IDENTIFICATION OF STEP AND NZIA TECHNOLOGIES THROUGH TEXT MINING: AN EMPIRICAL ANALYSIS OF PATENT DATA

Marco Cucculelli Noemi Giampaoli
Matteo Renghini

Working paper no. 188

September 2024

Identification of STEP and NZIA technologies through text mining: An empirical analysis of patent data

Marco Cucculelli
Department of Economics and Social Sciences, Università Politecnica delle Marche,
P.le Martelli 8, 60100 Ancona, Italy. E-mail: m.cucculelli@univpm.it

Money and Finance Research Group (MoFiR), Italy

Noemi Giampaoli
Department of Economics and Social Sciences, Università Politecnica delle Marche,
P.le Martelli 8, 60100 Ancona, Italy. E-mail: n.giampaoli@univpm.it

Money and Finance Research Group (MoFiR), Italy

Matteo Renghini
Department of Economics and Social Sciences, Università Politecnica delle Marche,
P.le Martelli 8, 60100 Ancona, Italy. E-mail: m.renghini@univpm.it

Department of Economics and Finance, LUISS Guido Carli,
Viale Pola 12, 00198 Roma, Italy. E-mail: mrenghini@luiss.it

Money and Finance Research Group (MoFiR), Italy

Abstract

Assessing the presence and distribution of strategic and net-zero technologies in companies is crucial for European competitiveness. However, due to the complexity and evolving nature of these technology areas, this is a challenging task. This paper presents a process for identifying and mapping strategic and net-zero technologies (as described in the Strategic Technologies for Europe Platform (STEP) and the Net-Zero Industry Act (NZIA)) in European companies. STEP and NZIA technologies are identified using text mining techniques based on the titles and abstracts of patents filed with the EPO and retrieved in PATSTAT for the years 2002 to 2022. The paper describes the classification process of STEP and NZIA technologies based on IPC codes of file patents. The IPC codes were then matched with the patent portfolio of almost 100,000 European companies to determine the company's technological profile and the distribution of these technologies by sector, geographic area, and company characteristics in the European panorama.

Keywords: PATSTAT, Orbis, Patents, Text mining, Innovation, STEP, NZIA, Unconventional data

1. Introduction

With the recently adopted Net-Zero Industry Act (NZIA) (2024), Europe has set itself the target of significantly decarbonizing production and the economy by 2030. This means that the next five years will be crucial for companies' market decisions, especially for European manufacturing companies that need to shift their product portfolio to zero environmental impact technologies and production. Companies that do not make this decision will be increasingly marginalized, both in the marketplace and in their relationships with institutions and the financial system.

Unfortunately, there is no tool or model to support companies in this crucial transition to decarbonization. It is also not easy to identify which companies are using STEP-NZIA technologies, as a list of these innovative technologies and associated industries is not easily available. Therefore, identifying and mapping these technologies relevant to the Strategic Technologies for Europe Platform (STEP) and the Net-Zero Industry Act (NZIA) is a challenging task, largely due to the complexity and evolving nature of these technologies and associated technological innovation.

Researching corporate innovation is a complex objective as measuring innovation in companies is far from straightforward. Consolidated measures of innovation refer to product introduction, process innovation, or organizational innovation (e.g., Laursen et al., 2012), the number of new products or processes (e.g., De Saá-Pérez et al., 2012), or the percentage of sales of new products and services in the total sales of the company (e.g., Wu, 2008). In addition to these measures, which are usually obtained through direct investigation or survey data, patents are an alternative indicator of technological innovation (e.g. Wu et al., 2005) and several patent-based measures have been used (e.g. Block and Spiegel, 2013; Block et al., 2013; Chirico et al., 2020). Nowadays, the use of patent data as an indicator of (technological) innovation is an established method for measuring a firm's innovation process, even if it is not without problems. The presence of other protection mechanisms (Arundel, 2001; Ejermo, 2009), the difficulty of matching patents with firm data, or the time lag in the availability of data can also make a timely

assessment of firms' innovation capabilities difficult. However, despite these shortcomings, the use of patents to assess the innovation profile of companies and geographical regions is widespread, especially in rapidly developing technologies and sectors such as STEP and NZIA technologies.

This paper presents a new methodology for identifying innovative companies in STEP and NZIA technologies. In line with recent research investigating unconventional methods[1], we rely on text mining techniques to identify specific types of innovations in patent data and track the STEP and NZIA content of the IPC codes associated with the patents. We aim to analyze the geographical presence and adoption of STEP and NZIA technologies in European companies (Arribas-Bel and Bakens, 2019; Gentzkow et al., 2019; Ash and Hansen, 2022).

To achieve this, we follow a sequential process of technology identification. First, we analyze the content of all patents published from 2002 to 2022, using PATSTAT as a data source. Then, we identify which patents are classified as STEP and NZIA applying text-mining techniques to the patents' information sources. Third, we identify IPC codes that can be categorized as STEP and NZIA codes using the IPC-patent association. Fourth, we search for IPC codes in the company's patent portfolio and evaluate the STEP NZIA intensity of the company's patent portfolio to describe the company's STEP and NZIA technology profile. Fifth, we match the company's technology profile with the company's financial data to map and describe which firms are innovating in these areas. Finally, we highlight the main sectors, geographical areas, and characteristics of innovative companies in the European economic panorama.

[1] Unconventional methods include those leveraging digital content and user-generated data (Gok et al., 2015; Ojanpera et al., 2017; Antons et al., 2020; Rammer and Es-Sadki, 2023), Natural Language Processing (NLP) techniques, and unstructured data sources (Nathan and Rosso 2015; Papagiannidis et al. 2018; Stich et al. 2022).

2. Empirical background

2.1. From conventional to unconventional approach

The measurement of innovation in companies has traditionally relied on conventional data sources such as surveys, financial statements, and patent data. These sources are fundamental in economic and innovation studies as they provide structured and standardized information on firms' innovation activities (Gault, 2013; OECD, 2015; OECD and Eurostat, 2018; Hall et al, 2010; Nagaoka et al., 2010; Mairesse and Mohnen, 2010). While these sources provide a coherent framework for analysis, they have several limitations, particularly in terms of their ability to capture more dynamic and intangible aspects of innovation.

For example, surveys are often limited by response bias, especially for small and medium-sized enterprises (SMEs). SMEs are only included in large innovation surveys - such as the Community Innovation Survey (CIS) - on a rotational basis, which makes longitudinal studies difficult. In addition, micro-enterprises with fewer than ten employees are not included at all, leading to significant gaps in innovation data for smaller firms (Arundel et al., 2013). Annual financial statements, another conventional source, provide critical insights into firms' investments in innovation, particularly research and development (R&D) expenditure. However, in SMEs, R&D expenditures are often informal or hidden in personnel costs, making them difficult to quantify (Santarelli and Sterlacchini, 1990). Patents have long been used as an indicator of technological innovation, as they provide a formal record of inventive activity (Griliches, 1990; Jaffe et al., 1993). However, the mere presence of patents within a firm cannot be a valid indicator of innovation due to the limitations associated with them (Balland and Boschma, 2020; McCann and Ortega-Argilés, 2015; Holgersson, 2013).

Given the limitations of conventional data sources, innovation studies have increasingly turned to unconventional approaches, such as large-scale text-mining in unstructured data sources. These methods have opened new avenues for understanding innovation by tapping into

underutilized or emerging data, including patent text analysis. Text mining allows researchers to extract detailed information about the technological content of patents, providing a more dynamic view of innovation processes than simple patent counts or classifications.

Text mining of patent documents involves analyzing the full text of patent applications - especially their titles and descriptions - to identify technological trends and track the development of specific innovations. This method enables a more granular and up-to-date understanding of innovations, particularly in areas, such as STEP and NZIA technologies. By using natural language processing (NLP) and machine learning, researchers can uncover patterns in technological development that would be difficult to detect using conventional patent metrics alone.

For these reasons, patent text analysis has become a key tool to overcome some of the limitations of traditional patent analysis. Antons et al. (2020) and Rammer and Es-Sadki (2023), for example, emphasize the potential of text-mining techniques to gain insights into the innovation process, especially in rapidly evolving sectors. Text mining of patents allows researchers to explore not only the number of patents filed, but also the nature and content of the innovations described in these patents.

In addition, applying topic modeling and other NLP techniques to patent data can help identify clusters of related technologies, providing a way to track the diffusion of innovations across industries and regions. Arora et al. (2020), for example, use text-mining to assess the dynamic capabilities of firms by analyzing their patent portfolios to understand how they adapt to new technological trends. This approach is particularly valuable in the context of emerging technologies related to STEP and NZIA, where innovation often takes the form of incremental advances that are difficult to capture using traditional measurement tools.

2.2. Text mining in patents data

As a structured form of textual data, patents are ideal for text-mining approaches that go beyond simple patent counting or classification. Recent advances in NLP and machine learning have enabled researchers to explore the textual content of patents and gain valuable insights into the specific technologies being developed and the relationships between different innovations.

One of the most promising areas of text mining in patents is the use of topic modeling to identify important technological trends. Topic modeling can uncover latent themes in patent texts, such as the development of certain environmental technologies or green innovations. This is particularly relevant for studies that track innovations in specific areas, such as the rapid development of new environmental solutions. For example, Papagiannidis et al. (2018) and Stich et al. (2022) apply text-mining techniques to investigate whether companies develop technologies that are in line with environmental policy and regulatory frameworks. Bishop et al. (2022) also show how advanced NLP algorithms can be used to classify companies based on their commitment to green technologies, providing a way to track the adoption and diffusion of these innovations.

In addition, text-mining enables the identification of specific keywords and phrases in patent texts that are associated with important technological advances. By analyzing the frequency and the context of these terms, researchers can track the emergence of new technologies and their impact on broader technological ecosystems. Guzman and Li (2023), for example, use text mining to measure the technological differentiation of companies based on their patent portfolios and how their innovations position them against competitors.

To summarize, text mining of patents is a powerful tool for overcoming the limitations of traditional innovation metrics. By analyzing the textual content of patent documents, researchers can gain deeper insights into the technological advances that drive innovation, especially in the context of STEP and NZIA technologies. The advancement of text-mining techniques holds

significant potential for providing timely, detailed, and dynamic analysis of innovation processes, paving the way for more informed policy decisions and strategic investments in new technologies.

3. Methods and Data

Our methodology for identifying innovative companies in the STEP and NZIA technologies comprises three different steps. First, we extract patent data from PATSTAT to build a comprehensive database with complete information on patents. Second, we define a method to analyze the content of each patent and classify it as "STEP or NZIA" rather than "other technology areas". Third, we build the patent portfolio at the company level and link the patent data to the characteristics of the companies.

3.1. The database PATSTAT

The patent database we use is the PATSTAT database from the European Patent Office (EPO).[2] PATSTAT is available in two versions: “PATSTAT Online” and PATSTAT as bulk data. On the one hand, “PATSTAT Online” is easily accessible via a web interface, but it has limitations such as read-only access and a limited number of patents that can be queried and exported simultaneously. On the other hand, the bulk data version provides PATSTAT in the form of multiple CSV files. This version is unrestricted but requires its hardware to operate and does not offer a web interface. We used the bulk data version in our analysis.

PATSTAT provides a wide range of information about patents, including bibliographic information (e.g., filing date, number of applicants, etc.) and patent classifications based on content, technological field, or industry. The classifications follow systems such as IPC2, CPC3

[2] PATSTAT is a statistical patent database provided by the EPO since 2006. Starting from the mid-19th century, it contains information on over 100 million patents from more than 90 patent authorities worldwide. Additionally, the database provides information on several hundred million legal events (e.g., changes in ownership, fee payments, etc.) related to the respective patents, provided by over 45 patent authorities.

or NACE4. Additional information is also provided, such as the number of citations, applicants and inventors, and the patent family to which the patent belongs (e.g., family affiliation, size, etc.). Importantly, PATSTAT also contains textual information such as the title or abstract of the patent.

To contain all this information, PATSTAT is based on a relational database consisting of 29 separate tables, each of which has a specific thematic focus (e.g., citation information). The tables are linked to each other via various identifiers that serve as key variables and link a patent to a range of information (e.g., classifications or citations). For example, the "TLS201_APPLN" table contains basic bibliographic data, while the "TLS229_APPLN_NACE2" table provides information on the sector to which the patent belongs. To illustrate the PATSTAT structure, Figure 1 shows an excerpt from the relevant PATSTAT tables and key attributes for the assignment of patent information to specific companies. Table TLS201_APPLN is the main table and contains basic information about the patent, which is uniquely identified by the application identifier (appln_id). This includes, for example, the patent authority (appln_auth), the filing date (appln_filing_date), etc. It also contains information on whether the patent has been granted or not (granted).

The table TLS202_APPLN_TITLE contains information about the title of each patent. In particular, this table contains the English title of the application, if available (APPLN_TITLE), or in another language if the English title is not available. The language of the title of the application selected and loaded for PATSTAT is indicated as APPLN_TITLE_LG. This table also contains the APPLN_ID to merge the information from table TLS202_APPLN_TITLE with that from table TLS201_APPLN.

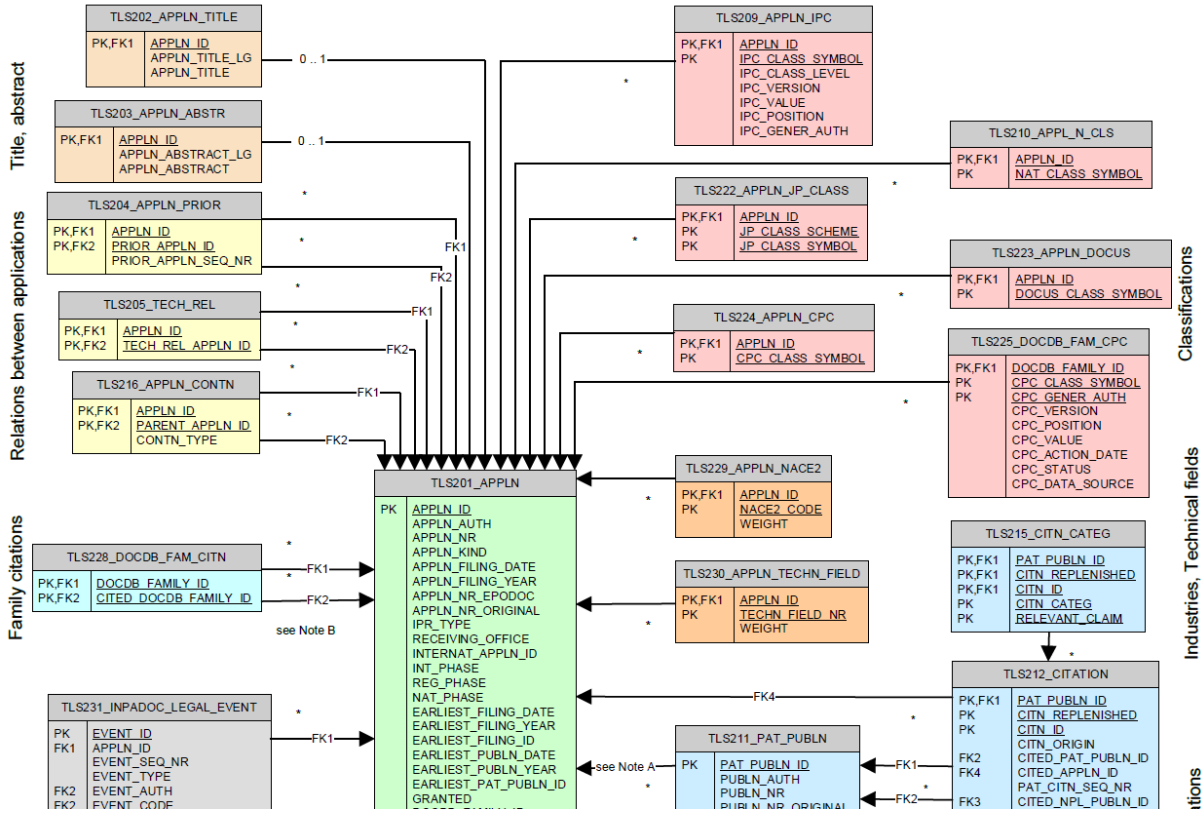
Similarly, table TLS203_APPLN_ABSTR contains the English summary, if available. If no English summary is available, it contains the most recent summary in another language. Like the title table, this table also consists of three parts: APPLN_ID to merge the table with TLS201_APPLN, APPLN_ABSTRACT_LG with the language of the abstract of the application

selected and loaded in PATSTAT, and APPLN_ABSTRACT with the full text of the abstract of the application.

It is also important to note that some tables can be linked directly to the main table TLS201_APPLN. In these cases, the key for the merge is usually APPLN_ID. In other cases, however, the connection is not so direct, and it is necessary to go through other secondary tables to reach TLS201_APPLN. In any case, each table has one or more columns that are repeated in other files so that the different information can be unified.

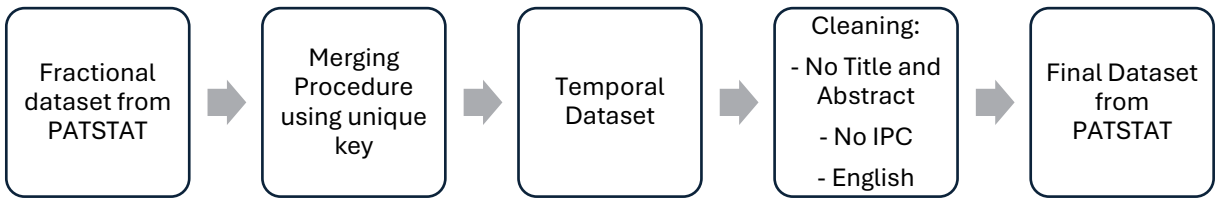
In our case, the first step in building the final database is to create a unique dataset containing all patents filed at the European Patent Office (recognized as EP patents) from 2002 to 2022. To achieve this, we use several merging procedures, taking advantage of the fact that each table has at least one common key with the others. In this way, we can create a dataset that contains the following information: application ID (APPLN_ID), application authorities (APPLN_AUTH), application number (APPLN_NR), application kind (APPLN_KIND), application year (APPLN_FILING_YEAR), application title (APPLN_TITLE), application abstract (APPLN_ABSTRACT), and IPC classification (IPC_CLASS_SYMBOL).

Figure 1: Example of tables from PATSTAT dataset.



When creating the database, we carried out three cleansing steps. First, we remove patents that do not have at least one title or abstract, as our next step will involve a detailed text analysis of these parts. Next, we exclude patents without IPC codes, as we want to find out which IPC codes contain patents related to STEP and NZIA technologies. Third, we exclude patents whose titles or abstracts were published in a language other than English. Our final database from PATSTAT comprises 10,006,326 observations. The main results of this method are summarized in Figure 2 below.

Figure 2: Summary of dataset's creation from PATSTAT.



After creating the first dataset focusing on patents registered at the European Patent Office as the subject of the text mining activity, we repeat the same procedure to consider all existing patents in the PATSTAT bulk version from 2002 to 2022. Given the large amount of data, this procedure is challenging from a software management perspective.

3.2. The Text Mining procedure

All data was read with Python 3.11.9. We uploaded the matrix containing the list of information described above for each patent in the dataset. To identify the IPCs related to STEP and NZIA technologies, we also uploaded a list of keywords related to these technologies (e.g., “smart grids”, “solar thermal”, “artificial intelligence”, etc.) to run the classification algorithm.

Before starting the text mining procedure, several critical aspects need to be emphasized. Since we are working with extremely noisy and unstructured texts, we need to perform text processing. First of all, punctuation is crucial. Since texts often lack uniform punctuation, we have developed a strategy to standardize punctuation by replacing the most important characters (points, commas, dash, parentheses, etc.), the new line symbol \n, and the new tab symbol \t with a space. Second, we made sure that the entire text was converted into a list of individual words to facilitate sequential reading. Third, we implemented an algorithm that not only reads individual words (e.g., "LEO") but also pairs of words (e.g., "smart grid") and triplets (e.g., " Low Earth Orbit").[3] Finally, since we are interested in finding words in both the patent’s title and the patent’s abstract, we built a routine to sum the words found in the title and abstract, rather than repeating the text mining procedure twice.[4]

[3] For instance, in the case of word pairs, we are not interested in capturing just "artificial" or just "intelligence". Instead, when the scraping algorithm reads "intelligence," it recognizes it as part of a word pair and also reads the immediately preceding word.

[4] For our analysis, we did not need to resort to BERT - Bidirectional Encoder Representations from Transformers (Devlin et al., 2018) - or other contextualization algorithms. While such algorithms are

We have searched for every single keyword in the dataset and got a CSV output with a list of IPC codes summarizing all patents containing a specific keyword. The results are compiled in tabular form. At this stage of the methodology, specific metrics are determined, such as the frequency with which each keyword appears in the entire group of patents belonging to the same IPC and the overall frequency. We then refine the results by cleaning up the raw results. We decided to aggregate the keywords according to two criteria. The first criterion is the singular and plural form. After searching for keywords in singular/plural and masculine/feminine form, we combine them into a single keyword and add up their frequencies.

The second criterion relates to acronyms. After searching for both the acronym (e.g., "LEO") and the full term (e.g., "Low Earth Orbit"), we simply add up their frequencies. Table 1 shows an example of a results table, while Figure 3 summarizes the proposed methodological framework for the PATSTAT dataset and the Python dataset.

Table 1: Example of the table collecting the frequencies of STEP NZIA keywords.

IPC Code	$Word1_s$	$Word1_p$	$Word2_s$	$Word2_p$	...	$WordN_p$	Tot. Keyword
D06M 13/325	0	31	2	1	...	18	$\sum_{i=1}^{n} Word_i = 52$
C13B 50/00	6	11	0	10	...	0	$\sum_{i=1}^{n} Word_i = 27$
...	...	...	...	...	...	...	...
A63B 69/02	100	0	45	0	...	0	$\sum_{i=1}^{n} Word_i = 145$

Note: Example of the resulting table after the text mining activity conducted with Python 3.11.9. The rows contain the different IPC codes, while the columns represent the keywords. Each cell contains the number of times a keyword appears within all patents related to a specific IPC code. Source: authors' elaboration.

generally crucial for contextualizing words and distinguishing their meanings in different contexts, this issue does not arise in our case. The words used pertain specifically to aspects of NZIA technologies, making it reasonable to assume they do not refer to other contexts beyond our intended scope.

Figure 3: Process to generate the dataset from Python.

3.3. From Patent Data to Company Data

Although PATSTAT contains all relevant information on patents, it lacks company-level information that is essential for understanding where STEP and NZIA technologies are located in the European business landscape. Therefore, a significant part of the data processing consists of using other databases that contain company-level variables and linking their records to the patent data.

To overcome this problem, we rely on the Orbis database (Moody's - Bureau van Dijk) to collect information on companies. Based on Orbis, we have obtained two different datasets. From Patent Orbis we download all companies that fulfill the following criteria: (1) the company was active in 2022; (2) the company is located in the European Union (27 countries); and (3) the company has at least one patent. We obtained 189,139 companies that reflect these criteria. For each observation, we have information on the application number, the company's BvD ID, and the four-digit IPC codes of its patents' portfolio. We keep the four-digit IPC codes at this stage. However, we remove them after the matching with PATSTAT as they have limited functionality in our case, and we delve into the granular 14-digit IPC codes. Second, we obtained qualitative and quantitative information from the Orbis database. In particular, detailed information is collected on the sector, size, geographical location, BvD identifier, and economic and financial data.

Once the sample of companies was drawn, the next step was to retrieve their patent data. This process essentially consists of performing a series of steps to create a unique dataset containing all the information on both patents and companies. The process focuses primarily on

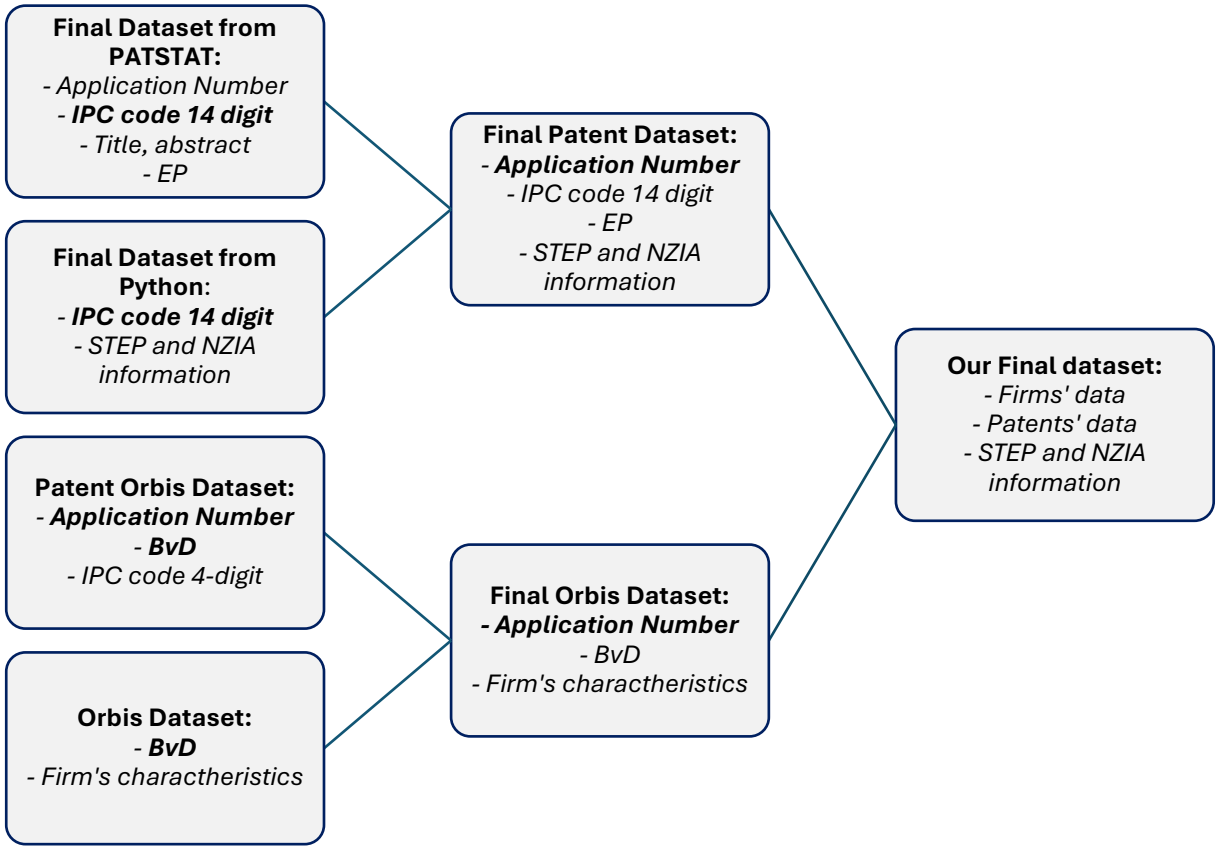
the information contained in the final dataset from Python and PATSTAT to merge these data with that from Patent Orbis.

As mentioned above, this process is complicated due to the fragmentation of the different datasets and requires a multi-step process. First, we need to match the data obtained from Patent Orbis and Orbis. Since in both datasets each company is uniquely represented by the BvD identifier, we can use this as a key to merge the data and create a unique dataset. In this way, for each BvD ID, we obtain the company name, all characteristics of the named companies, and the company's patent portfolio. As said above, we remove information on four-digit IPC codes, as we secondly keep only the 14-digit IPCs.

Next, we turn to the final extended dataset from PATSTAT and the final dataset from Python. Again, we need to find a common key to merge the two datasets. The common key is the application number. By merging using the APPLN_NR, we obtain a dataset that contains for each application number: the 14-digit IPC codes, the title, the abstract, whether it is an EP patent or not, and, most importantly, whether it is a STEP and NZIA patent. In addition, the number of occurrences of each keyword is stored for each IPC code. At this stage, the dataset comprises more than 250 million observations. To be able to manage this dataset, we remove the title and abstract as they are no longer relevant at this stage of the analysis but keep the rest of the data.

Finally, the third step is to assign the patents to the companies. To do this, we can use the company application numbers extracted from Orbis, and the application numbers extracted from PATSTAT. To enable the matching by using the application numbers, it is necessary to create an additional data set. This is because we need to drop the duplicate application numbers in the company dataset. This file links the company dataset with the patent dataset. We then perform the third merge procedure to create the final dataset, which is the focus of this paper. This dataset links the companies to the patents and contains all the necessary information for a quantitative analysis of both. The following Figure 4 gives an overview of the different phases and merges.

Figure 4: Summary of the main steps and merges to obtain the final dataset.



4. Results

4.1. Descriptive analysis of the dataset

After the text-mining procedure, we get information about all patents published from 2002 to 2022. If we group all the results by IPC codes, we cover a total of 73,290 codes. This means that the previously shown Table 1 contains a total of 73,290 rows and 118 columns. These are the 118 keywords we used in our text-mining analysis to identify the STEP and NZIA technologies. The last column of the table contains the total number of keywords found for each IPC code. We used this information in the first phase of the analysis. We will use the more detailed results for each keyword for further analysis.

The results show that 33,558 IPC codes out of 73,290 contain at least one of the words related to STEP and NZIA technologies. However, this is still not sufficient to identify the IPC codes

that can be considered STEP and NZIA. There are two reasons for this statement. The first reason is related to the rationality of the selection. If a word occurs only once in the text of a single patent, it would be too strong an assertion that the IPC code associated with that patent is STEP and NZIA. Therefore, we need to avoid too low frequencies associated with a particular keyword or pure randomness in the data. The second reason is related to the distribution of the results. Since the frequency distribution of the words is quite broad (ranging from one word to more than 6,500 occurrences in several keywords), further filtering of the dataset is required. In this context, the data reveals that 12,988 IPCs meet or exceed the threshold of 10 words. For IPCs with a minimum of 50 words, the number decreases to 3,720 IPCs, and further declines to 1,895 IPCs for those containing at least 100 words. Only 290 IPCs include a minimum of 500 words, and just 102 IPCs contain at least 1,000 words. This pattern demonstrates a clear decrease in the number of IPCs as the word count threshold increases.

Given this distribution, in this descriptive analysis, we consider only the IPC codes with at least 1000 keywords referring to STEP and NZIA as relevant codes (102 IPC codes). We will provide more detailed results based on different values of the distribution in future analysis.

4.2. Geographical distribution

The construction of our dataset is limited to companies with patents located in the European Union. Furthermore, only companies that were active at the time the dataset was constructed, i.e., in 2024, are included in the dataset. After the matching between PATSTAT and Orbis data sets, we identified 109,888 European companies (down from 189,139 in the original Orbis sample) that own at least one patent. After matching the BvD ID from Orbis with the linkage dataset, which is composed of the application number, IPC codes, and the corresponding text-mining results on STEP and NZIA, we found that 2,340 European companies hold a patent with an IPC code classified as STEP and NZIA, which is 2.13 percent of the total sample.

Table 2 shows the frequency distribution of companies in the different European Union countries included in the sample, both at the overall level (Panel A) and for STEP and NZIA companies only (Panel B). The sample covers all 27 European countries, with companies in Germany, Italy, France, and Spain accounting for a significant proportion of European companies with patents. Companies in these countries make up around 60 percent of the sample. Looking at Panel B, Germany, France, Italy, and Spain continue to account for about 60 percent of STEP and NZIA companies, confirming their importance as innovative countries.

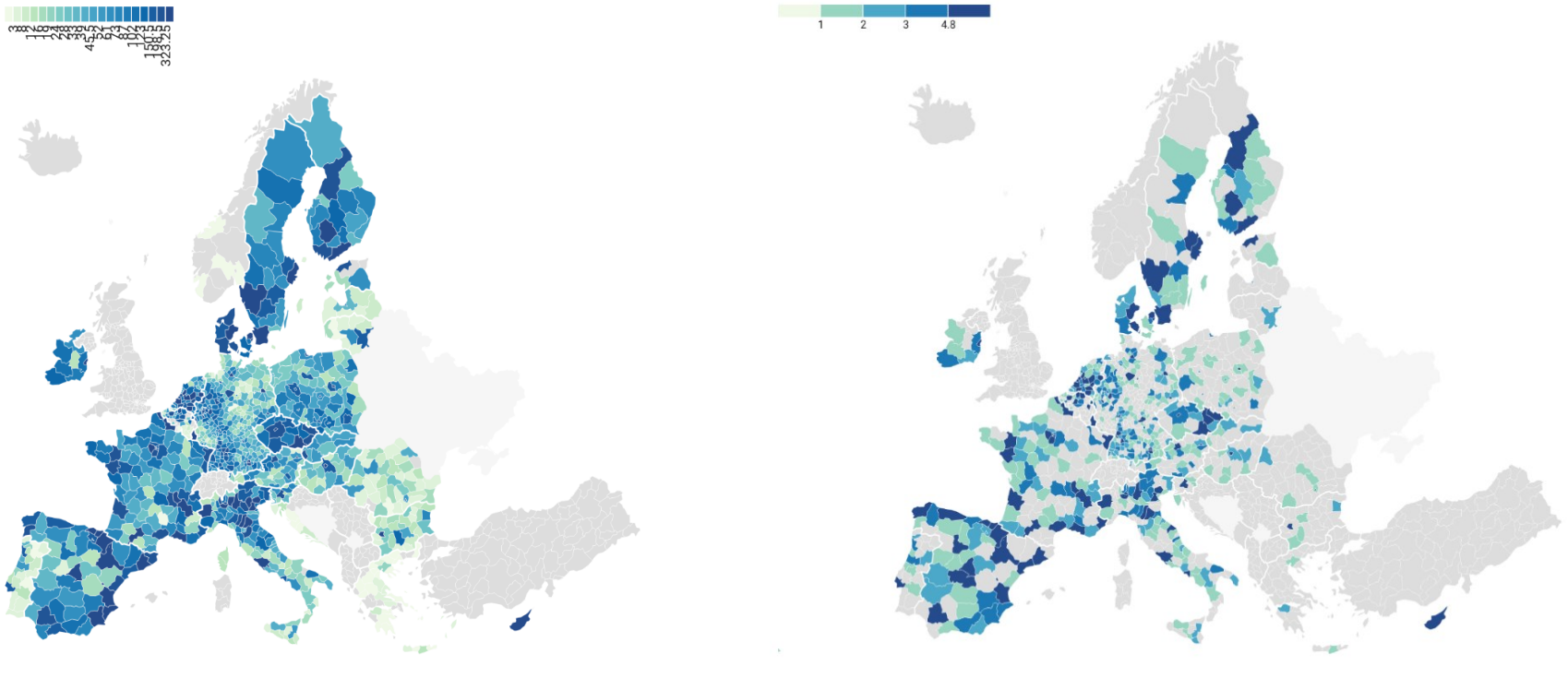
To obtain more precise information on the geographical distribution, Figure 5 shows the distribution of companies at the NUTS3 level for each country considered. The left-hand side of Figure 5 shows the overall distribution of European companies with patents, while the right-hand side shows the companies labeled STEP and NZIA. Figure 5 clearly shows the emergence of several concentration areas for these companies across Europe (dark blue) while other regions show either a low presence (light green) or no presence of STEP and NZIA companies.

Table 2: Firm distribution by country.

Panel A: Total sample			Panel B: STEP AND NZIA sample			
Country	Freq.	Freq. %	Country	Freq.	Freq. %	% of STEP NZIA companies on total firms in the country
Austria	3,099	2.82	Austria	73	3.12	2.36%
Belgium	2,884	2.62	Belgium	72	3.08	2.50%
Bulgaria	1,136	1.03	Bulgaria	21	0.9	1.85%
Croatia	192	0.17	Croatia	2	0.09	1.04%
Cyprus	371	0.34	Cyprus	17	0.73	4.58%
Czech Republic	3,233	2.94	Czech Republic	49	2.09	1.52%
Denmark	2,932	2.67	Denmark	68	2.91	2.32%
Estonia	327	0.3	Estonia	7	0.3	2.14%
Finland	3,203	2.91	Finland	77	3.29	2.40%
France	12,676	11.54	France	307	13.12	2.42%
Germany	28,795	26.2	Germany	657	28.08	2.28%
Greece	273	0.25	Greece	12	0.51	4.40%
Hungary	1,214	1.1	Hungary	19	0.81	1.57%
Ireland	1,687	1.54	Ireland	53	2.26	3.14%
Italy	13,108	11.93	Italy	231	9.87	1.76%
Latvia	191	0.17	Latvia	2	0.09	1.05%
Lithuania	271	0.25	Lithuania	2	0.09	0.74%
Luxembourg	567	0.52	Luxembourg	18	0.77	3.17%
Malta	150	0.14	Malta	5	0.21	3.33%
Netherlands	7,489	6.82	Netherlands	177	7.56	2.36%
Poland	6,573	5.98	Poland	65	2.78	0.99%
Portugal	650	0.59	Portugal	20	0.85	3.08%
Romania	932	0.85	Romania	17	0.73	1.82%
Slovakia	769	0.7	Slovakia	11	0.47	1.43%
Slovenia	728	0.66	Slovenia	7	0.3	0.96%
Spain	11,526	10.49	Spain	223	9.53	1.93%
Sweden	4,912	4.47	Sweden	128	5.47	2.61%
Total	109,888	100	Total	2,340	100	

Notes: The table reports the distribution of firms at the country level. On the left, the table exhibits the distribution of all firms in our dataset. On the right, the table shows the country distribution of STEP AND NZIA-labelled firms.

Figure 5. Firm distribution at NUTS3 level.



4.3. Sector analysis

Table 3 shows the sample disaggregated at the level of the main NACE sections. Enterprises operating in Sections C (Manufacturing), G (Wholesale and retail trade), and M (Professional, scientific, and technical activities) make up the majority of enterprises in the sample and, accordingly, the majority of STEP and NZIA companies.

However, it is worth noting that while companies in Section J (Information and Communication) account for only 5.42 percent of all companies, they make up around 22 percent of all STEP and NZIA companies, highlighting the relevance of Section J in the context of STEP and NZIA technologies.

Similarly, despite its small number of companies (0.58 percent of total companies and 2.8 percent of all STEP and NZIA companies), Section D (Energy supply) includes 10.3 percent of the companies in the sector that are associated with STEP and NZIA technologies. More detailed information on the individual NACE subsections can be found in Appendix A, Table A.1. Instead, Appendix A, Table A.2 shows the distribution of STEP and NZIA enterprises at the NACE level by disaggregating the main sections at 4 digits.

Table 3: Firms' distribution at NACE main section.

Panel A: Full sample			Panel B: STEP AND NZIA sample			
NACE main section	Freq.	%	NACE main section	Freq.	%	%
A	937	0.88	A	6	0.27	0.64%
B	231	0.22	B	0	0	0.00%
C	45,690	43.09	C	470	20.92	1.03%
D	614	0.58	D	63	2.8	10.26%
E	685	0.65	E	5	0.22	0.73%
F	4,010	3.78	F	73	3.25	1.82%
G	15,267	14.4	G	214	9.52	1.40%
H	793	0.75	H	10	0.45	1.26%
I	515	0.49	I	9	0.4	1.75%
J	5,751	5.42	J	499	22.21	8.68%
K	4,397	4.15	K	98	4.36	2.23%
L	2,074	1.96	L	33	1.47	1.59%
M	18,623	17.56	M	611	27.19	3.28%
N	3,174	2.99	N	73	3.25	2.30%
O	62	0.06	O	3	0.13	4.84%
P	820	0.77	P	32	1.42	3.90%
Q	922	0.87	Q	22	0.98	2.39%
R	419	0.4	R	6	0.27	1.43%
S	1,044	0.98	S	19	0.85	1.82%
T	5	0	T	0	0	0.00%
U	4	0	U	1	0.04	25.00%

Notes: The table provides the distribution of firms according to the main section of NACE Rev. 2. On the left-hand side, the table shows the overall distribution of firms in our dataset. On the right-hand side, tables show the distribution of STEP AND NZIA firms. The last column provides at the country level the ratio of STEP and NZIA firms over a total firm in a certain country.

4.4. Main differences between STEP NZIA firms and other firms

Finally, we describe the sample of firms identified by STEP and NZIA using three variables for firm size: number of employees, turnover, and total assets. We use log transformation to ensure comparability across firms.

Table 4 shows the main descriptive statistics for the three proxy variables for firm size. When comparing STEP and NZIA firms with non-STEP and NZIA firms, the mean values of turnover, employees, and total assets are higher than for the latter. Performing a t-test on the differences between the means of the two groups shows that these are significantly different from zero for all three proxies.

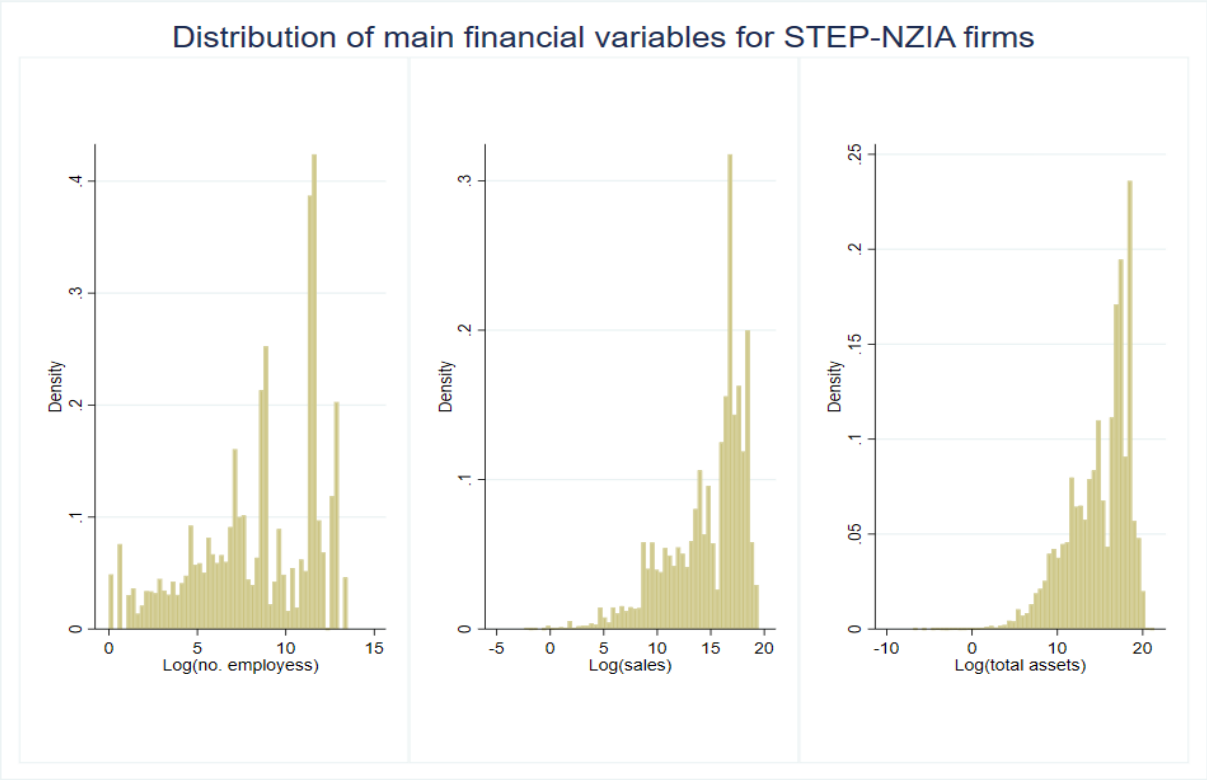
Table 4: Main descriptive statistics and t-test.

	NO STEP AND NZIA					STEP AND NZIA					T-test	
	N	Mean	Std.	Min.	Max.	N	Mean	Std.	Min.	Max.	Diff.	p-val.
Sales (log)	5640243	13.83	3.37	-6.91	19.45	280642	14.65	3.58	-2.41	19.45	-0.822	***
Emp (log)	6384604	7.644	3.31	0	13.49	297106	8.346	3.43	0	13.49	-0.702	***
Assets (log)	6428269	13.73	3.54	-7.12	21.7	312136	14.86	3.52	-6.91	21.36	-1.133	***

Notes: The table provides the main descriptive statistics splitting the sample into two groups: No STEP and NZIA firms and STEP and NZIA firms.

Figure 6 shows the distribution of these three variables when only STEP and NZIA companies are considered. Based on the frequency distribution shown in Figure 6, it can be shown that companies with STEP and NZIA labels are characterized by a larger size. As expected, company size is therefore a relevant factor when describing and searching for companies adopting these innovative technologies.

Figure 6: Distribution of financial variables for STEP and NZIA companies.



5. Conclusions

Nowadays, patents are regarded as one of the most powerful and widely used indicators of innovation in empirical studies (Wu et al., 2005; Benz et al., 2021; Berlemann and Jahn, 2016; Block and Spiegel, 2013). They are often used as a proxy for technological innovation. Patents have several advantages over alternative innovation metrics, in particular their public availability and quantitative nature.

However, to use patents effectively in research on business innovation, they need to be matched with firm-level data. However, this matching process is challenging. The reasons for this are manifold. Above all, matching requires considerable computing time: both the text-mining activities and the management of company and patent records are time-consuming, even with advanced computing tools. To avoid misclassifications and reduce the number of companies and patents affected by inaccurate matching procedures, merging these two data sets is technically challenging and, requires careful technical analysis.

We collect patent data for a selected sample of companies, we provide a new and unique dataset that identifies companies with STEP and NZIA technologies based on their patent portfolio. First, we developed a set of keywords for STEP and NZIA technologies. Then, instead of selecting IPC classes (Aiello et al., 2021; Ceipek et al., 2021), we performed text-mining using keywords in the abstracts and titles of patents included in PATSTAT and then clustered the IPC classes into a STEP and NZIA group. Our analysis covers the period from 2002 to 2022. At the same time, we collected information from Orbis Patent on patents held by active European companies and collected both qualitative (e.g., NACE codes, geographical location) and quantitative (e.g., financial data) company information. The publication and application numbers were recorded for each company. In the third step, all patents were merged into a unique dataset and the company-patent dataset was merged with the reconciliation file using the application number as a key. Since the application number appears multiple times in the Orbis patent dataset, we also created

a secondary match file consisting of unique application numbers and BvD IDs to ensure an accurate match between application numbers and IPC codes.

The final dataset includes almost 110,000 companies in the 27 countries of the European Union. For each company, we identify whether the company has STEP and NZIA technologies. In addition, the keyword matrix within each IPC allows us to set different thresholds for the definition of STEP and NZIA technologies and apply analytical weightings in our analyses. This dataset provides useful insights into the adoption of strategic technologies by European companies by mapping their geographic spread, industry affiliation, and key characteristics.

References

Aiello, F., Cardamone, P., Mannarino, L., Pupo, V. (2021). Green patenting and corporate social responsibility: Does family involvement in business matter?. Corporate Social Responsibility and Environmental Management, 28(4), 1386-1396. https://doi.org/10.1002/csr.2146

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

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

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

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. Research Policy, 30(4), 611-624. https://doi.org/10.1016/S0048-7333(00)00100-1

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

Ash, E., Hansen, S. (2022). Text algorithms in economics. Annual Review of Economics, 15, 659-688. https://doi.org/10.1146/annurev-economics-082222-074352

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

Benz, L., Block, J. H., Johann, M. S. (2021). Hidden champions as a determinant of regional development: An analysis of German districts. ZFW – Advances in Economic Geography. https://doi.org/10.1515/zfw-2020-0043

Berlemann, M., Jahn, V. (2016). Regional Importance of Mittelstand Firms and Innovation Performance. Regional Studies, 50(11), 1819-1833. https://doi.org/10.1080/00343404.2015.1058923

Bishop, A., Mateos-Garcia, J., Richardson, G. (2022). Using text data to improve industrial statistics in the UK. Economic Statistics Centre of Excellence (ESCoE) Discussion Papers ESCoE DP-2022-01, Economic Statistics Centre of Excellence (ESCoE).

Block, J., Miller, D., Jaskiewicz, P., Spiegel, F. (2013). Economic and technological importance of innovations in large family and founder firms: An analysis of patent data. Family Business Review, 26(2), 180-199. https://doi.org/10.1177/0894486513477454

Block, J. H., Spiegel, F. (2013). Family firm density and regional innovation output: An exploratory analysis. Journal of Family Business Strategy, 4(4), 270-280. https://doi.org/10.1016/j.jfbs.2013.10.003

Ceipek, R., Hautz, J., De Massis, A., Matzler, K., Ardito, L. (2021). Digital transformation through exploratory and exploitative internet of things innovations: the impact of family management and technological

diversification. Journal of Product Innovation Management, 38(1), 142-165. https://doi.org/10.1111/jpim.12551

Chirico, F., Criaco, G., Baù, M., Naldi, L., Gomez-Mejia, L. R., Kotlar, J. (2020). To patent or not to patent: That is the question. Intellectual property protection in family firms. Entrepreneurship Theory and Practice, 44(2), 339-367. https://doi.org/10.1177/1042258718806251

De Saá-Pérez, P., Díaz-Díaz, N. L., Luis Ballesteros-Rodriguez, J. (2012). The role of training to innovate in SMEs. Innovation, 14(2), 218-230. https://doi.org/10.5172/impp.2012.14.2.218

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Ejermo, O. (2009). Regional Innovation Measured by Patent Data—Does Quality Matter?. Industry and Innovation, 16(2), 141-165. https://doi.org/10.1080/13662710902764246

Gault, F. (Ed.), 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

Gok, A., Waterworth, A., Shapira, P. (2015). Use of web mining in studying innovation. Scientometrics, 102(1), 653–671. https://doi.org/10.1007/s11192-014-1434-0

Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey, Part II. National Bureau of Economic Research. http://doi.org/10.3386/w3301

Guzman, J., Li, A. (2023). Measuring founding strategy. Management Science, INFORMS, 69, 101–118. http://dx.doi.org/10.1287/mnsc.2022.4369

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

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

Jaffe, A. B., Trajtenberg, M., Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. The Quarterly Journal of Economics, 108(3), 577–598. https://www.jstor.org/stable/2118401

Laursen, K., Masciarelli, F., Prencipe, A. (2012). Trapped or spurred by the home region? The effects of potential social capital on involvement in foreign markets for goods and technology. Journal of International Business Studies, 43(9), 783-807. https://doi.org/10.1057/jibs.2012.27

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

McCann, P., Ortega-Argilles, R. (2015). Smart specialization, regional growth and applications to European Union cohesion policy. Regional Studies, 49(8), 1291–1302. https://doi.org/10.1080/00343404.2013.799769

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

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

OECD (2015). Frascati manual 2015: Guidelines for collecting and reporting data on research and experimental development. OECD Publishing. http://dx.doi.org/10.1787/9789264239012-en

OECD, Eurostat (2018). Oslo Manual 2018: Guidelines for Collecting, Reporting and using Data on Innovation, the Measurement of Scientific, Technological and Innovation Activities. OECD Publishing and Eurostat, http://dx.doi.org/10.1787/9789264304604-en

Ojanpera, S., Graham, M., Straumann, R. K., De Sabbata, S., Zook, M. (2017). Engagement in the knowledge economy: Regional patterns of content creation with a focus on sub-Saharan Africa. Information Technologies and International Development, 13.

Papagiannidis, S., See-To, E. W., Assimakopoulos, D. G., Yang, Y. (2018). Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the internet age?. Computers & Operations Research, 98, 355–366. https://doi.org/10.1016/j.cor.2017.06.010

Rammer, C., Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators - a literature review. Technological Forecasting and Social Change, 197, 122874. http://dx.doi.org/10.1016/j.techfore.2023.122874

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

Wu, H. L. (2008). When does internal governance make firms innovative? Journal of Business Research, 61(2), 141-153. https://doi.org/10.1016/j.jbusres.2007.06.010

Wu S., Levitas, E., Priem, R. L. (2005). CEO tenure and company invention under differing levels of technological dynamism. Academy of Management Journal, 48(5), 859-873. https://doi.org/10.5465/amj.2005.18803927

APPENDIX A

Table A.1 – Details on main section NACE.

Main section NACE	Extended description
A	Agriculture, forestry, and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam, and air conditioning supply
E	Water supply; sewerage, waste management, and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication services
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific, and technical activities
N	Rental, travel agencies, business support services
O	Public administration and defense; compulsory social security
P	Education
Q	Healthcare and social assistance
R	Arts, entertainment, and recreation
S	Other service activities
T	Activities of households as employers; undifferentiated activities
U	Extraterritorial organizations and bodies

Notes: The table provides a detailed description of the main section of NACE.

Table A.2 – Frequency of STEP NZIA firm at NACE 4-digit level.

NACE Rev. 2 code (4-Digit)	Freq.	Freq. %	NACE Rev. 2 code (4-Digit)	Freq.	Freq. %	NACE Rev. 2 code (4-Digit)	Freq.	Freq. %
119	1	0.04	3101	1	0.04	5630	3	0.13
121	1	0.04	3109	2	0.09	5800	1	0.04
141	1	0.04	3200	1	0.04	5811	2	0.09
143	1	0.04	3230	1	0.04	5814	4	0.18
146	1	0.04	3240	1	0.04	5819	1	0.04
322	1	0.04	3250	4	0.18	5820	3	0.13
1031	1	0.04	3290	1	0.04	5821	1	0.04
1062	1	0.04	3299	2	0.09	5829	54	2.4
1071	1	0.04	3300	1	0.04	5911	2	0.09
1082	1	0.04	3312	9	0.4	5913	1	0.04
1086	4	0.18	3314	3	0.13	5920	3	0.13
1089	2	0.09	3315	1	0.04	6010	1	0.04
1102	1	0.04	3319	1	0.04	6110	4	0.18
1300	1	0.04	3320	9	0.4	6120	3	0.13
1320	1	0.04	3500	3	0.13	6190	14	0.62
1391	1	0.04	3510	1	0.04	6200	21	0.93
1520	2	0.09	3511	41	1.82	6201	211	9.39
1610	1	0.04	3513	7	0.31	6202	72	3.2
1623	1	0.04	3514	2	0.09	6203	6	0.27
1629	1	0.04	3520	1	0.04	6209	65	2.89
1712	1	0.04	3521	2	0.09	6311	15	0.67
1721	1	0.04	3522	3	0.13	6312	10	0.45
1812	3	0.13	3523	2	0.09	6390	2	0.09
1813	2	0.09	3530	1	0.04	6399	3	0.13
1814	2	0.09	3811	1	0.04	6419	3	0.13
1820	1	0.04	3820	1	0.04	6420	71	3.16
1920	1	0.04	3821	1	0.04	6430	4	0.18
2000	3	0.13	3900	2	0.09	6491	1	0.04
2011	3	0.13	4110	2	0.09	6492	1	0.04
2013	2	0.09	4120	10	0.45	6499	7	0.31
2014	3	0.13	4221	2	0.09	6530	1	0.04
2016	1	0.04	4299	7	0.31	6612	1	0.04
2030	1	0.04	4310	1	0.04	6619	6	0.27
2041	1	0.04	4311	2	0.09	6622	1	0.04
2059	7	0.31	4312	2	0.09	6630	2	0.09
2100	2	0.09	4321	21	0.93	6800	1	0.04
2110	7	0.31	4322	9	0.4	6810	6	0.27
2120	31	1.38	4329	2	0.09	6820	22	0.98
2221	1	0.04	4332	3	0.13	6831	1	0.04
2222	1	0.04	4333	1	0.04	6832	3	0.13
2229	13	0.58	4334	1	0.04	6910	6	0.27
2312	1	0.04	4391	6	0.27	6920	1	0.04
2319	2	0.09	4399	4	0.18	7000	1	0.04
2320	1	0.04	4500	1	0.04	7010	37	1.65
2331	1	0.04	4510	2	0.09	7021	2	0.09

(continue to the next page)

2332	1	0.04	4511	1	0.04	7022	53	2.36
2344	1	0.04	4519	1	0.04	7110	5	0.22
2361	5	0.22	4520	4	0.18	7111	4	0.18
2370	1	0.04	4531	2	0.09	7112	162	7.21
2399	1	0.04	4540	1	0.04	7120	15	0.67
2420	1	0.04	4610	4	0.18	7200	6	0.27
2434	1	0.04	4612	1	0.04	7210	14	0.62
2442	4	0.18	4614	10	0.45	7211	74	3.29
2445	1	0.04	4617	1	0.04	7219	166	7.39
2453	1	0.04	4618	5	0.22	7220	6	0.27
2511	17	0.76	4619	1	0.04	7310	4	0.18
2512	3	0.13	4621	2	0.09	7311	12	0.53
2521	1	0.04	4623	1	0.04	7312	1	0.04
2529	2	0.09	4636	1	0.04	7320	4	0.18
2530	1	0.04	4638	4	0.18	7410	7	0.31
2550	5	0.22	4642	1	0.04	7420	2	0.09
2560	2	0.09	4643	11	0.49	7430	1	0.04
2561	5	0.22	4645	1	0.04	7490	28	1.25
2562	8	0.36	4646	16	0.71	7711	1	0.04
2572	1	0.04	4649	2	0.09	7712	1	0.04
2573	1	0.04	4650	1	0.04	7721	2	0.09
2590	1	0.04	4651	12	0.53	7722	1	0.04
2594	3	0.13	4652	8	0.36	7731	1	0.04
2599	7	0.31	4660	1	0.04	7732	1	0.04
2610	1	0.04	4661	1	0.04	7733	1	0.04
2611	26	1.16	4663	1	0.04	7735	1	0.04
2612	2	0.09	4665	1	0.04	7739	7	0.31
2620	14	0.62	4666	1	0.04	7740	25	1.11
2630	6	0.27	4669	32	1.42	7810	1	0.04
2640	1	0.04	4671	2	0.09	7820	1	0.04
2651	18	0.8	4673	2	0.09	7912	1	0.04
2660	4	0.18	4674	6	0.27	8020	2	0.09
2670	4	0.18	4675	4	0.18	8110	1	0.04
2680	1	0.04	4676	1	0.04	8121	1	0.04
2700	1	0.04	4677	1	0.04	8130	1	0.04
2711	11	0.49	4690	16	0.71	8200	1	0.04
2712	14	0.62	4710	1	0.04	8220	1	0.04
2720	23	1.02	4719	6	0.27	8230	3	0.13
2732	1	0.04	4724	1	0.04	8299	19	0.85
2733	2	0.09	4725	2	0.09	8411	1	0.04
2740	1	0.04	4726	1	0.04	8412	2	0.09
2751	3	0.13	4729	2	0.09	8540	1	0.04
2790	20	0.89	4730	4	0.18	8541	2	0.09
2800	3	0.13	4741	6	0.27	8542	20	0.89
2812	2	0.09	4742	2	0.09	8551	2	0.09
2813	3	0.13	4743	2	0.09	8559	6	0.27
2815	1	0.04	4751	1	0.04	8560	1	0.04
2820	2	0.09	4752	3	0.13	8610	10	0.45

(continue to the next page)

2821	9	0.4	4759	3	0.13	8621	2	0.09
2822	4	0.18	4763	1	0.04	8622	3	0.13
2825	3	0.13	4770	1	0.04	8623	1	0.04
2829	14	0.62	4771	1	0.04	8690	5	0.22
2830	4	0.18	4772	1	0.04	8899	1	0.04
2841	3	0.13	4773	3	0.13	9001	1	0.04
2849	6	0.27	4774	1	0.04	9002	1	0.04
2890	1	0.04	4775	1	0.04	9101	1	0.04
2891	1	0.04	4778	1	0.04	9200	2	0.09
2893	2	0.09	4791	6	0.27	9329	1	0.04
2896	1	0.04	4799	1	0.04	9412	1	0.04
2899	23	1.02	4910	1	0.04	9420	1	0.04
2910	1	0.04	4939	1	0.04	9499	2	0.09
2920	1	0.04	4941	4	0.18	9511	1	0.04
2931	2	0.09	4942	1	0.04	9604	1	0.04
2932	5	0.22	5221	2	0.09	9609	13	0.58
3020	2	0.09	5229	1	0.04	9900	1	0.04
3099	1	0.04	5510	2	0.09			
3100	1	0.04	5610	4	0.18	**Total**	**2247**	**100**

Notes: The table provides the frequency distribution at NUTS3 of STEP and NZIA-labelled firms.