directions could benefit from more research into the source code of websites to improve the evaluation of the technical quality of a website. For example, a more thorough investigation of JavaScript through the application of complexity measures and static analysis could yield important insights.

Second, the essential elements of a website change frequently, with multimedia elements such as images, video and sound experiencing a remarkable resurgence as a means of communication thanks to technological advances such as high-speed broadband connections. Furthermore, the design of a website has become a crucial mechanism for the quality of a website [58]. This means that the set of features adopted in this paper to evaluate the digital readiness of a company can be extended to full websites and updated and improved to meet new standards and technologies.

Third, our analysis is limited to Italy, but it can be expanded towards other countries in Europe and across the world, allowing researchers to perform comparative analysis and to unveil country-specific digital behaviours.

Fourth, although the data collection for this article is limited to a specific time period, the analysis can be repeated and renewed over several time periods by updating the extracted features of the website. In this way, the evolution of the company's digital capabilities can be tracked over time, eventually merging this data with more detailed (and/or other) balance sheet information.

Fifth, future analysis could benefit from an integrated approach that incorporates different data sources to triangulate the data obtained from the websites. This can be done by tracking external activities conducted by organisations on the Internet to obtain more robust measures of IT capabilities (e.g., job postings, see [47]). In addition, the analysis can be integrated and further validated with microdata from public institutions.

Sixth, previous literature has underlined that business goals drive IT development, even when websites are outsourced to external IT consultants [31]. This serves as a significant indication of a company's overall digital strategy. However, it is possible that in certain regions, the same external consultant may work with multiple clients. While the company's digital capabilities may persist, we could observe a standardisation of behaviour, which may confound the impact of regional expertise versus firm-level expertise. For example, we could create a website similarity measure to detect whether the same website developers are hired by multiple companies, relying on a clustering algorithm. Such an approach could reveal commonalities and localised patterns.

Seventh, from a methodological point of view, the extracted features have the potential to support various machine learning algorithms. For example, we could extract clusters of firms to highlight similar behaviours and develop a machine learning-based recommendation system that suggests improvements to companies in their interaction with the digital web.

All in all, our work paves the way with a methodological framework to evaluate the firm's digital capabilities at the micro level and reveals unprecedented details for scholars interested in applying data science methods to management and regional studies.

Beyond the difficulties of obtaining firm-level data on IT, the challenge is to develop sound proxies, from an epistemological point of view, to measure the link between technical resources and competencies, as the simple accounting of hardware or cloud infras-

tructure is unable to capture the leverage agency of the firm [47]. In this regard, software artifacts represent a valid alternative if crawled technical metrics are comparable across different types of organisations to partially mitigate the problem of the quality level as requested by the website owner(s) [88].[21]

Our research marks an initial step towards developing firm-level business intelligence that can be effectively used for micro-level benchmarking analysis. In future implementations of this work, we could also explore whether and how digital aspects captured by websites are evolving and how this might impact our assessment of digital strategy in the longer term.

In summary, the development of proxies that are able to capture the digital behaviour of companies in economic terms is an important tool for policymakers. A company's digital footprint is an important measure for defining targeted policies and development strategies for its replicability, unobtrusiveness, frequent updating, extension to new website information, and industrial benchmarking.

This is particularly important in the context of the Italian National Recovery and Resilience Plan, which was set up to help the country recover from the Covid-19 pandemic. As part of this plan, 21% of the total budget is earmarked for digitalisation (with measures such as the fastest connection through ultra-broadband, incentives for the adoption of innovative technologies by the private sector, revitalisation of the tourism and cultural sectors).[22]

Aware of the limitations and opportunities of our analysis, we contribute to applied research on the digital divide with an innovative and original approach that can be further leveraged through joint strategic partnerships between the public and private sectors.

Appendix

In the appendix, we report tables regarding the list of acronyms (see Table 8), the percentages of crawled firms by Nace Rev.2 (see Table 9), and finally the percentages of crawled firms by size (see Table 10).

Table 8 The list of abbreviations used in the paper and their relative explanation

Full text	Acronym
AI	Artificial Intelligence
AIDA	Advanced Industrial Document Automator
B2B	Business to Business
B2C	Business to Client
CoDAI	Corporate Digital Assessment Index
ICT	Information and Communication Technology
ISO	International Organisation for Standardisation
IT	Information Technology
NACE REV.2	Nomenclature Statistique des activités économiques
NUTS	Nomenclature of Territorial Units for Statistics.
SEO	Search Engine Optimisation
SME	Small and Medium Enterprise
TOE	Technology, Organisation, Environment
WAI	Web Assessment Index

[21] Accordingly, corporate websites serve different purposes and are created with different logic, which includes web designers' and external stakeholders' choices [88].

[22] see https://www.governo.it/it/approfondimento/le-missioni-e-le-componenti-del-pnrr/16700.

Table 9 In the table we report the percentage of crawled firms by Nace Rev.2. Please notice that with active firms we include active firms with at least one employee excluding sole proprietorship

Nace Rev.2	N. of Crawled firms	Percentage of Active Firms	Percentage of Active Firms with Website
A	2113	15.20%	64.80%
B	260	21.70%	62.40%
C	50,067	40.10%	69.20%
D	686	31.80%	70.70%
E	1914	37.80%	69.40%
F	14,727	15.10%	62.60%
G	42,360	26.90%	64.40%
H	5695	17.10%	63.50%
I	12,460	17.90%	59.30%
J	11,700	34.90%	66.90%
K	1847	22.70%	55.10%
L	3602	16.40%	63.00%
M	13,332	31.20%	66.30%
N	8838	23.70%	62.70%
O	10	43.50%	66.70%
P	2450	33.70%	66.00%
Q	5636	28.10%	66.20%
R	2649	19.60%	60.20%
S	2352	19.20%	59.10%

Table 10 In the table, we report the percentage of crawled firms by size. Please notice that with active firms, we include active firms with at least one employee, excluding sole proprietorship

Size	N. of Crawled firms	Percentage of Active Firms	Percentage of Active Firms with Website
Micro	109,015	21%	64%
Small	57,304	37%	64%
Medium	13,239	54%	71%
Big	3147	73%	83%

Acknowledgements
We would like to thank Valentina Pieroni for her support with the data collection.

Author contributions
ML designed the research, performed the experiments, wrote the paper, and reviewed drafts of the paper. PF collected the data, performed the experiments, wrote the paper, and reviewed drafts of the paper. RM designed the research, wrote the paper, and reviewed drafts of the paper. All authors read and approved the final manuscript.

Funding
This work was partially supported by the consortium Artes 4.0 - Advanced Robotics and Enabling Digital Technologies and Systems and the Department of Excellence "Economic and Digital Resilience (RED)" project of IMT Lucca.

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]European University Institute, Centre for a Digital Society/Robert Schuman Centre, Florence, Italy. [2]IMT School of Advanced Studies Lucca, Lucca, Italy. [3]IUSS Pavia, Pavia, Italy.

Received: 2 October 2023 Accepted: 19 July 2024 Published online: 30 July 2024

References

1. Blanka C, Krumay B, Rueckel D (2022) The interplay of digital transformation and employee competency: a design science approach. Technol Forecast Soc Change 178(C). https://doi.org/10.1016/j.techfore.2022.1
2. Verhoef PC, Broekhuizen T, Bart Y, Bhattacharya A, Dong JQ, Fabian N, Haenlein M (2021) Digital transformation: a multidisciplinary reflection and research agenda. J Bus Res 122(C):889–901

3. Volberda HW, Khanagha S, Baden-Fuller C, Mihalache OR, Birkinshaw J (2021) Strategizing in a digital world: overcoming cognitive barriers, reconfiguring routines and introducing new organizational forms. Long Range Plan 54(5):102110
4. Castells M (1996) He rise of the network society, the information age: economy, society and culture. Blackwell Sci., Oxford
5. Antonelli C (2003) The digital divide: understanding the economics of new information and communication technology in the global economy. Inf Econ Policy 15(2):173–199
6. Baskerville RL, Myers MD, Yoo Y (2020) Digital first: the ontological reversal and new challenges for information systems research. MIS Q 44(2):509–523
7. Van Dijk J, Hacker K (2003) The digital divide as a complex and dynamic phenomenon. Inf Soc 19(4):315–326
8. Fink C, Kenny JC (2003) W(h)ither the digital divide?. J Policy, Regul Strat Telecommun 5(6):15–24
9. Aissaoui N (2022) The digital divide: a literature review and some directions for future research in light of covid-19. Global Know Mem Commun 71(8/9):686–708
10. Corrocher N, Ordanini A (2002) Measuring the digital divide: a framework for the analysis of cross-country differences. J Inf Technol 17(1):9–19
11. Kyriakidou V, Michalakelis C, Sphicopoulos T (2011) Digital divide gap convergence in Europe. Technol Soc 33(3):265–270
12. Szeles MR (2018) New insights from a multilevel approach to the regional digital divide in the European Union. Telecommun Policy 42(6):452–463
13. Matthess M, Kunkel S (2020) Structural change and digitalization in developing countries: Conceptually linking the two transformations. Technol Soc 63
14. Ellinger AE, Lynch DF, Hansen JD (2003) Firm size, web site content, and financial performance in the transportation industry. Ind Mark Manage 32(3):177–185
15. Shakina E, Parshakov P, Alsufiev A (2021) Rethinking the corporate digital divide: the complementarity of technologies and the demand for digital skills. Technol Forecast Soc Change 162(C)
16. Lythreatis S, Singh SK, El-Kassar A-N (2022) The digital divide: a review and future research agenda. Technol Forecast Soc Change 175(C)
17. Thonipara A, Sternberg GR, Proeger T, Haefner L (2022) Digital divide, craft firms' websites and urban-rural disparities—empirical evidence from a web-scraping approach. Rev Reg Res, 1–14
18. Forman C (2005) The corporate digital divide: determinants of Internet adoption. Manag Sci 51(4):641–654
19. Vehovar V, Sicherl P, Hüsing T, Dolnicar V (2006) Methodological challenges of digital divide measurements. Inf Soc 22:279–290
20. Tambe P, Hitt L, Rock D, Brynjolfsson E (2020) Digital capital and superstar firms. National Bureau of Economic Research, Inc. NBER Working Papers 28285. https://ideas.repec.org/p/nbr/nberwo/28285.html
21. Blazquez D, Domenech J (2018) Big data sources and methods for social and economic analyses. Technol Forecast Soc Change 130(C):99–113
22. Gok A, Waterworth A, Shapira P (2015) Use of web mining in studying innovation. Scientometrics 102(1):653–671
23. Blazquez D, Domenech J, Debón A (2018) Do corporate websites' changes reflect firms' survival? Online Inf Rev 42(6):956–970
24. Kinne J, Resch B (2018) Generating big spatial data on firm innovation activity from text-mined firm websites. Int J Geogr Inf Sci 6(1):82–89
25. Youtie J, Hicks D, Shapira P, Horsley T (2012) Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. Technol Anal Strateg Manag 24(10):981–995
26. Li Y, Arora S, Youtie J, Shapira P (2018) Using web mining to explore triple helix influences on growth in small and mid-size firms. Technovation 76:3–14
27. Saridakis G, Lai Y, Mohammed AM, Hansen JM (2018) Industry characteristics, stages of e-commerce communications and entrepreneurs and smes revenue growth. Technol Forecast Soc Change 128(C):56–66
28. Pukelis L, Stanciauskas V (2019) Sing Internet data to complement traditional innovation indicators. Technical report, Public Policy and Management Institute (PPMI). https://www.ippapublicpolicy.org/file/paper/5d073ea805eb6.pdf
29. Auger P (2005) The impact of interactivity and design sophistication on the performance of commercial websites for small businesses. J Small Bus Manag 43(2):119–137
30. Faulkner P, Runde J (2009) On the identity of technological objects and user innovations in function. Acad Manag Rev 34(3):442–462
31. Miozzo M, Grimshaw D (2005) Modularity and innovation in knowledge-intensive business services: it outsourcing in Germany and the UK. Res Policy 34(9):1419–1439
32. Wells JD, Valacich JS, Hess TJ (2011) What signal are you sending? How website quality influences perceptions of product quality and purchase intentions. MIS Q 373–396
33. Abeysekera I (2019) How best to communicate intangible resources on websites to inform corporate-growth reputation of small entrepreneurial businesses. J Small Bus Manag 57(3):738–756
34. Arora SK, 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(8):1904–1915
35. Axenbeck J, Breithaupt P (2019) Web-based innovation indicators: Which firm website characteristics relate to firm-level innovation activity? ZEW Discussion Papers 19-063. https://ideas.repec.org/p/zbw/zewdip/19063.html
36. Arora SK, 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 Transf 45(5):1451–1480
37. George G, Osinga EC, Lavie D, Scott BA (2016) Big data and data science methods for management research. Acad Manag J 59(5):1493–1507
38. Brynjolfsson E, Wang C, Zhang X (2021) The economics of it and digitization: eight questions for research. MIS Q 45(1):473–477
39. Daniele V (2021) Territorial disparities in labour productivity, wages and prices in Italy: what does the data show? Eur Urban Reg Stud 28(4):431–449

40. Floridi L (2014) The fourth revolution: how the infosphere is reshaping human reality. Oxford University Press, London
41. Giustiziero G, Kretschmer T, Somaya D, Wu B (2021) Hyperspecialization and hyperscaling: a resource-based theory of the digital firm. Strateg Manag J
42. Warner KS, Wäger M (2019) Building dynamic capabilities for digital transformation: an ongoing process of strategic renewal. Long Range Plan 52(3):326–349
43. Herhausen D, Miočević D, Morgan RE, Kleijnen MH (2020) The digital marketing capabilities gap. Ind Mark Manage 90:276–290
44. Liu Y, Dong J, Mei L, Shen R (2022) Digital innovation and performance of manufacturing firms: an affordance perspective. Technovation 102458
45. Hanelt A, Bohnsack R, Marz D, Antunes Marante C (2021) A systematic review of the literature on digital transformation: insights and implications for strategy and organizational change. J Manag Stud 58(5):1159–1197
46. Brynjolfsson E, Rock D, Syverson C (2019) Artificial intelligence and the modern productivity paradox. Econ Art Int Agenda 23:23–57
47. Brynjolfsson E, Jin W, Steffen S (2024) Do it capabilities still drive productivity and innovation in the digital age? Available at SSRN 4765508
48. Kinne J, Axenbeck J (2020) Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics 125(3):2011–2041
49. Ageeva E, Melewar T, Foroudi CP, Dennis JZ (2018) Examining the influence of corporate website favorability on corporate image and corporate reputation: findings from fsqca. J Bus Res 89:287–304
50. Beaudry C, Heroux-Vaillancourt M, Rietsch C (2016) Validation of a web mining technique to measure innovation in high technology Canadian industries pp 1–25
51. Krüger M, Kinne J, Lenz D, Resch B (2020) The digital layer: how innovative firms relate on the web. ZEW Discussion Papers 20-003
52. Abbasiharofteh M, Krüger M, Kinne J, Lenz D, Resch B (2023) The digital layer: alternative data for regional and innovation studies. Spatial Econ Anal 18(4):507–529
53. Abbasiharofteh M, Kinne J, Krüger M (2024) Leveraging the digital layer: the strength of weak and strong ties in bridging geographic and cognitive distances. J Econ Geogr 24(2):241–262
54. Dahlke J, Beck M, Kinne J, Lenz D, Dehghan R, Wörter M, Ebersberger B (2024) Epidemic effects in the diffusion of emerging digital technologies: evidence from artificial intelligence adoption. Res Policy 53(2):104917
55. Billon M, Ezcurra R, Lera-López F (2009) Spatial effects in website adoption by firms in European regions. Growth Change 40(1):54–84. https://doi.org/10.1111/j.1468-2257.2008
56. Heroux-Vaillancourt M, Beaudry C, Rietsch C (2020) Using web content analysis to create innovation indicators - what do we really measure? Quant Sci Stud 1(4):1601–1637
57. Mateos MB, Mera AC, González FJM, López ORG (2001) A new web assessment index: Spanish universities analysis. Internet Res 11(3):226–234
58. Rasmussen KB, Thimm H (2015) Circumventing nonresponse-upgrading traditional company survey data with unobtrusive data from company websites. Bull Soc Method 127(1):85–96
59. Rammer C, Es-Sadki N (2023) Using big data for generating firm-level innovation indicators-a literature review. Technol Forecast Soc Change 197:122874
60. Stich C, Tranos E, Nathan M (2023) Modeling clusters from the ground up: a web data approach. Environ Plan B: Urban Anal City Sci 50(1):244–267
61. Hernàndez B, Jiménez J, Martìn MJ (2009) Key website factors in e-business strategy. Int J Inf Manag 29(5):362–371
62. Arifi D, Resch B, Kinne J, Lenz D (2023) Innovation in hyperlink and social media networks: comparing connection strategies of innovative companies in hyperlink and social media networks. PLoS ONE 18(3):0283372
63. Kinne J, Lenz D (2021) Predicting innovative firms using web mining and deep learning. PLoS ONE 16(4):0249071
64. Sanders JW, Galloway L (2013) Rural small firms' website quality in transition and market economies. Jo Small Bus Enter Dev 20(4):788–806
65. Blazquez D, Domenech J (2018) Web data mining for monitoring business export orientation. Technol Econ Dev Econ 24(2):406–428
66. Blazquez D, Domenech J, Gil JA, Pont A (2019) Monitoring e-commerce adoption from online data. Knowl Inf Syst 60(1):227–245
67. Ziakis C, Vlachopoulou M, Kyrkoudis T, Karagkiozidou M (2019) Important factors for improving Google search rank. Future Internet 11(2):32
68. Król K, Zdonek D (2020) Aggregated indices in website quality assessment. Future Internet 12(4):72
69. Axenbeck J, Breithaupt P (2021) Innovation indicators based on firm websites—which website characteristics predict firm-level innovation activity? PLoS ONE 16(4):0249583
70. Zhang S, Cabage N (2017) Search engine optimization: comparison of link building and social sharing. J Comput Inf Syst 57(2):148–159
71. Koski H, Pajarinen M, Rouvinen P (2019) What company characteristics are associated with the adoption of social media? Ind Innov 26(8):880–897
72. Aronica M, Bonfanti RC, Piacentino D (2021) Social media adoption in Italian firms. Opportunities and challenges for lagging regions. Pap Reg Sci 100(4):959–979
73. Yamada K, Yamana H (2019) Effectiveness of usability & performance features for web credibility evaluation. In: 2019 IEEE international conference on big data (big data). IEEE, Los Alamitos, pp 6257–6259
74. Heričko T, Šumak B, Brdnik S (2021) Towards representative web performance measurements with Google lighthouse. In: Proceedings of the 2021 7th student computer science research conference, p 39
75. Shan TS, Obit JH, Alfred R, Tahir A (2017) Enhancing the performance of university's website for mobile devices based on responsive web design approach. Adv Sci Lett 23(11):10969–10973
76. Le TD, Le-Dinh T, Uwizeyemungu S (2024) Search engine optimization poisoning: a cybersecurity threat analysis and mitigation strategies for small and medium-sized enterprises. Technol Soc 102470
77. Buchanan WJ, Helme S, Woodward A (2018) Analysis of the adoption of security headers in http. IET Inf Secur 12(2):118–126

78. Alashwali ES, Szalachowski P, Martin A (2020) Exploring https security inconsistencies: a cross-regional perspective. Comput Secur 97:101975
79. Lavrenovs A, Melón FJR (2018) Http security headers analysis of top one million websites. In: 2018 10th International Conference on Cyber Conflict (CyCon). IEEE, Los Alamitos, pp 345–370
80. Mendoza A, Chinprutthiwong P, Gu G (2018) Uncovering http header inconsistencies and the impact on desktop/mobile websites. In: Proceedings of the 2018 World Wide Web Conference, pp 247–256
81. Brynjolfsson E, Jin W, Wang X (2023) Information technology, firm size, and industrial concentration. Technical report, National Bureau of Economic Research
82. Forman C, Goldfarb A, Greenstein S (2005) How did location affect adoption of the commercial Internet? Global village vs. urban leadership. J Urban Econ 58(3):389–420
83. Vaughan L (2004) Exploring website features for business information. Scientometrics 61(3):467–477
84. Dickinger A, Stangl B (2013) Website performance and behavioral consequences: a formative measurement approach. J Bus Res 66(6):771–777
85. Mirtsch M, Kinne J, Blind K (2020) Exploring the adoption of the international information security management system standard iso/iec 27001: a web mining-based analysis. IEEE Trans Eng Manag 68(1):87–100
86. Rungi A, Biancalani F (2019) Heterogeneous firms and the north–south divide in Italy. Ital Econ J 5(3):325–347
87. Odoardi I, D'Ingiullo D, Furia D (2020) Human capital and the reduction of inequalities: an intra-national analysis in Italy. Appl Econ 52(57):6215–6228
88. Fogli D, Guida G, et al (2013) Assessing corporate web sites: quality model and methodology. Lect Notes Comput Sci 8295:92–103

Publisher's Note