CARMA 2025

7th Int. Conf. on Advanced Research Methods and Analytics
2-4 July 2025, Rome, Italy
DOI: https://doi.org/10.4995/CARMA2025.2025.20497

Antecedents and drivers of Net-Zero Technologies using web scraping: an empirical analysis from Italian corporate websites

Marco Cucculelli , Noemi Giampaoli, Matteo Renghini
Dipartimento di Scienze economiche e Sociali, Università Politecnica delle Marche, Italy.

How to cite: Cucculelli M.; Giampaoli N., Renghini M. 2025. Antecedents and drivers of Net-Zero Technologies using web scraping: an empirical analysis from Italian corporate websites. In: 7th International Conference on Advanced Research Methods and Analytics (CARMA 2025). Rome, 2-4 July 2025. https://doi.org/10.4995/CARMA2025.2025.20497

Abstract
The Net Zero Industry Act (NZIA) represents a breakthrough in fostering innovation and green transition among firms. However, information on companies adopting technologies related to the NZIA framework is not readily available, making it necessary to rely on unstructured database and web-scraping techniques to identify them. This paper proposes the use of commercial websites to detect NZIA-related technologies. First, we demonstrate how websites can be leveraged to identify companies involved NZIA technologies. Second, we explore the antecedents and drivers that influence the likelihood of NZIA technologies adoption. Third, we distinguish between companies producing and companies using NZIA technologies. This approach offers a bottom-up perspective that can support policymakers in mapping relevant industrial ecosystems, while also highlighting the value of advanced NLP techniques in economic and business research.

Keywords: *Web Data; Innovation; NZIA; Web Scraping; Unconventional data; Corporate websites.*

1. Introduction

The Net Zero Industry Act (NZIA) plays a crucial role in fostering innovation and accelerating the transition toward sustainable industrial practices. It is a key component of the EU's Green Deal Industrial Plan, aiming to boost domestic manufacturing of net-zero technologies. Alongside the Critical Raw Materials Act (CRMA), it reduces reliance on unstable global supply chains and strengthens energy security. By fostering a competitive cleantech sector, NZIA advances Europe's climate goals, industrial leadership, and job creation.

Scholarly interest in these EU strategies is growing. Hool et al. (2024) examine the CRMA's opportunities and risks, while Lofgren et al. (2024) discuss green policy design for basic materials. Tagliapietra et al. (2023) address NZIA legislation, and Kleimann et al. (2023) compare NZIA to the U.S. IRA. Yet, identifying net-zero technologies at the firm level remains challenging. These innovations are fluid, interdisciplinary, and often invisible to traditional metrics. Standard tools, surveys, patents, financials (Gault, 2013; Hall et al., 2010), struggle to capture early-stage, informal, or non-patented activity, particularly among SMEs and in fast-evolving sectors (Santarelli & Sterlacchini, 1990; Balland & Boschma, 2020). To overcome these limits, researchers increasingly rely on digital traces and unstructured data (Nathan & Rosso, 2015; Stich et al., 2022). Corporate websites, reflecting firms' projects and goals, offer a valuable and underused source for identifying technology adoption, including net-zero innovations. This study addresses these gaps by posing the research questions: Where are the NZIA Technologies? Which factors determine the adoption of Net-Zero Industry Act technologies by Italian firms?. The study provides a detailed mapping, differentiates producers from users, and identifies firm-level drivers. It contributes to the literature on European strategic tech and green plans (Lofgren et al., 2024; Tagliapietra et al., 2023) by using web scraping to uncover new data (Arribas-Bel and Bakens, 2019; Gentzkow et al., 2019), offering timely and granular insights. We combine unstructured website data with structured firm-level data from Orbis (Moody's). Through web scraping, we identify NZIA firms and classify them as producers or users. The structured data captures financials, ownership, sector, and location to analyze adoption drivers. The method includes two main steps: mapping NZIA presence across regions and sectors; and estimating a probit model to identify firm-level determinants of adoption. A further round of scraping helps isolate producers, offering a bottom-up view of the Italian industrial ecosystem. We find that NZIA technologies are concentrated in the Adriatic coast, around Rome, and in Central-Northern Italy. Larger firms, market power, and local wealth boost adoption, while debt, especially short-term, limits it. Equity and liquidity facilitate it. Family ownership and long CEO tenure reduce the likelihood of adoption, while STEM-qualified staff and managerial turnover increase it. Producers are typically larger, equity-rich, and STEM-intensive, highlighting structural differences between users and producers.

The rest of the paper runs as follow. Section 2 outlines data and scraping strategy. Section 3 presents the methodology. Section 4 reports findings. Section 5 offers robustness checks. Section 6 concludes with policy insights and future research directions.

2. Web scraping: Unveiling NZIA technologies through company websites

2.1. Data Collection Strategy

We implement a large-scale web scraping exercise to get NZIA information from Italian company websites. The URLs of these websites are retrieved from the AIDA dataset (Moody's Orbis). In addition, we enrich our dataset with firm-level innovation data, namely, the presence and count of patents, sourced from the Patent Orbit module. Starting from an initial and comprehensive sample of 1,702,893 active Italian companies, we isolate those that report a valid corporate URL. After removing websites associated with multiple firms to ensure one-to-one matching, we retain 477,812 unique companies with a valid and uniquely identifiable web domain. These companies represent the core sample for our web scraping procedure.

2.2. Web structure mapping and scraping architecture

The scraping was conducted in Python 3.11.9, hosted on a Virtual Machine for greater scalability. We began by mapping each company's homepage (the "parent") to identify its structure and retrieve all intra-site hyperlinks, the so-called "child pages". We define depth 1 as direct links from the homepage (i.e., one click away), and depth 2 as links reachable through a child page (two clicks from the homepage). To minimize connection or timeout errors, we fed every web page URL to the scraping algorithm at least twice. Following this mapping, we perform an initial data cleaning before proceeding with the keyword search within the websites. Specifically, we dropped all non-functional URLs, URLs with excessive loading times, adn links pointing outside the company's domain or to social media platforms. Finally, we keep only unique links to remove any duplicate web pages. This rigorous filtering process results in a refined sample of 6,541,465 unique links across 366,440 companies, ready for keyword mining.

2.3. Keyword-based text extraction and processing

The second stage consists of extracting relevant textual content and identifying keywords linked to NZIA technologies. We create a keywords' dictionary based on the European Commission's NZIA documentation, encompassing both technological terms and semantic variations. For further details regarding the definition of the key terms used, see *Implementing the EU's Net-Zero Industry Act* provided by the European Parliament (Ragonnaud, 2025). Given the noisy and unstructured nature of web text, we develop a robust pre-processing pipeline. First, the standardization of punctuation and whitespace is vital. Since website text often lacks consistent punctuation, we devised a strategy to standardize punctuation by replacing key characters (points, comma, dash, parentheses, etc.), the new line symbol \n, and the new tab symbol \t in the raw HTML output of the scraper with a space. Second, the tokenization into single words, bigrams and trigrams has been done. We ensured that the entire text in the raw HTML is transformed into a list of individual words to facilitate sequential reading. Specifically, we implemented an algorithm to read not only individual words (e.g., "RAN") but also bigrams (e.g., "artificial intelligence") and trigrams (e.g., "radio access network"). Finally, as website text is often “contaminated” with words from other languages, we followed a dual encoding (UTF-8 and UTF-16) to handle multilingual content and avoid data loss.

After two rounds of scraping, we obtained information for a total of 203,717 companies. The results about keywords are compiled in tabular form and aggregated at the firm level, with two post-processing refinements: morphological harmonization and semantic alignment. The first refinement refers to singular/plural forms and gendered forms. The second refinement pertains to acronyms vs. full form. After searching for both the acronyms (e.g., RAN) and the full term (e.g., radio access network), we grouped them and summed their frequencies.

2.4. Identifying producers vs. users of NZIA technologies

To distinguish producers from users of NZIA technologies, we extend our scraping methodology with a second keyword dictionary focusing on innovation signaling and potential technological producers. The methodological steps performed are identical to those previously shown in Figure 2, from point four onwards, but with a different set of keywords. Specifically, we look for terms linked to intellectual property (e.g., patents, trademarks), in-house R&D, research centers and proprietary technologies. The classification logic is as follows: a company is defined as a producer (Producer) if its website contains both NZIA-related terms and innovation-related keywords.

3. Antecedents and drivers

3.1. Firms’ Antecedents and Drivers – Data Collection

To gather both financial and qualitative firm-level information, we draw on data from the AIDA dataset (Moody’s Orbis, Bureau van Dijk, 2025). We collect data related to financial aspects, sectoral features, ownership structures, capital structure composition, and other characteristics. The determinants of NZIA adoption are categorized in antecedents and drivers.

<u>Antecedents</u>. We classify antecedents into two main groups of control variables: (i) firm-internal factors, reflecting the supply-side (i.e., firm and market characteristics), and (ii) firm-external factors, reflecting the demand-side (i.e., market pressures and external incentives).

Drivers. We identify four groups of drivers that may influence the likelihood of NZIA adoption among Italian firms: (1) financial drivers, (2) profitability-related drivers, (3) organizational and relational drivers, and (4) knowledge and human capital drivers.

To ensure the reliability of the sample, we apply several data cleaning steps. We exclude firms if leverage ratio is greater than one and lower than zero and those with equity greater than one and less than zero. We also remove firms with negative revenues. After excluding missing and inconsistent values, the final cleaned sample includes 105,443 firms. We are aware that the web scraping techniques adopted in this study have substantially reduced the size of the original sample. However, this attrition does not compromise the validity of our analysis.

3.2. Antecedents and drivers - Empirical Methods

In this section, we outline our empirical strategy. We estimate the following probit model:

$$NZIA_i = \alpha + \beta Driver_i + \gamma Antecedent_i + \varepsilon_i \tag{1}$$

where i=1…105,443 corresponds to the number of firms. The dependent variable NZIA equals 1 if the firm is identified as a NZIA adopter based on our web scraping analysis, and 0 otherwise. We first test our baseline hypotheses regarding the role of antecedents in facilitating NZIA adoption. Second, we examine the role of drivers in shaping the likelihood of NZIA adoption. Thus, we look at the β coefficient and we have mixed outlook about the direction of the effects.

4. Results

4.1. Mapping the NZIA technologies

We first provides a descriptive overview of the adoption of NZIA technologies across Italian firms from geographical, sectoral, and firm-level perspectives. Geographically, NZIA adoption is concentrated mainly in Northern Italy, particularly in the wealthier North-Eastern regions like Lombardy, Veneto, and Emilia-Romagna, indicating a link between local economic prosperity and technological uptake. Sectorally, NZIA firms are mainly found in manufacturing, wholesale and retail trade, and technologically intensive sectors such as public administration and energy supply. Market competition appears to favor NZIA adoption, with most adopters located in highly competitive sectors. At the firm level, adoption rates increase with firm size and age, suggesting that larger and older firms are more likely to invest in these strategic technologies, possibly due to fewer financial constraints. Overall, NZIA adoption is strongly shaped by regional economic conditions, sectoral characteristics, and firm-specific factors. See Figure 1.

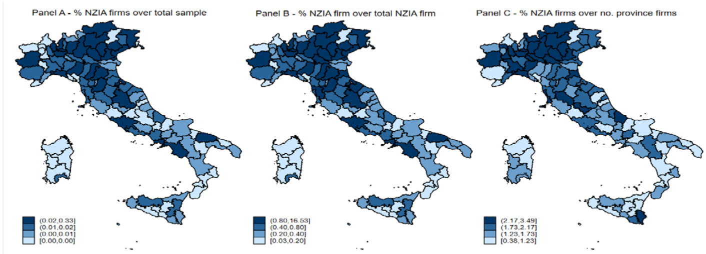

Figure 1. Nuts3 distribution of NZIA firms. Source: Authors' elaboration.

4.2. Antecedents and drivers

The empirical results show that larger firms are more likely to adopt NZIA technologies, confirming that firm size enhances investment capacity. Contrary to expectations, higher market concentration also increases adoption, likely due to greater market power enabling innovation. Firm age has a nonlinear effect: both younger and incumbent firms are more inclined to adopt NZIA technologies, driven by dynamic growth strategies and accumulated knowledge, respectively. Wealthier local economies positively influence adoption by supporting innovation demand and institutions. Financially, higher debt reduces adoption likelihood, especially long-term debt, while equity capital and profitability promote it, indicating firms mainly rely on internal funds for innovation. Organizationally, family ownership and longer managerial tenure negatively affect adoption, reflecting conservative strategies, whereas firms with more STEM-educated employees and managers are more likely to innovate. Managerial incentives aligned with long-term goals are also important for fostering adoption. See Table 1 nad Table 2.

Table 1. Results: Antecedents, financial drivers and profitability-related drivers of NZIA firms.
Source: authors' elaboration.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep.Var.	NZIA	NZIA	NZIA	NZIA	NZIA	NZIA	NZIA
Driver =	*LevRatio*	*ShortDebt*	*LongDebt*	*Equity*	*ROE*	*LnEBITDA*	*SalesGrowth*
Driver	-0.008***	-0.005**	-0.011***	0.014***	-0.001	0.005***	0.003**
Antecedent	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	104,798	104,798	104,798	104,916	102,223	97,068	102,623
SectorFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RegionalFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 2. Results: Organizational and relational drivers, knowledge and human capital drivers of NZIA firms. Source: authors' elaboration.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep.Var.	NZIA	NZIA	NZIA	NZIA	NZIA	NZIA	NZIA
Driver =	*District*	*FamilyD*	*FamilyG*	*Duration*	*%STEM*	*D_STEM*	*Man_Age*
Driver	-0.002	-0.005***	-0.004***	-0.001***	0.026***	0.026***	0.001
Antecedent	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	105,122	105,122	105,122	105,122	105,122	105,122	59,993
SectorFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RegionalFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

4.3. Producers versus users

The analysis shows that firms with stronger equity and lower short-term debt are more likely to produce NZIA technologies, highlighting the need for solid financial capacity to support capital-intensive R&D. Surprisingly, traditional profitability measures have little influence on production orientation. Firm size and market power also positively affect the likelihood of producing these technologies. Organizationally, membership in industrial districts and a high share of STEM-qualified employees significantly increase production chances, while family ownership has no major effect. Longer managerial tenure tends to reduce production likelihood, except when combined with STEM expertise, which supports sustained production capabilities.

5. Robustness

The robustness of the results is confirmed across various model specifications, including regressions without and with fixed effects, OLS, and logit models, all yielding consistent findings. Moreover, alternative definitions of producers and users based on patent portfolios produce stable results, further reinforcing the main conclusions. Lastly, the analysis highlights the importance of conducting separate examinations for micro, small, medium, and large firms to better capture structural and strategic heterogeneity across different firm sizes.

6. Conclusion

This study develops a novel method to identify firms adopting and producing Net-Zero Industry Act (NZIA) technologies by combining large-scale web scraping of corporate websites with structured firm-level data. This approach overcomes limitations of traditional tools like patent analysis, enabling real-time, scalable mapping of NZIA engagement across Italian firms. The analysis reveals that NZIA adoption is concentrated in wealthier northern regions and technologically intensive sectors, with larger, older firms and those in more concentrated markets more likely to adopt these technologies. Financial strength, especially equity capital, and organizational factors, such as STEM workforce presence and managerial tenure, significantly influence adoption and production. Producers require strong capital buffers and benefit from industrial district networks and STEM-educated management. The findings suggest key policy priorities including improving green finance access, supporting industrial clusters, investing in STEM skills, and fostering competitive frameworks to boost domestic clean-tech innovation. Despite data challenges, this integrative methodology offers a powerful tool for monitoring technological transitions and informing policy in the net-zero transition.

References

Arribas-Bel, D., Bakens, J. (2019). *Use and validation of location-based services in urban research: An example with Dutch restaurants*. Urban Studies, 56(5), 868–884. https://doi.org/10.1177/0042098018779554

Balland, P., Boschma, R. (2020). *Smart specialisation: beyond patents*. European Commission. Directorate General for Regional and Urban Policy.

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

Gentzkow, M., Kelly, B., Taddy, M. (2019). *Text as data*. Journal of Economic Literature, 57(3), 535–74. DOI: 10.1257/jel.20181020

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

Hool., A., Helbig, C. (2023). *Challenges and opportunities of the European Critical Raw Materials Act*. Mineral Economics, 37, 661-668. https://doi.org/10.1007/s13563-023-00394-y

Kleimann, D., Poitiers, N., Sapir, A., Tagliapietra, S., Veron, N., Veugelers, R., Zettelmeyer, J. (2023). *Green tech race? The US Inflation Reduction Act and the EU Net Zero Industry Act*. World Economy. https://doi.org/10.1111/twec.13469

Lofgren, A., Ahlvik, L., van den Bijgaart I., Coria, J., Jaraitè, J., Johnsson, F., Rootzen, J. (2024). *Green industrial policy for climate action in the basic materials industry*. Climate Change, 147-177. https://doi.org/10.1007/s10584-024-03801-7

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

Ragonnaud, G. (2025). *Implementing the EU's Net-Zero Industry Act*. Europen Parliament, EPRS – European Parliament Research Service.

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

Stich, C., Tranos, E., Nathan, M. (2022). *Modeling clusters from the ground up: A web data approach*. Environment and Planning B: Urban Analytics and City Science. https://doi.org/10.1177/23998083221108185

Tagliapietra, S., Veugelers, R., Zettelmeyer, J. (2023). *Rebooting the European Union's Net Zero Industry Act*. Bruegel Policy Brief, No. 15/2023.