EIF Research
and Market Analysis
Working Paper 2023/91

Using machine learning to map the European Cleantech sector

Matteo Ambrois, Vincenzo Butticè, Federico Caviggioli, Giovanni Cerulli, Annalisa Croce, Antonio De Marco, Andrea Giordano, Giuliano Resce, Laura Toschi, Elisa Ughetto, Antonio Zinilli

Matteo Ambrois
Politecnico di Milano

Vincenzo Butticè
Politecnico di Milano

Federico Caviggioli
Politecnico di Torino

Giovanni Cerulli,
IRCrES-CNR

Annalisa Croce
Politecnico di Milano

Antonio De Marco
Politecnico di Torino

Andrea Giordano
Politecnico di Torino

Giuliano Resce
Università del Molise

Laura Toschi
Università degli Studi di Bologna

Elisa Ughetto
Politecnico di Torino

Antonio Zinilli
IRCrES-CNR

Editor:
Helmut Kraemer-Eis,
Head of EIF's Research & Market Analysis, Chief Economist

Contact:
European Investment Fund
37B, avenue J.F. Kennedy, L-2968 Luxembourg

Tel.: +352 248581 596
http://www.eif.org/news_centre/research/index.htm
Luxembourg, July 2023

Scan above to obtain a PDF version of this working paper

Disclaimer

Preface

This paper is the introductory chapter of a series of analyses that will result from the CLEU[1] project, a collaboration between the universities of Politecnico di Torino, Politecnico di Milan and Università degli Studi di Bologna. The project focuses on Cleantech, an industry sector that develops and deployes sustainable and environmentally friendly solutions for various target applications. It aims to: i) analyse the actions that are undertaken by European Cleantech firms to engage in transformative climate and innovation actions to align with the European Green Deal-inspired policies; ii) examine the association of environmental innovation and the number of new investments made by venture capital (VC) investors in Cleantech companies on environmental indicators; iii) analyse the enabling factors for the development of European Cleantech firms, with a focus on EU-level and country-level targeted policies and regulations and the different sources of financing; iv) analyse the extent to which the implementation of policies and regulations affect both the propensity of cleantech firms to seek external equity financing and the equity offering by VC funds.

This paper presents a methodology to identify Cleantech companies in Bureau van Dijck's Orbis, a business database that provides detailed balance sheet information on millions of companies worldwide. The resulting dataset will shed a novel light on the European Cleantech sector, which thus far has mostly been analysed from the perspective of investment databases. The subsequent analyses of the CLEU project will rely heavily on the Cleantech data outlined in this paper.

This project was funded by the European Investment Bank (EIB)'s University Research Sponsorship (EIBURS) programme. The EIBURS provides grants to help EU universities and academic research centres to develop activities in selected research areas in addition to those that would normally be carried out by the beneficiary and on topics of major interest to the EIB Group (European Investment Bank and European Investment Fund). The CLEU project is coordinated by the Research and Market Analysis Division of the European Investment Fund (EIF).

Understanding the financing needs of the EU Cleantech sector is of particular importance to the EIF for several reasons. Firstly, Cleantech companies often require significant upfront capital investments due to the high costs associated with developing and scaling clean technologies. By understanding their financing needs, the EIF can tailor its funding programs and financial instruments to provide appropriate support, such as VC, loans or guarantees, to help these companies overcome financial barriers and access the necessary capital.

The EIF plays a crucial role in promoting Cleantech companies and initiatives in Europe. As the EIB Group's specialist provider of risk finance to benefit SMEs across Europe, the EIF provides financing and support to enhance access to capital for Cleantech startups and businesses. Today, the EIF offers various financial instruments, such as VC funds, equity investments and loan guarantees, specifically targeting Cleantech sectors including renewable energy, energy efficiency, sustainable mobility, and circular economy. Currently, the EIF manages several initiatives that aim to stimulate investments in EU Cleantech companies. For example, the InvestEU SME Window,

[1] The **Cl**eantech industry in the **Eu**ropean Green Deal: policy challenges and the finance landscape for SMEs: https://shorturl.at/jqZO2

through its Climate and Environment products, contains a EUR 900m pocket to increase access to equity finance for innovative SMEs that develop or adopt Cleantech solutions, while the EIF's RCR mandate provides EUR 300m annually, on average, over the period EUR 2022-2027, to European funds investing in Cleantech companies. This amount was recently topped up by approximately EUR 300m per annum through the Commission's REPowerEU plan.

By expanding our understanding of the Cleantech sector through the CLEU project, policymakers can further improve the design of targeted support schemes to accelerate the adoption of clean technologies, reduce greenhouse gas emissions, improve environmental quality, and promote sustainable resource utilization, thus driving the European green transition and positioning the EU at the forefront of the global Cleantech industry.

We therefor invite you to delve into this introductory analysis and explore the first findings of the CLEU project and wish you an informative and engaging reading experience.

Helmut Kraemer-Eis
Editor

Wouter Torfs
Project coordinator

Non-technical Summary[2]

This study provides a new perspective on European Cleantech, a sector that develops and deployes sustainable and environmentally friendly solutions for various target applications. It presents a novel solution to classify Cleantech companies based on a supervised machine learning (ML) algorithm applied to the extended business description of European companies, as found in Bureau van Dijck's Orbis database, a comprehensive and global business database that provides detailed information on millions of companies worldwide.

The process of using ML to classify Cleantech companies based on business descriptions involves a two-step approach. First, a small set of companies is extracted from the database and manually identified and labeled as Cleantech or non-Cleantech. This labeled dataset serves as a training set for machine learning models. By analysing the training data, the machine learning model can learn to assign a probability or confidence score to new, unseen company descriptions, indicating the likelihood of them belonging to the Cleantech category. The model essentially learns to generalise from the patterns it observed in the training data and to apply that knowledge to classify new descriptions. For example, it might learn that terms like "sustainable," "renewable," "energy efficiency," "waste management," "environmental conservation," or "carbon footprint reduction" are often indicative of Cleantech businesses. Once the model is trained and validated, it can be deployed to automatically classify large volumes of company descriptions, helping researchers, investors, or policymakers to quickly identify and analyse Cleantech companies.

The resulting dataset will shed new light on the European Cleantech sector. Earlier studies, typically based on investment databases, provided only a partial perspective of the Cleantech phenomenon, as such databases only include Cleantech companies that have been involved in an investment transaction. By employing a general database of administrative balance sheet data with coverage of the vast majority of the population of companies, such as the one applied in this paper, we are able to broaden the scope of our analysis. Furthermore, matching our sample of Cleantech companies to a variety of other databases led to a number of new valuable insights into the European Cleantech sector, related to sectoral and geographic distribution, innovative capacity, size, VC investment activity, and others.

Comparing our newly developed Cleantech classification to the traditional NACE sector classification, we found that Cleantech companies are predominantly active in the manufacturing, wholesale and retail trade, water supply and waste management, and construction sectors. Examining the spatial distribution of Cleantech in Europe, Germany, Italy, and France emerge as the key countries with the highest concentration of Cleantech companies. We also found Cleantech to be a well-established phenomenon, pre-dating to a large extent the two important Cleantech investment cycles, as a significant portion of the companies were established before

[2] This paper benefited from comments and input of Adelaide Cracco, Merilin Hörats, Helmut Kraemer-Eis, Andrea Hermida Parapar, Wouter Torfs and Virginie Varga. All remaining errors are our own.

the 2000s. We also analysed patenting activity of our Cleantech sample and found that Austria's Cleantech ecosystem is the most innovation-intensive, followed by Sweden and Germany, with sustainable energy production, energy-efficient industrial technologies, and air/water/soil pollution being the prominent technological categories for patenting. Investigating a selection of essential financial key performance indicators (KPIs) led us to conclude that cleantech innovators tend to operate at a larger scale compared to their ecosystem counterparts, in terms of total assets, sales and employee count. Finally, concerning VC financing, Finland, Sweden, France and Spain emerge as the geographical areas with a high concentration of VC-backed companies

Table of Contents

1 | Introduction

The European Union (EU) has long pursued a leading role in the development of policies to address climate change and environmental degradation. Recently, the Commission introduced the European Green Deal, which provides a roadmap to steer the EU towards a path of sustainable development. In addition to the 2050 carbon neutrality commitment, the Green Deal focuses on a broad range of environmental targets.

The European Green Deal emphasises the importance of green innovations and the role of private companies that focus on the development of sustainability solutions (Cleantech, also commonly referred to as Greentech).[3] Cleantech companies play a critical role in addressing environmental challenges such as climate change, pollution and resource depletion. Cleantech companies have the potential to drive economic growth by creating new jobs, generating revenue and attracting investment. By identifying and supporting Cleantech companies, policymakers can help foster environmental innovation, which is likely to generate positive economic ripple throughout the broader economy, boosting growth and employment. Moreover, supporting Cleantech companies can safeguard Europe's competitive position on global markets, where clean technology is likely to gain significant importance.

Identifying Cleantech companies is challenging, as there is no universally accepted definition of what constitutes Cleantech. Current definitions of Cleantech are often too generic and broad, leading to confusion and inconsistencies in how the industry is defined.[4] Existing classification methods based upon industry labels (i.e., NACE codes) have proven to be inefficient in properly identifying Cleantech firms, as they are not able to capture the cross-cutting nature of the sector (Christensen & Hain, 2017; Criscuolo & Menon, 2015; Cumming et al., 2016). Other classification attempts, such as the EU Taxonomy, translate the EU's climate and environmental objectives into criteria for specific economic activities for investment purposes. By nature, however, such classification approaches are too rigid and risk not considering the dynamic nature of the sector.

This study provides an alternative solution to classify Cleantech companies by developing a robust and fully replicable original methodology to identify European Cleantech companies in the Orbis database, a large company-level database commercialised by Bureau Van Dijk, through the use of a supervised machine learning (ML) algorithm applied to the extended business description of European companies. The resulting dataset will shed new light on the European Cleantech sector. Earlier studies, typically based on investment databases, provided only a partial perspective of the Cleantech phenomenon. By using a balance sheet database, such as the one provided by Bureau Van Dijck, we are able to broaden the scope of our analysis and include Cleantech companies that have never received an equity capital injection. Furthermore, by matching our sample of Cleantech companies to a variety of other databases, we are able to

[3] The term Cleantech will be used throughout the remainder of the paper.

[4] See Annex A for a brief history of the Cleantech concept.

provide a number of valuable insights into the functioning of the European Cleantech sector, related to geographic distribution, innovative capacity, size, VC investment activity, and others.[5]

In the remainder of this paper, we describe the methodology we employed to identify Cleantech companies in Europe and we provide preliminary descriptives about their distribution in terms of country, industry and foundation year, innovative performance (i.e. patents), accounting data and VC financing, to provide an initial overview of the Cleantech phenomenon in European countries using our novel database.

[5] See for example De Haan-Montes et al. (2023), Kraemer-Eis et al. (2022) or Pitchbook (2023).

2 | The identification of Cleantech companies

2.1 | Three steps methodology[6]

Currently, to the best of our knowledge, no comprehensive database provides a list of companies that fit the Cleantech definition illustrated above. Given the limitations of the existing classification methods (e.g., NACE industrial classification, EU Taxonomy), we developed a novel methodology to identify Cleantech companies, starting from the enterprise business description reported in the Orbis database (managed by Bureau Van Dijk).

The method first implies to manually identify the Cleantech nature of a relatively small set of companies. Then, this "manually-classified" dataset is used to let a machine learn the link between a company description and its Cleantech status. By letting the machine learn this mapping, it is possible to predict the classification of non-manually classified companies. We used a series of machine learning methods whose prediction error is particularly small for non-manually classified firms.

More specifically, the methodology to perform this task was based on three main steps:

1. Supervised machine learning (ML) algorithm applied to each company's extended business description retrieved from Orbis;
2. Computer-aided filter of false positive Cleantech instances applied to each company labelled as Cleantech in the previous step;
3. Manual checks, ecosystem segmentation, technological classification, and definition of the role of Cleantech companies (selected in step two) in the Cleantech ecosystem.

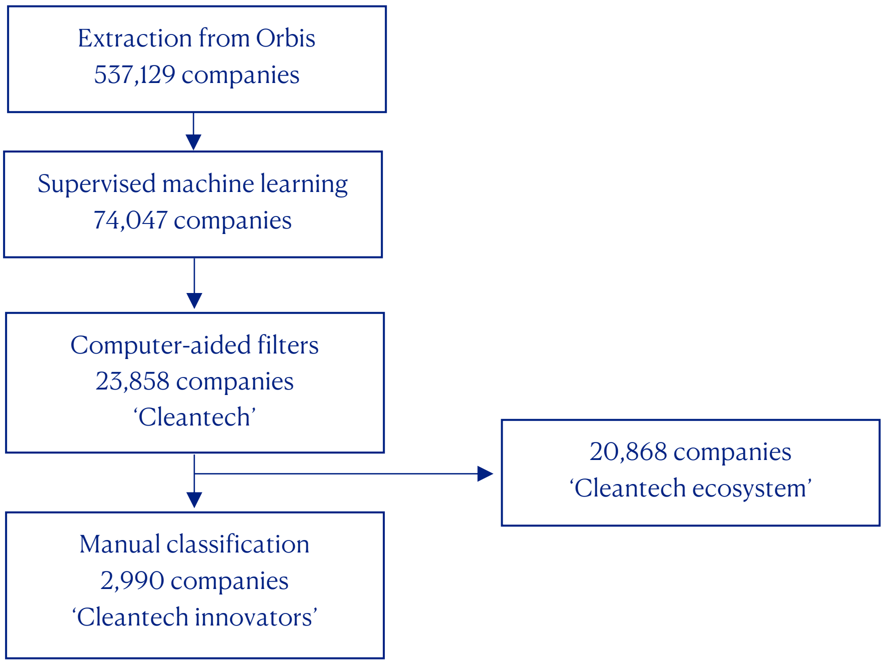
We applied the methodology to the entire sample of companies available in Orbis, a dataset handled by Bureau Van Dijk that includes financial information on over 40 million companies worldwide. Orbis has been selected because of company coverage, financial data availability and harmonisation potential. We selected all companies that met the following criteria:

1. Companies located in Europe;
2. Companies which had recorded accounting data for at least one business year;
3. Companies with an available extended business description.

[6] A detailed discussion of the methodology is provided in Annex B.

After applying the above criteria, 537,129 companies remained. The sample was reduced to 74,047 companies after running the ML algorithm and to 23,858 companies after applying computer-aided filters. Overall, out of 23,858 identified Cleantech companies, we manually classified 2,990 companies as "Cleantech innovators", focusing on clean technology development, and the remaining 20,868 as "Cleantech ecosystem", referring to companies that adopt Cleantech technologies, sell services based on Cleantech technologies, or provide inputs for the development of Cleantech technologies. Figure 1 provides the graphical representation of the methodology used to screen Cleantech companies.

Figure 1: Graphical representation of the methodology used to screen Cleantech companies



We fitted a series of machine learning algorithms that we trained on our training dataset, i.e. the one manually created and optimally tuned via cross-validation. For this purpose, we first generated a series of features (or predictors) from companies' description texts using text-mining techniques and then regressed these features to predict whether a company is or is not Cleantech. Assessing predictive performance in an out-of-sample mode, we found that the Gradient Boosting Machine (GBM), an ML method aggregating many classification trees, was the most predictive with a prediction accuracy larger than 90%. This accuracy was robust also to the presence of imbalanced data as in our training dataset the number of Cleantech firms is much smaller than the number of non-Cleantech ones. Using our optimal GBM fit, we finally predicted the Cleantech status completely out-of-sample over the larger dataset of 537,129 companies.

2.2 | Computer-aided filters

This phase aimed at reducing the number of false positives from the dataset of 74,047 companies obtained from the previous step, i.e. to include only Cleantech companies in the sample. We developed a set of computer-aided filters based on keywords using functions embedded in the Stata software. We developed an iterative process for elaborating these filters. First, we analysed the Cleantech literature to define appropriate keywords about clean technologies. Second, we went through company descriptions to validate the keywords identified in the previous step and to search for new ones. As with the ML step, we prepared the descriptions to reduce errors. Thus, all punctuation marks were removed and the text was reduced to lowercase. In addition, we chose the keywords by examining the word's root where necessary. The result of this phase was a new reduced sample, including 23,858 companies which were broadly defined as "Cleantech".

2.3 | Manual classification

While the first step allowed for accurate filtering of non-negative Cleantech companies from the starting population and the second step identified Cleantech companies (i.e., only Cleantech companies were included in the final sample), the third and final step had a twofold objective: i) to assign each Cleantech company to one or more technological categories; ii) to distinguish each Cleantech company into "Cleantech ecosystem" (i.e., referring to companies that adopt Cleantech technologies, sell services based on Cleantech technologies, or provide inputs for the development of Cleantech technologies) and "Cleantech innovators" (i.e., referring to companies committed to develop clean technologies). To this aim, two research assistants manually analysed the companies' business descriptions to ensure the maximum accuracy.

2.3.1 | Classification of Cleantech companies into technological categories

Each Cleantech company was classified into seven different technological categories reflecting the pillars of the European Green Deal and the EU Taxonomy. Some of these categories were further divided into sub-categories. We assigned each company to a technological category inspired by Haščič & Migotto (2015), in which technologies are classified according to their contribution to environmental sustainability. It should be noted that a company may be involved in one or more sub-categories. The structure of the seven technological categories (and sub-categories) is presented below:

1) Environmental management
    1. *Air/water/soil pollution abatement/remediation*
    2. *Waste management*
2) Resources preservation
    1. *Water conservation/availability*
    2. *Sustainable agri-food technologies*
    3. *Sustainable raw materials*

3) Industrial energy management
    1. *Sustainable energy production*
    2. *Sustainable fuels*
    3. *Energy-efficient industrial technologies*
4) Capture, storage, sequestration or disposal of GHG
5) Sustainable modes of transport
6) Sustainable buildings
7) Other categories

Environmental management

The first category, *Environmental management,* includes technologies dealing with pollution abatement, remediation, and waste management. In particular, the sub-category 1.1 *Air/water/soil pollution abatement/remediation* includes those technologies related to the abatement and remediation of air, water, and soil pollution, such as soil or water purification treatments. The subcategory 1.2 *Waste management* includes companies that produce technologies for solid waste collection, recovery, recycling and reusing raw materials, creation of fertilisers from waste, technologies used for incineration, and energy recovery.

Resources preservation

The second category, *Resource conservation,* includes technologies that can contribute to the preservation of ecosystems. This class includes sub-category 2.1 *Water conservation/availability,* referring to technologies that deal with water conservation. Examples include water-saving technologies such as taps that control water flow, valves that close automatically after actuation, or valves that close after releasing a predetermined amount of water. The second sub-category 2.2 *Sustainable agri-food technologies* concerns sustainable agri-food technologies such as hydroponic solutions or precision farming technologies. The third sub-category, 2.3 *Sustainable raw materials*, relates to the development of sustainable raw materials such as bioplastics or biodegradable raw materials obtained from natural materials such as sugar canes or potatoes (Ezgi Bezirhan Arikan & Havva Duygu Ozsoy, 2015).

Industrial energy management

The third category, *Industrial energy management,* includes technologies for energy production and energy efficiency. Specifically, the sub-category 3.1 *Sustainable energy production* includes clean energy generation technologies such as wind, solar thermal, photovoltaic, geothermal, marine, and hydroelectric. Other types of power generation considered were new nuclear technologies, fuel cells and co-generation technologies. The sub-category 3.2 *Sustainable fuels* includes fuels from renewable sources that minimise the environmental impact, e.g. fuels deriving from renewable biomass or from waste. Finally, the sub-category 3.3 *Energy-efficient industrial technologies* includes battery storage, capacitor and thermal storage. This class also includes technologies related to superconductors, pressurised fluid, mechanical & pumped, recyclable products, and reduction of materials in manufacturing.

Capture, storage, sequestration, or disposal of GHG

The fourth category, *Capture, storage, sequestration or disposal of GHG,* includes technologies that deal with the capture of GHG (e.g. carbon dioxide) from the atmosphere and their treatment.

Sustainable modes of transport

The fifth category, *Sustainable modes of transport,* includes technologies that can decarbonise the transportation sectors, such as technologies for electric vehicles, fuel cell vehicles or co_2 saving ducts.

Sustainable buildings

The sixth category, *Sustainable buildings,* includes technologies for energy efficiency management in buildings, both from an electricity and a thermal point of view. This category includes, among other things, technologies used to insulate buildings, such as expanded polystyrene.

2.3.2 | Classification of Cleantech companies into ecosystem segments

The second part of the manual classification acknowledges the complexity of the supply chain structure of the Cleantech ecosystem and accordingly segments the 23,858 Cleantech companies into:

1. **"Cleantech innovators":** these companies create (and eventually use) the clean technology as their core business. They are at the centre of the supply chain.
2. **"Cleantech ecosystem"**: these companies adopt clean technologies, sell services based on clean technologies, or provide inputs for the development of clean technologies. We further distinguished such companies into "experimenters" and "manufacturers", which support the realisation of the technology and "distributors", "integrators", and "operators", that make the technology available in the market.
   I. *Experimenters:* companies involved in performing experimental tasks that can lead to discoveries and advances in the science of the Cleantech supply chain (both private and public);[7]

[7] It is possible that some experimenters also carry out manufacturing activities.

II. *Manufacturers:* companies involved in the Cleantech supply chain, dealing with ancillary services concerning actual innovation; in other words, they deal with manufacturing, fabrication, and production of necessary and auxiliary components or raw materials to the clean technology;[8]

III. *Distributors:* companies that only distribute or are involved in the commercial provision of certain Cleantech products or technologies. Their primary role is to make clean technologies available on the market;[9]

IV. *Integrators:* companies involved in the Cleantech supply chain, dealing with accessory services concerning actual innovation; in other words, they deal with engineering, installation, procurement, design, conception, and planning. Their prominent role is to make the clean technology ready to use for the adopters;[10]

V. *Operators:* companies involved in the Cleantech supply chain that deals with the construction, implementation, and maintenance of facilities where clean technology is used; in other words, they are ancillary services for actual innovation. In addition, adopters who use technology as the primary tool to achieve their output (e.g. energy production) are also considered operators.[11]

Although each company could fit into more than one definition, we decided to assign a unique class according to the company's primary activity.

2.4 | Limitations

Some limitations of the applied methodology need to be taken into consideration. First, it is well known that AI systems are not entirely objective as they learn to make decisions based on data, which can include biased human decisions. For instance, during the creation of the training set, the human involvement in the provisioning and selection of data, the way of reasoning and the understanding of what constitutes Cleantech applied by the researchers involved in the process can make the model's predictions susceptible to bias. In order to alleviate this type of bias, we have adopted one of the techniques generally suggested by the literature: the involvement of more

[8] They may be manufacturers of components exclusively for clean technologies, but in most cases, their components can be used across several sectors, including the Cleantech sector.
[9] Since commercial activities are also carried out by other actors, starting with the innovators themselves, this class includes those companies that only engage in commercial activities.
[10] Since integration activities are also performed by other actors, starting with the innovators themselves, this class includes companies that exclusively perform integration activities using clean technologies provided by other companies.
[11] Since management activities are also performed by other actors, starting with the "Cleantech innovators" themselves, this category includes companies that exclusively perform management activities using clean technologies provided by other companies. Some operators can also be considered integrators, but not vice versa.

researchers to compare perspectives and the application of a rigorous and systematic approach to address inconsistencies. However, this remains a point of attention, not only for this study, but for AI methods in general.

A second bias can also be introduced by the underlying data itself. In our specific case, the outcome of our algorithm is based on the details provided by the companies in their business description. Accordingly, if a company is able to describe its business with an appropriate use of Cleantech-related terms, then the company is correctly depicted as Cleantech. Otherwise, if it is not sufficiently effective in communicating its business model, the terms related to Cleantech might not be mentioned, even if the company is Cleantech (false negative). Moreover, if a company falsely claims to run a Cleantech business model, it will be classified as such, effectively constituting a false positive. This type of problem, however, cannot be addressed for the case at hand, as it depends on a different human bias, the one of the company team which provides description of the business and this is not under the control of the researchers applying the ML technique.

Finally, as a procedural choice, we have decided to focus our attention toward corporates as isolated entitities. However, the Cleantech industry is more complex, with several interconnections among players. Consider, for example, the relationships that exist along the supply chains where players interact for the creation and delivery of products and services from raw materials to the end consumer. Similarly, there are Cleantech clusters that bring together different entities, such as businesses, research institutions, and other stakeholders to develop and commercialise green technologies. These clusters fall beyond the scope of this analysis, even if they are typically located in areas with a strong concentration of green technology companies. Failing to properly identify and classify them is a missing part of the exercise which deserves more attention in the future. Indeed, given the important role played in fostering collaboration, innovation, and knowledge sharing among the different stakeholders involved in the development and deployment of clean technologies, Cleantech clusters can help to accelerate the development and adoption of clean technologies, create new business opportunities, and drive economic growth while addressing environmental challenges.

3 | The mapping of Cleantech in Europe

3.1 | Descriptive statistics on Cleantech companies

3.1.1 | By ecosystem segment and technological category

Our final sample comprises 23,858 companies broadly identified as Cleantech, 2,990 of which are Cleantech *innovators* (12.5%). Table 1 reports the distribution of sample companies according to the segmentation described in section 2.3.2 |. Within the Cleantech *ecosystem* group, companies are equally distributed among *integrators* (27.5%), *operators* (23%) and *manufacturers* (22.6%). The two remaining groups, distributors and experimenters account of 14% and 0.4%, respectively.

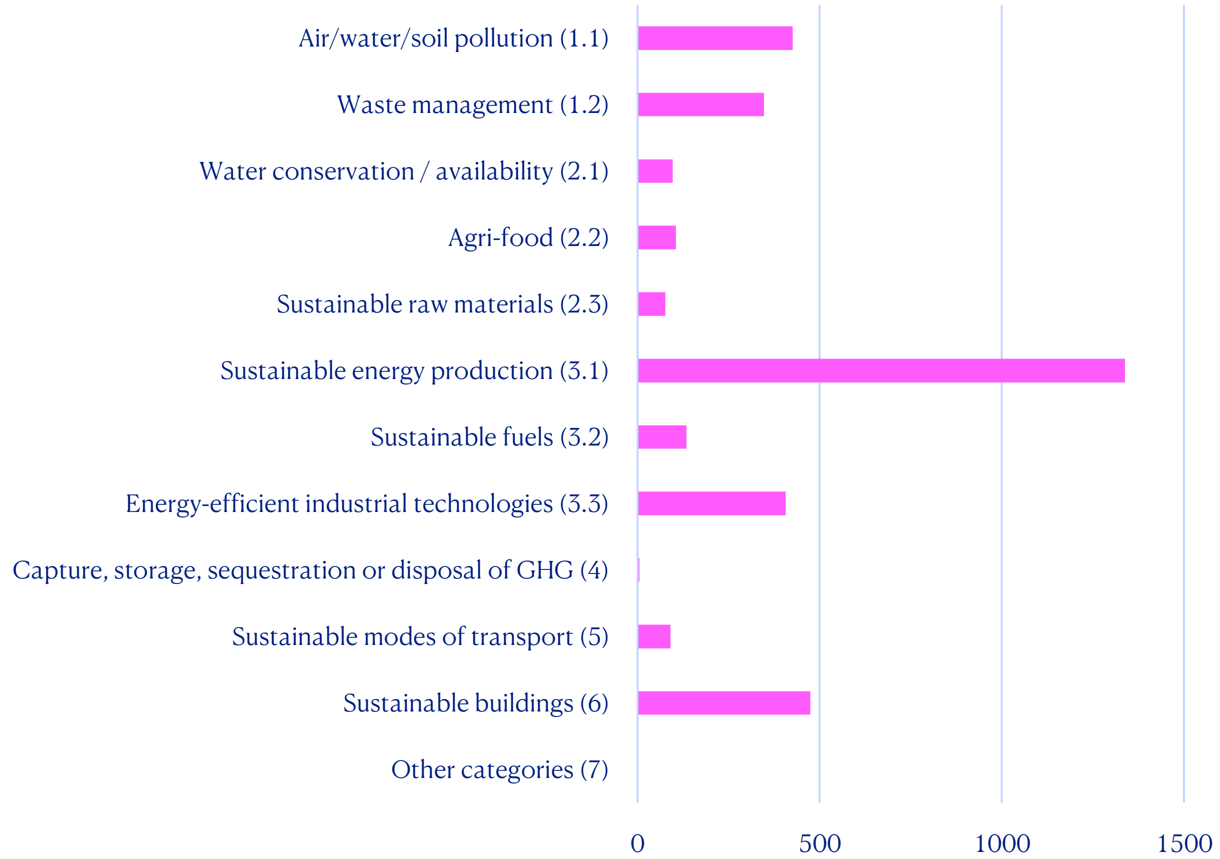
Table 1: Classification of Cleantech companies into different ecosystem segments

	# companies	*%*
Cleantech innovators	2,990	12.5%
Cleantech ecosystem	20,868	87.5%
Experimenters	103	0.4%
Manufacturers	5,380	22.6%
Distributors	3,337	14.0%
Integrators	6,558	27.5%
Operators	5,490	23.0%
Total	23,858	100%

Source: Orbis, authors' calculations

The distribution of Cleantech companies in the sample according to the different segments and technological categories described in Section 2.3 |, is reported in Figures 2 and Figure 3.

Figure 2: Cleantech innovators by technological categories

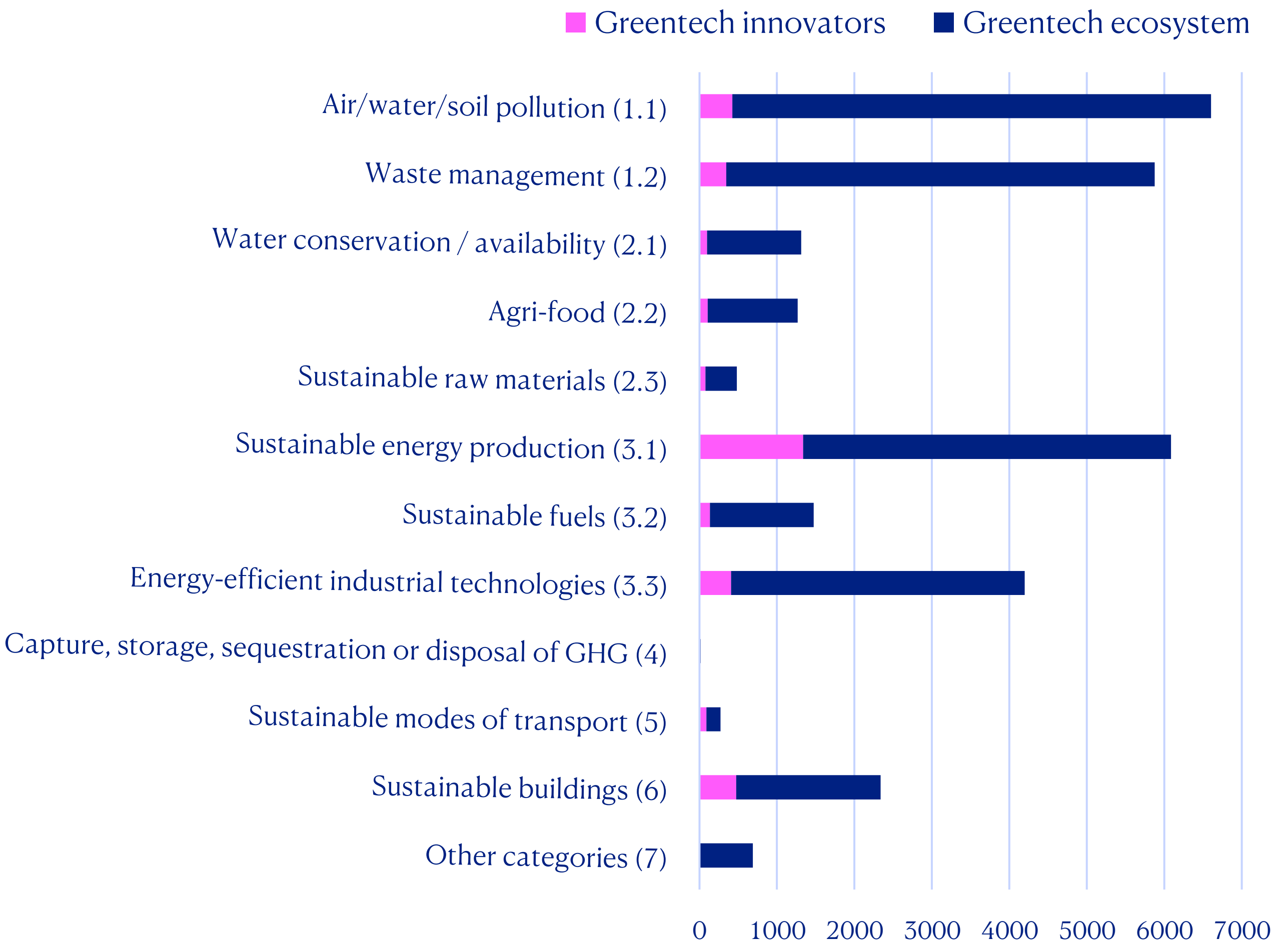


Source: Orbis, authors' calculations

The decarbonization of energy production is arguably the most significant environmental challenge of our time, reflected by the fact that 45% of companies fall under the *Sustainable Energy Production* sub-category (3.1) in the Cleantech innovators group. Among the other most prevalent sub-categories are end-of-pipe technologies for environmental management, including *air/water/soil pollution abatement/remediation* (1.1) at 14% and *waste management* (1.2) at 12%. Additionally, *energy efficiency in industrial processes* (3.3) represents 14% of companies, while *sustainable buildings* (6) make up 16%. The remaining categories, such as *water conservation/availability* (2.1), *sustainable agri-food technologies* (2.2), *sustainable raw materials* (2.3), *sustainable fuels* (3.2), and *sustainable modes of transport* (5), each account for approximately less than 5%. Lastly, the *Capture, Storage, Sequestration, or Disposal of GHG category* (4) is the least represented, constituting less than 1% of the group.

Figure 3 reports the distribution across technological categories for the broader sample of Cleantech companies (innovators and eco-system combined). When considering the broader sample, sub-category 3.1 (*Sustainable energy production)* remains one of the most represented technologies (26%). Compared to the group of innovators, however, the two sub-categories related to *Environmental management* are significantly better represented, accounting for 28% (1.1, *air/water/soil pollution*) and 25% (1.2, *waste management*), respectively.

Figure 3: Cleantech innovators + ecosystem by technological categories



Source: Orbis, authors' calculations

3.1.2 | By geography

Examining the spatial distribution of Cleantech companies provides valuable insights for policymakers, investors, and Cleantech firms themselves. It helps in identifying clusters, understanding regional advantages, evaluating policy effectiveness, identifying market gaps, and informing strategic planning and resource allocation decisions.

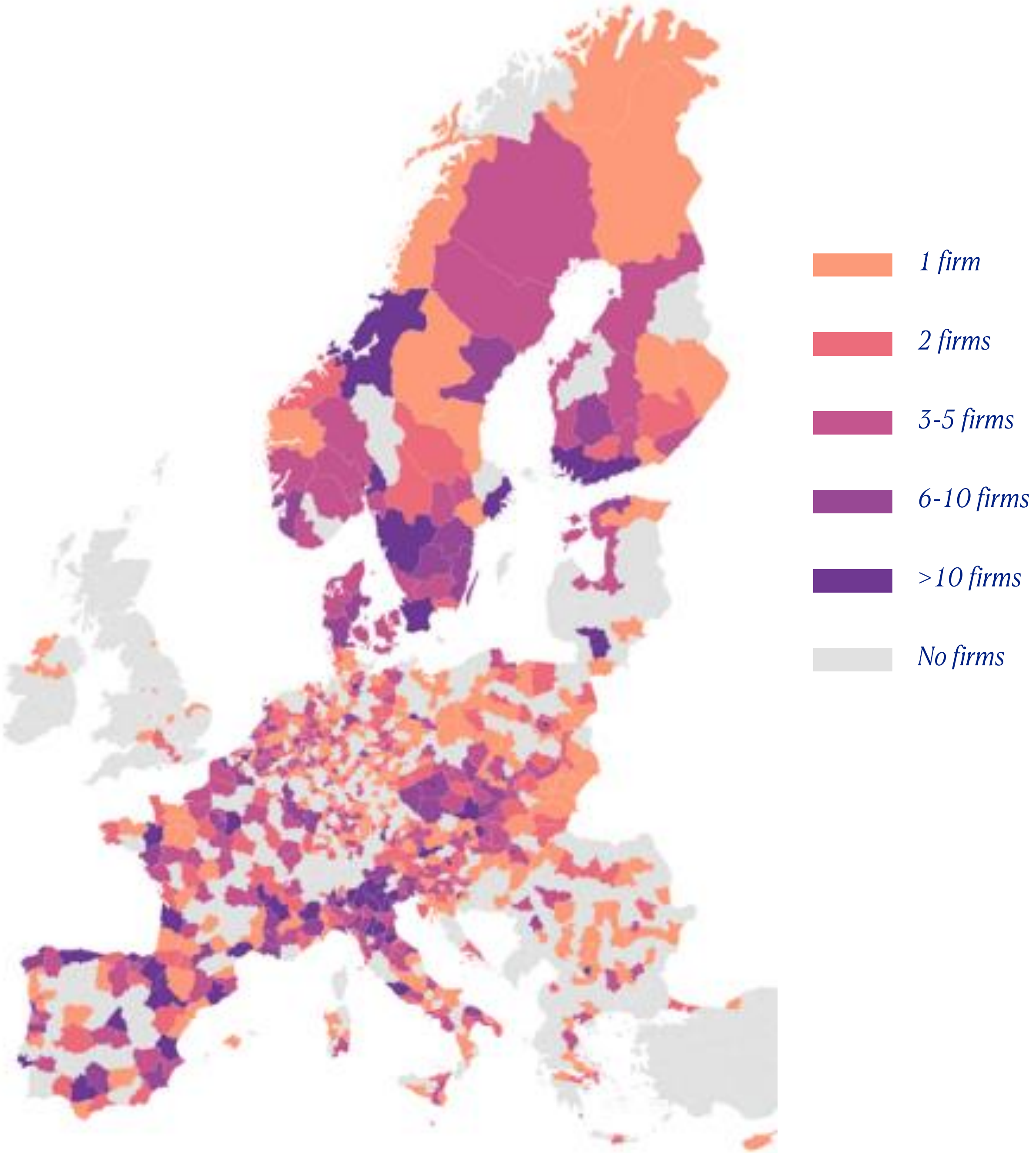
More than half of Cleantech companies (51.18%) are located in just three countries: Germany (18.65%), Italy (17.85%), and France (14.33%), with the remaining companies distributed over the other European countries according to the data provided in Table 2. No significant differences appear in the geographical distribution of Cleantech innovators and Cleantech ecosystem companies. Figure 4 and Figure 5 illustrate the regional distribution of Cleantech firms by NUTS-3 area, respectively for Cleantech innovators and Cleantech ecosystem. The regional distribution maps reveal a dense concentration of Cleantech companies in Europe's traditional VC hubs (see Kraemer-Eis et al., 2016).

Table 2: Distribution of Cleantech companies by country

	Cleantech companies		Cleantech innovators		Cleantech ecosystem	
	# companies	%	# companies	%	# companies	%
Germany	4,444	18.7%	515	17.3%	3,929	18.9%
Italy	4,254	17.9%	559	18.7%	3,695	17.7%
France	3,414	14.3%	371	12.4%	3,043	14.6%
Spain	2,072	8.7%	329	11%	1,743	8.4%
Poland	1,443	6.1%	152	5.1%	1,291	6.2%
Sweden	845	3.6%	141	4.7%	704	3.4%
Czech Republic	743	3.1%	99	3.3%	644	3.1%
Belgium	706	2.9%	101	3.4%	605	2.9%
Norway	677	2.8%	79	2.7%	598	2.9%
Austria	598	2.5%	85	2.9%	513	2.5%
Romania	550	2.3%	47	1.6%	503	2.4%
Finland	500	2.1%	71	2.4%	429	2.1%
Portugal	456	1.9%	47	1.6%	409	2%
Hungary	413	1.7%	30	1%	383	1.8%
Netherlands	387	1.6%	66	2.2%	321	1.5%
Denmark	334	1.4%	51	1.7%	283	1.4%
Bulgaria	312	1.3%	27	0.9%	285	1.4%
Slovakia	267	1.1%	29	1%	238	1.1%
Serbia	239	1%	18	0.6%	221	1.1%
Greece	226	0.9%	41	1.4%	185	0.9%
Croatia	192	0.8%	24	0.8%	168	0.8%
Lithuania	153	0.6%	18	0.6%	135	0.7%
Slovenia	151	0.6%	24	0.8%	127	0.6%
Latvia	112	0.5%	5	0.2%	107	0.5%
Estonia	83	0.4%	13	0.4%	70	0.3%
United Kingdom	70	0.3%	22	0.7%	48	0.2%
Luxembourg	45	0.2%	7	0.2%	38	0.2%
North Macedonia	43	0.2%	2	0.1%	41	0.2%
Switzerland	41	0.2%	3	0.1%	38	0.18%
Iceland	15	0.06%	1	0.03%	14	0.07%
Malta	13	0.05%	3	0.1%	10	0.05%
Turkey	13	0.05%	3	0.1%	10	0.05%
Montenegro	11	0.05%	0	0%	11	0.05%
Ireland	5	0.02%	1	0.03%	4	0.02%
Cyprus	1	0.00%	1	0.03%	0	0%
Total	23,828	100%	2,985	100%	20,843	100%

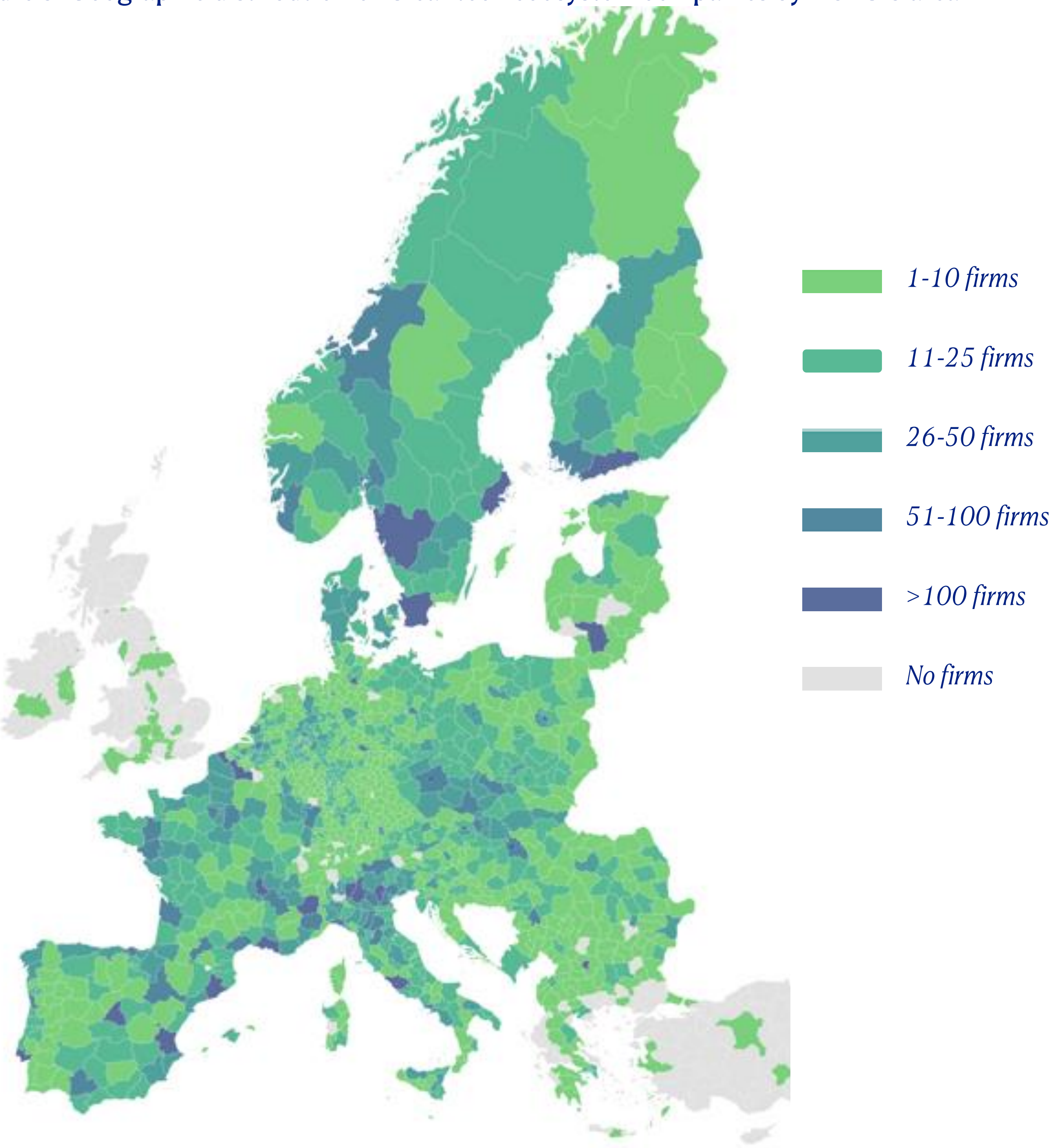
Source: Orbis, authors' calculations

Figure 4: Geographic distribution of Cleantech innovators by NUTS-3 area



Source: Orbis, authors' calculations

Figure 5: Geographic distribution of Cleantech ecosystem companies by NUTS-3 area



Source: Orbis, authors' calculations

To assess the prevalence of regional technological specialisation patterns, the Balassa-index was calculated for each NUTS3-region (Figure 6 for innovators and Figure 7 for ecosystem companies). The Balassa-index is an empirical proxy for regional specialisation and is calculated for each region, r_0, and technology, *j*, as follows:

$$BI_j^{r_0} = \frac{companies_j^r \,/\, \sum_j^J companies_j^r}{\sum_r^R companies_j^r \,/\, \sum_j^J \sum_r^R companies_j^r}$$

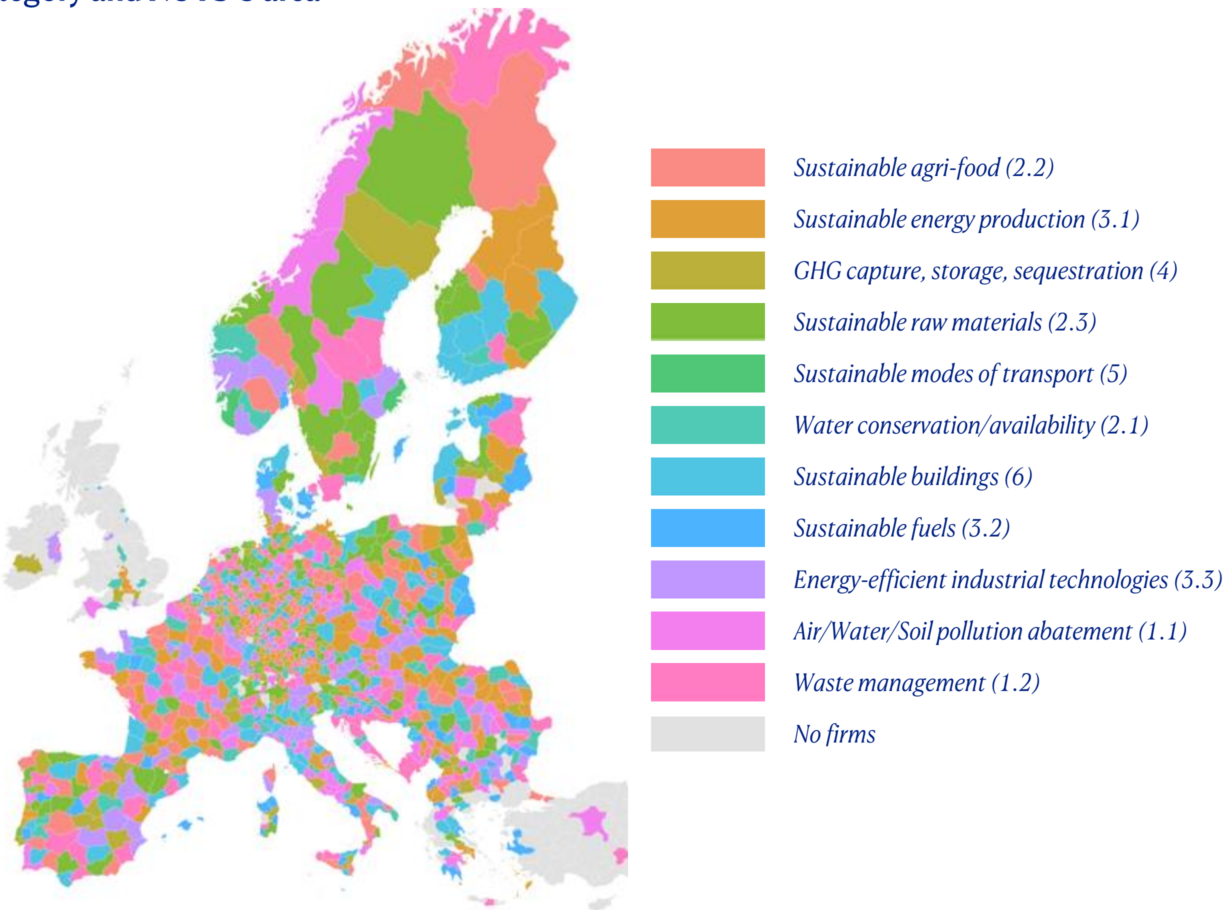
The index exceeds one if the extend of regional concentration for a specific technology exceeds the concentration of the sample average, and thereby is assumed to reveal a region's relative technological strength (Soete, 1987), and hence, indicate a comparative production advantage for a specific clean technology. In this context, Figure 6 and Figure 7 illustrate for each NUTS-3 region the technology category with the highest index value, and therefor, the technology category in which the region has a revealed comparative advantage.

Figure 6: Regional specialisation patterns of Cleantech innovators by technological category and NUTS-3 area

Note: The map depicts the technology category with the highest Bassala index value among all technology categories within a given region.

Source: Orbis, authors' calculations

Figure 7: Regional specialisation patterns of Cleantech ecosystem companies by technological category and NUTS-3 area*



* The map depicts the technology category with the highest Bassala-index among all technology categories within a given region.

Source: Orbis, authors' calculations

3.1.3 | By sector

This section illustrates the cross-sectoral nature of Cleantech, by describing the distribution of companies over the different NACE[12] rev.2 section. Cleantech companies are indeed active across a wide range of sectors and some interesting conclusions regarding sectoral concentration emerge.

When considering the entire sample (Table 3, columns 2 and 3), the majority of Cleantech companies (innovators + ecosystem) operate in the manufacturing (C), wholesale and retail trade (G), water supply and waste management (E), and construction (F) sectors. Focusing only on Cleantech innovators (Table 3, columns 4 and 5), there is a significantly stronger concentration in manufacturing (42.91%), which indicates that Cleantech innovation occurs predominantly in hardware-intensive sectors. Examining in closer detail the sectoral distribution of Cleantech innovators, by looking at 4-digit NACE codes, we observe some narrow sectors with a relatively

[12] The NACE classification (Statistical Classification of Economic Activities in the European Community) is a sectoral classification system used to categorise economic activities and industries within the EU.

dense concentration of Cleantech innovators. For example, 8.52% of Cleantech innovators operate in 35.11 NACE code (i.e. Production of electricity), while another 6.2% belong to the 71.12 NACE code (Engineering activities and related technical consultancy). Other NACE 4-digit code with a significant overrepresentation of Cleantech innovators are 46.69 (Wholesale of other machinery and equipment), 27.11 (Manufacture of electric motors, generators and transformers), 43.21 (Electrical installation) and 28.29 (Manufacturing of other general-purpose machinery n.e.c.), which all represent around 2.5% of the Cleantech innovators sample.

Table 3: Distribution of Cleantech companies by industry (NACE rev. 2)

NACE rev.2 section	Cleantech companies		Cleantech innovators		Cleantech ecosystem	
	# companies	*%*	*# companies*	*%*	*# companies*	*%*
A - Agriculture, forestry and fishing	171	0.7%	17	0.6%	154	0.7%
B - Mining and quarrying	171	0.7%	9	0.3%	162	0.8%
C - Manufacturing	5,686	23.9%	1,281	42.9%	4,405	21.1%
D - Electricity, gas, steam and air conditioning supply	1,917	8%	318	10.7%	1,599	7.7%
E - Water supply; sewerage, waste management and remediation activities	3,759	15.8%	140	4.7%	3,619	17.4%
F - Construction	3,376	14.2%	265	8.9%	3,111	14.9%
G - Wholesale and retail trade; repair of motor vehicles and motorcycles	4,792	20.1%	369	12.4%	4,423	21.2%
H - Transportation and storage	344	1.4%	18	0.6%	326	1.6%
I - Accommodation and food service activities	84	0.4%	9	0.3%	75	0.4%
J - Information and communication	206	0.9%	28	0.9%	178	0.9%
K - Financial and insurance activities	390	1.6%	75	2.5%	315	1.5%
L - Real estate activities	280	1.2%	29	1%	251	1.2%
M - Professional, scientific and technical activities	1,446	6%	359	12%	1,087	5.2%
N - Administrative and support service activities	968	4%	48	1.6%	920	4.4%
O - Public administration and defence; compulsory social security	21	0.1%	2	0.1%	19	0.1%
P - Education	28	0.1%	0	0%	28	0.1%
Q - Human health and social work activities	71	0.3%	5	0.2%	66	0.3%
R - Arts, entertainment and recreation	37	0.2%	3	0.1%	34	0.2%
S - Other service activities	81	0.3%	10	0.3%	71	0.3%
Total	23,828	100%	2,985	100%	20,843	100%

Source: Orbis, authors' calculations

Table 4. Distribution of Cleantech companies (innovators + ecosystem) by industry (NACE rev. 2) and technological categories

NACE rev.2 Section	Technological categories*													
	1		*2*		*3*		*4*		*5*		*6*		*7*	
A - Agriculture, forestry and fishing	73	0.6%	69	2.3%	73	0.6%	0	0.0%	0	0.0%	11	0.5%	7	1.0%
B - Mining and quarrying	84	0.7%	25	0.8%	67	0.6%	5	29.4%	0	0.0%	9	0.4%	12	1.7%
C - Manufacturing	2416	19.4%	712	23.3%	2696	23.0%	0	0.0%	129	47.4%	838	35.9%	150	21.8%
D - Electricity, gas, steam and air conditioning supply	295	2.4%	118	3.9%	1754	14.9%	0	0.0%	8	2.9%	29	1.2%	98	14.3%
E - Water supply; sewerage, waste management and remediation activities	3614	29%	418	13.7%	1708	14.5%	0	0.0%	1	0.4%	19	0.8%	8	1.2%
F - Construction	1281	10.3%	669	21.9%	1590	13.5%	1	5.9%	19	7.0%	455	19.5%	63	9.2%
G - Wholesale and retail trade; repair of motor vehicles and motorcycles	2680	21.5%	469	15.3%	1936	16.5%	1	5.9%	43	15.8%	722	30.9%	226	32.9%
H - Transportation and storage	199	1.6%	34	1.1%	145	1.2%	0	0.0%	11	4.0%	20	0.9%	23	3.3%
I - Accommodation and food service activities	39	0.3%	15	0.5%	34	0.3%	0	0.0%	1	0.4%	15	0.6%	1	0.1%
J - Information and communication	92	0.7%	18	0.6%	107	0.9%	0	0.0%	9	3.3%	27	1.2%	9	1.3%
K - Financial and insurance activities	156	1.2%	34	1.1%	233	2.0%	3	17.6%	7	2.6%	27	1.2%	17	2.5%
L - Real estate activities	141	1.1%	34	1.1%	148	1.3%	0	0.0%	2	0.7%	25	1.1%	11	1.6%
M - Professional, scientific and technical activities	601	4.8%	187	6.1%	888	7.6%	4	23.5%	34	12.5%	89	3.8%	42	6.1%
N - Administrative and support service activities	662	5.3%	212	6.9%	259	2.2%	2	11.8%	6	2.2%	27	1.2%	12	1.7%
O - Public administration and defence; compulsory social security	9	0.1%	1	0.0%	6	0.1%	1	5.9%	0	0.0%	3	0.1%	2	0.3%
P - Education	14	0.1%	5	0.2%	9	0.1%	0	0.0%	0	0.0%	6	0.3%	2	0.3%
Q - Human health and social work activities	40	0.3%	14	0.5%	29	0.2%	0	0.0%	1	0.4%	4	0.2%	2	0.3%
R - Arts, entertainment and recreation	26	0.2%	5	0.2%	12	0.1%	0	0.0%	0	0.0%	3	0.1%	0	0.0%
S - Other service activities	37	0.3%	17	0.6%	48	0.4%	0	0.0%	1	0.4%	5	0.2%	2	0.3%
	12,459		3,056		11,742		17		272		2,334		687	

*1. Environmental management, 2. Resources preservation, 3. Industrial energy management, 4. Capture, storage, sequestration or disposal of GHG, 5. Sustainable modes of transport, 6. Sustainable buildings, 7. Other categories

Source: Orbis, authors' calculations

Table 4 illustrates the sectoral distribution (according to the NACE rev.2 classification) of the entire sample of Cleantech companies for the technological categories defined in section 2.3.1 |. Companies active in our largest category, *Environmental management* (category 1) are predominantly active in the NACE sector water supply and waste management (29%), wholesale (21.5%) and manufacturing (19.4%). Concerning Cleantech companies active in *Industrial energy management* (category 3), the second largest technological category, the NACE sector manufacturing accounts for over 1 in 5 companies (22.96%), while electricity, gas, steam and air conditioning supply, water supply and waste management, construction and wholesale each account for around 15% of *industrial energy management* companies. Finally, companies active in *Resources preservation* (category 2), the third largest technological category, operate mostly in manufacturing (23.3%), construction (21.9%), wholesale (15.3%) and water supply and waste management (13.7%).

3.1.4 | By year of incorporation

While interest in Cleantech has experienced a strong surge in recent years, driven , among other things, by evolutions on the regulatory front, environmental technology is not a new phenomenon and predates the current green regulatory wave. This is confirmed by our data, as illustrated in Table 4, which illustrates the distribution of Cleantech companies by year of incorporation and shows that over half (63.77%) of the companies were founded in the previous decennium, prior to the first Cleantech investment wave.

Noteworthy is the gradual decline in the number of Cleantech companies after 2010. One potential explanation for initial downward trend is the global economic downturn that occurred during this period, commonly referred to as the "cleantech crash" when the aftermath of the Great Financial Crisis led to reduced investment and funding opportunities for innovative companies. These challenges were further aggravated by the sovereign debt crisis, as public budgetary constraints led to reduced incentives and subsidies for green technology, creating a less favorable environment for Cleantech to flourish. Similarly, the table shows a drop from 2016 onwords explained by the methodological approach adopted for constructing the initial sampling based on censoring of the most recent companies. This choice was driven by the need to track the evolution of these companies over time for a significant number of years.

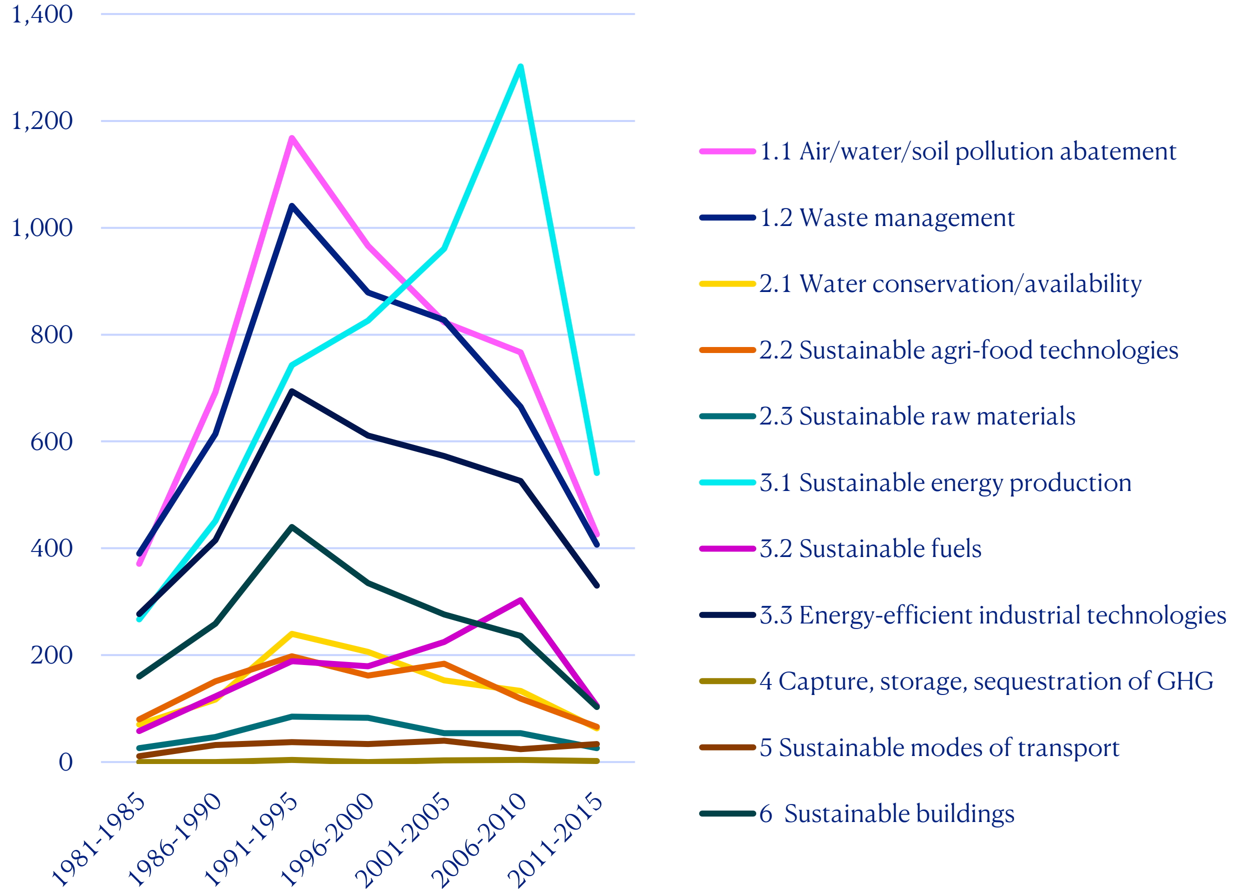
Table 4: Distribution of Cleantech companies by year of incorporation

Year of incorporation	Cleantech companies		Cleantech innovators		Cleantech ecosystem	
	# companies	*%*	*# companies*	*%*	*# companies*	*%*
Before 1980	4,326	18.1%	529	17.7%	3,797	18.2%
1981-1985	1,348	5.7%	143	4.8%	1,205	5.8%
1986-1990	2,308	9.7%	257	8.6%	2,051	9.8%
1991-1995	3,803	16%	350	11.7%	3,453	16.6%
1996-2000	3,409	14.3%	416	13.9%	2,993	14.4%
2001-2005	3,277	13.8%	433	14.5%	2,844	13.6%
2006-2010	3,320	13.9%	560	18.8%	2,760	13.2%
2011-2015	1,669	7%	250	8.4%	1,419	6.8%
2016 onwards	368	1.5%	47	1.6%	321	1.5%
Total	23,828	100%	2,985	100%	20,843	100%

Source: Orbis, authors' calculations

By plotting the evolution of Cleantech start-ups by technological category, Figure 8 reveals diverging trends in the emergence (and decline) of specific green technologies. Interestingly, the declining trend in the number of Cleantech firm births initiated already prior to 2010 in all but two categories: *sustainable energy production* (3.1) and *sustainable fuels* (3.2), as both sectors flourished beyond 2010, although their exponential growth was eventually countered by the financial crisis as well.

Figure 8: Evolution of Cleantech start-ups by technological category (1981-2015)



Source: Orbis, authors' calculations

3.2 | Patent data

Patents play a crucial role in analysing the innovative potential of Cleantech companies, providing valuable insights into their technological advancements and competitive advantage. Examining patterns in Cleantech patenting activity allows to assess their capacity to develop novel solutions, thereby addressing environmental challenges and driving sustainable practices.

Starting from the identified sample of Cleantech companies, we matched firm-level data with patent data in the Orbis Intellectual Property (Orbis IP) database by using the Bureau van Dijk company identifiers. We selected all patent applications filed by our sample companies at the European Patent Office (EPO) to ensure that the data is comparable across countries. Table 5 illustrates the distribution of EPO patenting Cleantech companies for Cleantech innovators and ecosystem companies. Just over 2,700 companies (11.3% of the sample) filed at least one EPO patent. Cleantech innovators appear to patent inventions at a significantly higher rate than the ecosystem companies.

To retrieve additional information on patenting activity (e.g., application dates, technological codes, etc.), the publication number of the patents collected from Orbis IP are matched to those in the Worldwide Patent Statistical Database (PATSTAT) of the EPO.[13] In particular, we identified those patents that report a Cooperative Patent Classification (CPC) code equal to Y02, indicating the technologies or applications for mitigation or adaptation against climate change (Climate Change Mitigation Technologies, CCMT). Among patenting Cleantech companies (2,705), 43.1% have at least one patent in a CCMT-related field. Distinguishing between the segments of Innovators and ecosystem companies, this share becomes 61.9% and 35.3%, respectively.

Table 5: EPO patenting activity of Cleantech companies

	Cleantech companies		Cleantech innovators		Cleantech ecosystem	
	# companies	*%*	*# companies*	*%*	*# companies*	*%*
At least one in any field	2,705	11.3%*	792	26.5%*	1,913	9.2%*
At least one in a CCMT field	1,166	43.1%**	490	61.9%**	676	35.3%**

*Of all Cleantech companies
**Of patenting Cleantech companies

Source: Orbis, authors' calculations

The patenting intensity of specific technology categories typically depends on a number of factors, such as the technology's novelty, its market potential, the intellectual property landscape it operates in, regulatory considerations, or strategic motives. Novel and inventive technologies with commercial potential are more likely to be eligible for patents.

Table 6 and Table 7 report the distribution of EPO patenting firms by technological category for the two sub-samples of companies belonging to Cleantech innovators and the Cleantech ecosystem, respectively. The technological category with the largest share of patenting companies is that of *Sustainable raw materials* (2.3) in the former sample and that of *Sustainable modes of transport* (5) in the latter one. We also computed the share of patenting firms that filed at least one patent application in a CCMT field, for each technological category. After excluding the smallest domains in terms of the number of firms, the technological category with the largest proportion of companies with CCMT patents is that of *Sustainable fuels* (3.2) in the sub-sample of Cleantech innovators and *Sustainable modes of transport* (5) in the sub-sample of Cleantech ecosystem.

[13] We used the Autumn Edition of 2021.

Table 6: Distribution of EPO patenting Cleantech companies by technological category in the segment of Cleantech innovators*

Technological category	At least one EPO patent in any field		At least one EPO patent in a CCMT field	
	# companies	*%***	*# companies*	*%****
Air/water/soil pollution abatement/remediation (1.1)	138	32.4%	73	52.9%
Waste management (1.2)	102	29.4%	59	57.8%
Water conservation/availability (2.1)	24	25%	5	20.8%
Sustainable agri-food technologies (2.2)	25	23.8%	9	36%
Sustainable raw materials (2.3)	34	44.7%	18	52.9%
Sustainable energy production (3.1)	284	21.2%	221	77.8%
Sustainable fuels (3.2)	27	20.2%	23	85.2%
Energy-efficient industrial technologies (3.3)	145	35.6%	87	60%
Capture, storage, sequestration or disposal of GHG (4)	2	40%	2	100%
Sustainable modes of transport (5)	36	40%	24	66.7%
Sustainable buildings (6)	128	27%	61	47.7%
Others	0		0	
Total	792	26.5%	490	61.9%

*Each company can be associated to multiple technological categories; hence, the totals are not the sum of the row values.
**Of all Cleantech companies
***Of patenting Cleantech companies

Source: Orbis, authors' calculations

Table 7: Distribution of EPO patenting Cleantech companies by technological category in the segment of Cleantech ecosystem*

Technological category	At least one EPO patent in any field		At least one EPO patent in a CCMT field	
	# companies	*%***	*# companies*	*%****
Air/water/soil pollution abatement/remediation (1.1)	591	9.6%	191	32.3%
Waste management (1.2)	369	6.7%	115	31.2%
Water conservation/availability (2.1)	164	13.5%	69	42.1%
Sustainable agri-food technologies (2.2)	68	5.9%	19	27.9%
Sustainable raw materials (2.3)	72	17.7%	24	33.3%
Sustainable energy production (3.1)	462	9.7%	206	44.6%
Sustainable fuels (3.2)	155	11.6%	71	45.8%
Energy-efficient industrial technologies (3.3)	327	8.6%	123	37.6%
Capture, storage, sequestration or disposal of GHG (4)	1	8.3%	1	100%
Sustainable modes of transport (5)	38	20.9%	19	50%
Sustainable buildings (6)	186	10%	67	36%
Others	59	8.6%	18	30.5%
Total	1,913	9.2%	676	35.3%

* each company can be associated to multiple technological categories; hence, the totals are not the sum of the row values.
**Of all Cleantech companies.
***Of patenting Cleantech companies.

Source: Orbis, authors' calculations

The innovative capacity of a local Cleantech ecosystem differs between countries. According to Table 8, Austria's Cleantech companies engage most intensely in patented innovation, with 22.2% of Cleantech companies owning at least one patented innovation. Sweden follows closely behind with 19.4%, while Germany ranks third with 17.7%. When specifically considering EPO patents in the CCMT fields, Austria also emerges as the leading country, with 50.4% of the patenting companies having at least one EPO patent in a CCMT field. Spain takes the second position with 48.4%, closely followed by Germany with 48.2%. The data highlights the strong performance of Austria in both overall EPO patenting by Cleantech companies and specifically in the CCMT fields. This suggests that Austria's Cleantech ecosystem is particularly innovation-intensive.

Table 8: Distribution of EPO patenting Cleantech companies by country

Country	At least one EPO patent in any field		At least one EPO patent in a CCMT field	
	# companies	*%**	*# companies*	*%***
Germany	786	17.7%	379	48.2%
Italy	558	13.1%	186	33.3%
France	259	7.6%	95	36.7%
Spain	182	8.8%	88	48.4%
Poland	59	4.1%	17	28.8%
Sweden	164	19.4%	73	44.5%
Czech Republic	34	4.6%	15	44.1%
Belgium	88	12.5%	31	35.2%
Norway	70	10.3%	31	44.3%
Austria	133	22.2%	67	50.4%
*Others****	365	7.9%	181	49.6%
Total	2,698	11.3%	1,163	43.1%

*Of all Cleantech companies in a given country.

**Of patenting Cleantech companies in a given country.

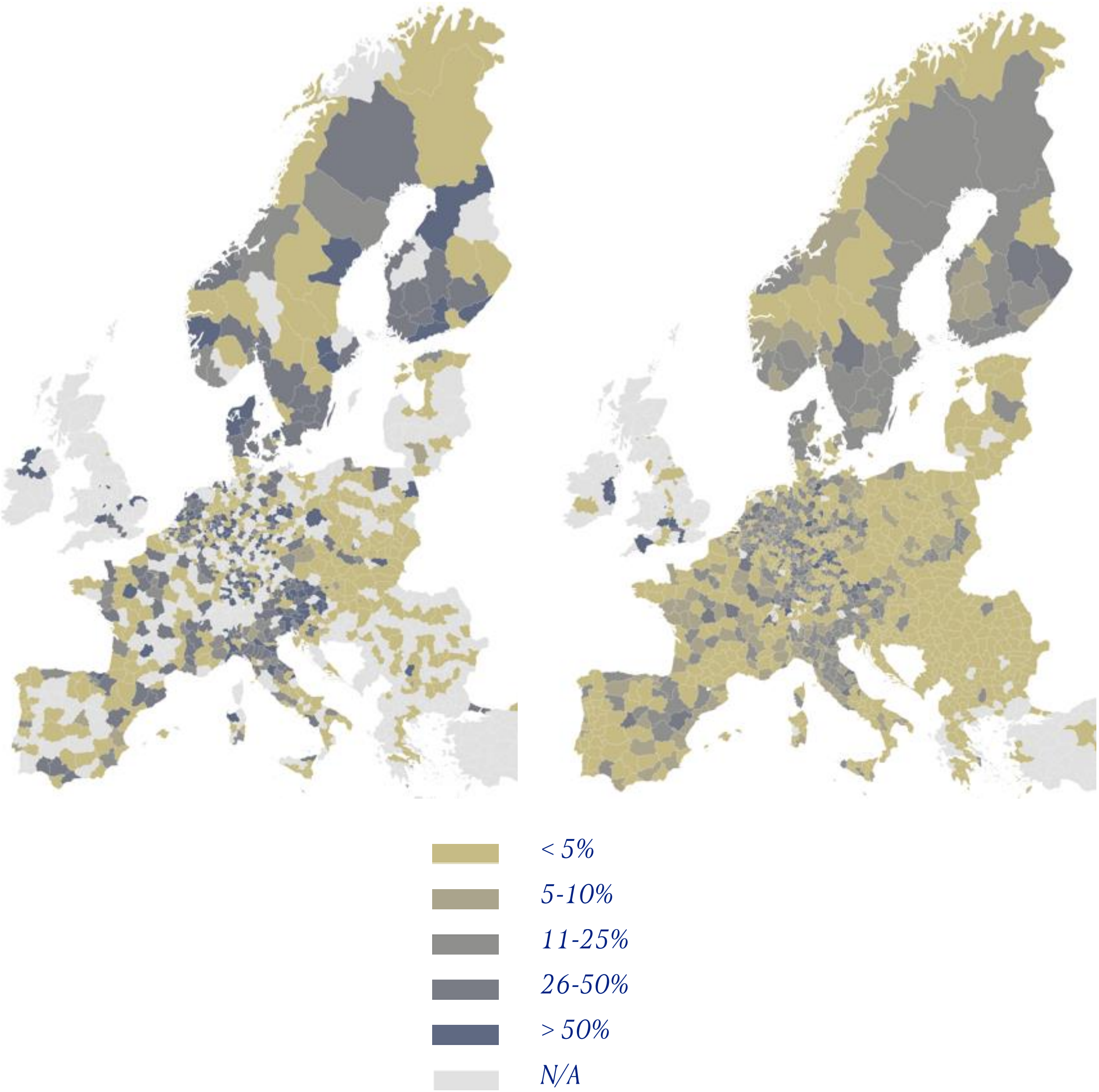
***The residual category includes Romania, Finland, Portugal, Hungary, Netherlands, Denmark, Bulgaria, Slovakia, Serbia, Greece, Croatia, Lithuania, Slovenia, Latvia, Estonia, United Kingdom, Luxembourg, North Macedonia, Switzerland, Iceland, Malta, Turkey, Montenegro, Ireland, and Cyprus; the information on the NUTS area is missing for 7 patenting companies.

Source: Orbis, authors' calculations

Innovative economic activities tend to cluster spatially due to knowledge spill-overs and access to skilled labour, among other things. Clustering facilitates the exchange of ideas, promotes efficiency and creates scale effects in innovative activities. Figures 9 illustrates, respectively, the geographic distribution of the share of patenting firms in the sub-samples of Cleantech innovators and Cleantech ecosystem by NUTS-3 area and indeed reveals clear clustering patterns in patenting intensity. The share of patenting tends to be higher in areas with more Cleantech activity, typically around Europe's traditional VC-hubs.

Figure 9: The share of patenting Cleantech companies by NUTS3 area

a) Cleantech innovators *b) Cleantech ecosystem*



Source: Orbis, authors' calculations

Finally, patenting intensity also differs depending on the ecosystem segments identified in section 2.3 | (Table 9). *Experimenters* engage most intensely in patenting activity (30.1%). They are followed by *Innovators* (26.49%) and *Manufacturers* (21.91%).

Table 9: Distribution of EPO patenting Cleantech companies by ecosystem segment

Category	At least one EPO patent in any field		At least one EPO patent in a CCMT field	
	# companies	%*	*# companies*	%**
Innovators	792	26.49%	490	61.87%
Experimenters	31	30.10%	11	35.48%
Manufacturers	1,179	21.91%	396	33.59%
Distributors	174	5.21%	56	32.18%
Integrators	307	4.68%	106	34.53%
Operators	222	4.04%	107	48.20%
Total	*2,705*	*11.34%*	*1,166*	*43.11%*

*Of all Cleantech companies.
**Of patenting Cleantech companies.

Source: Orbis, authors' calculations

3.3 | Financial KPIs

By collecting information on several essential financial KPIs from Orbis, we are able to construct a more detailed overview on the average scale of the Cleantech companies in our sample (Table 10). There were 239,450 data points relating to 23,828 Cleantech companies over the period 2009 until 2022, with a company being covered for an average time span of 10 years.

Table 10: Descriptive statistics for selected financial KPIs for the full sample of Cleantech companies

	Cleantech companies					
	# obs	*Mean*	*Median*	*St. dev.*	*Min*	*Max*
Sales (000'€)	239,450	149,137	9,707	2,474,602	0	291,000,000
Total assets (000'€)	184,867	196,288	8,664	3,442,078	0	283,000,000
Net profit (000'€)	231,074	5,419	195	145,218	-16,500,000	19,900,000
EBITDA (000'€)	218,641	18,541	695	372,836	-7,245,000	40,100,000
# Employees	190,340	413	48	3,980	0	291,000

Source: Orbis, authors' calculations

In Table 11, we report the mean values of the main accounting variables separately for Cleantech innovators and Cleantech ecosystem companies. We also performed a t-test comparison, as reported in the last columns of the table below, reporting differences and significance. Cleantech innovators typically operate at a larger scale, compared to their ecosystem counterparts.

Table 11: Mean values for selected financial KPIs: Innovators vs Ecosystem

	Cleantech innovators	Cleantech ecosystem	Difference	Significance-level*
Sales (000'€)	215,393	139,534	75,860	(***)
Total assets (000'€)	258,346	187,377	70,969	(***)
Net profit (000'€)	6,384	5,278	1,107	
EBITDA (000'€)	22,659	17,943	4,716	(**)
# Employees	688	372	317	(***)

* (**) and (***) indicate significantly different means at the 5%- and 1%-level respectively.

Source: Orbis, authors' calculations

3.4 | Venture capital investment in Cleantech companies

VC financing is a crucial source of external funding for Cleantech companies. Cleantech innovations often require significant upfront investments in research and development, prototype development, and scaling up production. Venture capitalists are able to provide the necessary funding to support such high-cost activities, which may be difficult to obtain from other traditional financing sources. In addition, Cleantech companies often face longer time horizons for commercialisation and profitability, due, for example, to regulatory complexities, market adoption challenges, and technological advancements. VC investors, with their risk-tolerant approach and longer investment horizons, are well positioned to provide patient capital and support the company's growth over the long term. In addition, they bring industry expertise, networks and business guidance, which can help Cleantech companies navigate the complex landscape, access markets, and achieve scalability.

Matching our data on Cleantech companies from the Orbis database with data on VC investment from VICO 4.0,[14] a pan-European dataset on VC investment activity developed by the RISIS2 EU-funded Horizon 2020 project (comprising more than 54,910 European VC investment deals), allows us to develop a better understanding of the role of VC in European Cleantech development. Table 13 illustrates the involvement of VC investors in the Cleantech companies identified in Orbis.

A total of 170 companies (0.71% of the 23,828 Cleantech companies) received at least one investment from VC investors. Unsurprisingly, VC involvement is substantially larger for Cleantech innovators (2.2%), compared to the ecosystem companies (0.5%). These relatively modest shares evidence that our identification methodology and resulting database is complementary to existing investment-based databases.

[14] The VICO dataset collects VC deals as contained in the commercial databases Zephyr, Crunchbase and Thomson Eikon. For more information, see https://rcf.risis2.eu/dataset/12/metadata.

Table 12: Distribution of VC-backed Cleantech companies by category

	Cleantech companies		Cleantech innovators		Cleantech ecosystem	
	# companies	*%*	*# companies*	*%*	*# companies*	*%*
# VC-backed companies	170	0.71%	66	2.21%	104	0.50%
Total # companies	23,828		2,985		20,843	

Source: Orbis, VICO4.0, authors' calculations

TablesTable 14 andTable 15 report the distribution of VC-backed companies by technological category for the two sub-samples of companies belonging to Cleantech innovators and the Cleantech ecosystem, respectively. In both samples, the category with the highest share of VC-backed companies is "*Sustainable energy production* (3.1)". Moreover, VC-involvement is particularly prevalent for Cleantech innovator active in the technological category *"Energy-efficient industrial technologies* (3.3)". These findings align with a recent EIF report (de Haan Montes et al., 2023) that analysed Cleantech VC deals from the Pitchbook database and found that energy-related Cleantech companies represent a substantial portion of Cleantech VC and PE growth investment activity across the EU.

Table 13: Distribution of VC-backed Cleantech innovator companies by technological category

Technological category	VC-backed companies	
	# companies	*%*
Air/water/soil pollution abatement/remediation (1.1)	5	7.6%
Waste management (1.2)	4	6.1%
Water conservation/availability (2.1)	0	0%
Sustainable agri-food technologies (2.2)	1	1.5%
Sustainable raw materials (2.3)	4	6.1%
Sustainable energy production (3.1)	36	54.6%
Sustainable fuels (3.2)	4	6.1%
Energy-efficient industrial technologies (3.3)	17	25.8%
Capture, storage, sequestration or disposal of GHG (4)	0	0%
Sustainable modes of transport (5)	0	0%
Sustainable buildings (6)	6	9.1%
Others	0	0%
Total	*66*	*2.2%*

*Each company can be associated to multiple technological categories; hence, the totals are not the sum of the row values.

Source: Orbis, VICO 4.0, authors' calculations

Table 14: Distribution of VC-backed Cleantech ecosystem companies by technological category

Technological category	VC-backed companies	
	# companies	*%*
Air/water/soil pollution abatement/remediation (1.1)	24	18.3%
Waste management (1.2)	24	19.2%
Water conservation/availability (2.1)	4	3.9%
Sustainable agri-food technologies (2.2)	2	1%
Sustainable raw materials (2.3)	11	6.7%
Sustainable energy production (3.1)	85	47.1%
Sustainable fuels (3.2)	14	9.6%
Energy-efficient industrial technologies (3.3)	36	18.3%
Capture, storage, sequestration or disposal of GHG (4)	0	0%
Sustainable modes of transport (5)	3	2.9%
Sustainable buildings (6)	13	6.7%
Others	4	0.6%
Total	104	0.5%

*Each company can be associated to multiple technological categories; hence, the totals are not the sum of the row values.

Source: Orbis, VICO 4.0, authors' calculations

Within the group of Cleantech ecosystem companies, two additional technological categories, namely "*Air/water/soil pollution abatement/remediation* (1.1)" and "*Waste management* (1.2)", exhibit a comparable level of VC financing as the "*Energy-efficient industrial technologies* (3.3)" sub-category. Approximately 18-19% of VC-backed companies in the Cleantech ecosystem belong to these categories, indicating that aside from supporting energy-related technology, VC investors also recognise the importance of solutions addressing environmental pollution and waste management challenges.

Patents are important for VC investors as they provide a means to protect the intellectual property rights of innovative technologies, thereby ensuring a competitive advantage and improved growth prospects for investee companies. Patents also enhance the attractiveness of companies to potential investors, by signalling the innovative potential of the patenting company. This holds true also for green technologies (Belucci et al., 2021). The data in Table 15 presents the distribution of EPO patenting Cleantech firms categorized by their VC-backed status. Among the VC-backed firms, 5.8% of the companies have filed at least one EPO patent in any field. Notably, 63.7% of the patenting companies that received VC investments have obtained at least one EPO patent specifically in the CCMT (Clean and Climate-friendly Mobile Technologies) fields.

Table 15: Distribution of EPO patenting Cleantech companies by VC-backed status

	At least one EPO patent in any field		At least one EPO patent in a CCMT field	
	# companies	*%*	*# companies*	*%*
VC-backed firms	88	5.8%	56	63.6%
Non-VC-backed firms	2,610	11%	1,107	42.4%
Total	*2,698*	*11.3%*	*1,163*	*43.1%*

*Each company can be associated to multiple technological categories; hence, the totals are not the sum of the row values.

Source: Orbis, VICO 4.0, authors' calculations

Table 16 reports the distribution of VC-backed Cleantech companies by country. Among the VC-backed Cleantech companies, the 23.5% is located in France, followed by Finland (18.8%), Germany (11.8%), and Sweden (8.8%). Considering only Cleantech innovators, it is worth noting that Finland holds the highest representation with 31.8% of VC-backed companies belonging to this group, followed by France (19.7%) and Estonia (15.6%). VC investment intensity, calculated as the share of Cleantech companies that received VC investments, differs significantly between countries, and ranges from less than a percentage point (Spain, 0.43%; Germany, 0.45%) to more than 10% (Estonia, 13.3%).

Table 16: Distribution of VC-backed Cleantech companies by country

Country	Cleantech companies			Cleantech innovators		Cleantech ecosystem	
	# companies	*% of total VC-backed*	*% of national Cleantechs**	*# companies*	*%*	*# companies*	*%*
France	40	23.5%	1.2%	13	19.7%	27	26%
Finland	32	18.8%	6.4%	21	31.8%	11	10.6%
Germany	20	11.8%	0.45%	1	1.5%	19	18.3%
Sweden	15	8.8%	1.8%	6	9.1%	9	8.7%
Estonia	11	6.5%	13.3%	10	15.6%	1	0.96%
Austria	9	5.3%	1.5%	7	10.6%	2	1.9%
Spain	9	5.3%	0.43%	1	1.5%	8	7.7%
Italy	5	2.9%		1	1.5%	4	3.9%
Netherlands	5	2.9%		1	1.5%	4	3.9%
United Kingdom	5	2.9%		1	1.5%	4	3.9%
Belgium	4	2.4%		1	1.5%	3	2.9%
Denmark	4	2.4%		0	0%	4	3.9%
Lithuania	2	1.2%		2	3%	0	0%
Czech Republic	1	0.6%		1	1.5%	0	0%
*Others***	8	4.7%		0	0%	8	7.7%
Total	*170*	*100%*		*66*	*100%*	*104*	*100%*

*The share of VC-backed companies in the number of national Cleantech companies is only reported for those countries with more than 5 VC-backed companies.
**The residual category includes Greece, Hungary, Poland, Portugal, Slovakia and Slovenia.

Source: Orbis, VICO 4.0, authors' calculations

Table 17 reports the distribution of VC-backed Cleantech companies by ecosystem segment. The findings reflect the fact that VC investor typically look for highly innovative companies with maximum growth potential. Hence, nearly 40% of VC-backed companies were classified as Innovators, followed by Manufacturers (28.2%) and Operators (12.9%).

Table 17: Distribution of VC-backed Cleantech companies by ecosystem segment

Category	Cleantech companies		Cleantech innovators		Cleantech ecosystem	
	# companies	*%*	*# companies*	*%*	*# companies*	*%*
Innovators	66	38.8%	66	100 %	0	0%
Experimenters	1	0.59%	0	0%	1	0.96%
Manufacturers	48	28.2%	0	0%	48	46.2%
Distributors	15	8.8%	0	0%	15	14.4%
Integrators	18	10.6%	0	0%	18	17.3%
Operators	22	12.9%	0	0%	22	21.2%
Total	170	100%	66	100%	104	100%

Source: Orbis, VICO 4.0, authors' calculations

4 | Conclusion

The European Green Deal is a comprehensive strategy aimed at achieving carbon-neutrality by 2050. It encompasses various interventions to drive sustainable development, with a particular emphasis on technological innovation. By focusing on decarbonising the energy sector, improving energy efficiency, promoting circular economy practices, and advancing sustainable transportation solutions, the Green Deal seeks to foster a "green" or "clean" economy. Innovations in the field of environmental sustainability, commonly referred to as Cleantech, are a key element of Europe's environmental and net-zero strategy. Unfortunately, today, no suitable framework, classification or method exists to identify and classify companies engaged in Cleantech innovation.

To accommodate this shortcoming, we developed a three-steps methodology based on: (i) a supervised ML algorithm applied to the extended business description of the full sample of European companies retrieved from Orbis dataset (ii) computer-aided filter of false positive Cleantech instances applied to each company labelled as (iii) a manual classification to identify companies that are committed to develop clean technologies, which led to the identification of 23,858 Cleantech companies, 2,990 of which were classified as "Cleantech innovators", while the remaining 20,868 companies were designated to the "Cleantech ecosystem". Identified companies were furthermore classified into different technology categories, such as environmental management, resources preservation, industrial energy management, capture, storage, sequestration or disposal of GHG, sustainable modes of transport and sustainable buildings. The database was subsequently enriched with financial accounting data (Orbis), patent information provided by (Orbis IP) and VC-investment data (VICO 4.0). The resulting novel database provides a unique perspective on the European Cleantech landscape.

The results presented in this paper provide a brief description of this database and serve as the introductory analysis of the EIBURS-funded[15] CLEU research project.[16] In doing so, we provided a number of interesting insights into the European Cleantech ecosystem.

Comparing our newly developed Cleantech classification to the traditional NACE sector classification, we found that Cleantech companies are predominantly active in the manufacturing, wholesale and retail trade, water supply and waste management, and construction sectors. Examining the spatial distribution of Cleantech in Europe, Germany, Italy, and France emerge as the key countries with the highest concentration of Cleantech companies. We also found Cleantech to be a well-established phenomenon, pre-dating to a large extent the two important Cleantech investment cycles, as a significant portion of the companies were established before the 2000s. We also analysed patenting activity of our Cleantech sample and found that Austria's Cleantech ecosystem is the most innovation-intensive, followed by Sweden and Germany, with

[15] The EIB University Research Sponsorship (EIBURS) programme provides grants to help EU universities and academic research centres to develop activities in selected research areas in addition to those that would normally be carried out by the beneficiary and on topics of major interest to the EIB Group (EIB and EIF).

[16] CLEU: The cleantech industry in the European Green Deal: policy challenges and the finance landscape for SMEs. For more information, see: *https://scienzeaziendali.unibo.it/en/research/research-projects/european-projects/cleu-the-cleantech-industry-in-the-european-green-deal-policy-challenges-and-the-finance-landscape-for-smes.*

sustainable energy production, energy-efficient industrial technologies, and air/water/soil pollution being the prominent technological categories for patenting. Investigating a selection of essential financial KPIs led us to conclude that Cleantech innovators tend to operate at a larger scale compared to their ecosystem counterparts, in terms of total assets, sales, and employee count. Finally, concerning VC financing, Finland, Sweden, France, and Spain emerge as the geographical areas with a high concentration of VC-backed companies.

The data presented in our study reaffirms the vibrant nature of the Cleantech industry in Europe. With a flourishing startup ecosystem, the sector is well-positioned to contribute significantly to the shift towards a low-carbon economy and tackle the urgent global challenge of climate change. Our findings also highlight notable variations among European countries, reflecting the diverse landscape in terms of investment levels, innovation and regulatory support across Member States. These differences underscore the need for targeted strategies tailored to each country's specific context to foster sustainable growth and maximize the potential of the Cleantech industry in Europe.

Annexes

Annex A: The evolution of the Cleantech concept

Cleantech, also known as clean technology, is an industry sector that focuses on developing and deploying sustainable and environmentally friendly solutions for various sectors. One of the most debated topics among academics and practitioners relates to the concept's exact definition. The origin of the term "cleantech" is somewhat difficult to trace to a single originator, as it has been used in various forms and contexts by different individuals and organisations over time. However, one of the earliest known uses of the term "clean technology" can be traced back to the early 1990s, when it was used by business and industry leaders in California to describe a new wave of environmentally friendly and sustainable technologies. According to some sources, John Balbach, a California-based consultant and entrepreneur, is sometimes credited with coining the term "cleantech". Balbach reportedly used the term to describe a broad range of emerging technologies and business models that focused on environmental sustainability and energy efficiency.

Given the importance of the topic, in 2002, the Cleantech Venture Network (nowadays known as Cleantech Group), a network of investors, was founded in San Francisco, California, "to address the lack of capital for clean technology ventures and the absence of a coordinated support network for the industry." (https://www.cleantech.com/). As part of its mission to help entrepreneurs, investors, corporations, service providers, government agencies, academic institutions, and non-profit organisations to connect, collaborate, and access the resources they needed to support and grow the cleantech industry, the organisation began using the term "cleantech" as a way of defining and promoting the sector, becoming an important segment of the broader technology and investment landscape. Lately, investors, in order to attract more capital, have chosen to use a different term referring to competitive (in terms of returns) investments that aim to solve the environmental problem (Caprotti, 2011). This strategy has led VC investments to grow by about 50% a year from 2004 to 2008, reaching a value of about $ 5 billion (Mills, 2015). However, the win-win strategy[17] has failed since the crisis of 2008.

From an academic perspective, one early and influential work on clean technology was the book "Soft Energy Paths: Toward a Durable Peace" by Amory Lovins (1977), which argued for a transition away from centralised, fossil-fuel-based energy systems towards decentralised, renewable-based systems. Later on, in addition to works investigating technical aspects of cleantech technologies, the management literature of cleantech started to be interested in the topic too, covering a wide range of topics, including: i) innovation and technology management (i.e., to explore the role of R&D, intellectual property, open innovation, and collaborative innovation in the development of cleantech products and services); ii) financing and investment (i.e., to investigate VC, public funding, and corporate investments as sources of capital available to cleantech firms and the factors that influence investment decisions, such as risk, uncertainty, and environmental regulations); iii) business models and strategy (i.e., to understand how firms create and capture value from their clean technologies); iv) environmental sustainability and

[17] Win-win because of the double goal: sustainability and profitability.

corporate social responsibility (i.e., to figure out how cleantech firms address sustainability challenges, such as energy and resource efficiency, waste reduction, and carbon emissions); v) supply chain management (i.e., to explore the complexities of managing global supply chains for cleantech products, including issues related to supplier selection, quality control, and sustainability); vi) organisational behaviour (i.e., to analyse the motivation, values, and beliefs of cleantech entrepreneurs and employees, as well as the impact of organisational culture and leadership on innovation and performance); vii) policy and regulation (i.e., to explore the effectiveness of different policy instruments, such as subsidies, tax incentives, and carbon pricing, in promoting the transition to a low-carbon economy, how firms comply with regulations and how regulatory frameworks affect the profitability of cleantech firms).

Annex B: Machine learning-based classification

The Phases of supervised Machine Learning

The initial step of our methodology was developing a ML classification tool to filter off all the companies which were certainly not Cleantechs. Specifically, we resorted to text classification, which is the process of automatically assigning documents to one or more predefined categories based on their content (Yang & Liu, 1999). Text classification can be naturally modelled as a supervised learning task in which a decision function is first derived by applying a machine-learning method on a set of labelled documents (training set), is evaluated on another set of labelled documents (test set), and then is applied to predict the category of incoming texts whose class is unknown.

To create the training set and the test set, we randomly selected from the initial sample 8,501 companies' descriptions. Each description was classified as Cleantech or non-Cleantech by two independent research assistants. As noted by Sebastiani (2002), it is easier for an analyst to characterize a concept extensionally, i.e., to select instances of it, rather than intentionally, i.e., to describe the concept in words. Consistently, we provided the research assistants a general definition of Cleantech as those technologies that have a positive environmental impact, e.g. in terms of reduction of the consumption of non-renewable resources or of waste produced (Pernick and Wilder, 2007). Any doubt about the classification of a particular technology, as well as any inconsistency was discussed between the research assistants and the authors until an agreement was reached. We provided research assistants with the textual description of the companies' business activity only in order to avoid the manual classification process being biased by information gathered from external, not standardised sources. Classified documents have been split between training and test set following a 70/30 rule. Table B.1 provides details upon this split.

Table B.1: Division of the sample between training and testing

	Negative	Positive	Sum
Train	5,833	118	5,951
Test	2,499	51	2,550
Sum	8,332	169	8,501

Feature extraction

Following the creation of the training and test sets, we prepared each document to generate a list of features used in the subsequent ML predictive task. Feature extraction techniques stem from text mining, which is converting free text into numerical variables that can be then analyzed using statistical tools (Feinerer et al., 2008).

The first step of this procedure is text pre-processing, which converts a vector of documents into a corpus, which is then converted into a *document term matrix* (DTM). The algorithm is called "bag-of-words". This first step also includes the text cleaning process (e.g. converting the entire text to lowercase characters and removing all unnecessary punctuation and symbols). In our analysis, we used different text-cleaning procedures, specifically:

- converting upper case letters to lower case;
- spell checking;
- substitution of contractions, such as converting “I’m” to “I am”;
- removing numbers;
- removing punctuation and special characters (for instance @,°,# [,§ etc.);
- converting acronyms to regular expressions, such as “IT” to “Italy”.

Furthermore, *stop words* were removed (a list of the stop words is reported in Annex C). Stop words are frequently used in texts but are poorly useful for predictions; as such, they can be omitted from the analysis.

After carrying out the pre-processing phase, we performed text normalisation to convert words into their simplest form. In fact, it is known that any language includes words of various tenses, plurals, or derived from other words. Our goal was to reduce the words to their common root. A “lemmatisation” process was developed to group the various inflected forms of a word so that they can be analysed as a single lemma. An alternative to lemmatisation is “stemming”. Unlike lemmatisation, stemming generates meaningless roots from words whose semantic nature can be very poor.

Using the abovementioned corpus of documents, we constructed our Document Term Matrix. This object is a simple matrix structure, with each document as a row and each n-gram (or term) as a column. Once we constructed the DTM, we calculated the frequencies (total times each n-gram appears in all documents), with the n-grams as the names of the vector. In our analysis, we created two different tokeniser functions to construct the DTM for 1-gram and 2-grams.

Only n-grams with a frequency higher or equal to a specific threshold are kept in the model. The threshold is given using sparsity optimisation, which refers to the relative document frequency threshold for a term. In our case, with a sparseness equal to 0.50, only terms occurring in half of the documents were retained. At the end of this process, we extrapolated 251 1-gram and 181 2-grams for a total of 432 features.[18] These features are used for training the Cleantech predictive model.

Prediction Task

The prediction task aims to produce an accurate predicting map between the features generated in phase 2 and the “true” Cleantech label (1 = Cleantech firm; 0 = non-Cleantech firm) using the available training dataset. This mapping is then projected to the unlabelled firms to provide them with an appropriate Cleantech predicted label. We used and compared various popular ML

[18] Examples of features extrapolated and used for training the Cleantech predictive model are: 1) for 1-gram features: Taxonomy, clean, electr, electron, energy, fuel, gas, power, process, produc, product, service, storag, util, vehicle, water; 2) for 2-grams features: building_materi, company_engag, company_produc, company_provid, construction_materi, construction_servic, engineering_servic, engineering_work, food_product, fresh_fruit, frozen_food, high_qual, industrial_build, information_technolog, maintenance_servic, management_servic, metal_product, motor_vehicl, necessary_assist, new_work, operation_specialis, personal_car, petroleum_product, pharmaceutical_product, predominantly_oper, primarily_engag, public_organ, raw_materi, related_servic, repair_servic, residential_build, retail_distribut, retail_sal, retail_trad, stateoftheart_technolog, transportation_servic, utility_vehicl, wholesale_distribut, wholesale_trad, wide_rang, wide_varieti.

methods to carry out this task by paying particular attention to their out-of-sample predicting performance. We selected the classifier providing the best test accuracy.

Consider a firm i with an associated target binary variable P_i ("being Cleantech") that takes values one (positive occurrence) if the firm is (manually) classified as Cleantech and value zero (negative occurrence) otherwise. Based on the set of features created by the previous bag-of-words algorithm ($Features_i$) and referring to the firm i, our prediction task was to find a mapping function $f(\cdot)$ (i.e., a ML binary classifier) that predicts as best as possible the Cleantech event (i.e., the two-class variable P_i):

$$Features_i \xrightarrow{f(.)} P_i \tag{1}$$

The features used to predict P_i are those described in the previous section and refer to the frequency of lemmatised words (251 1-gram features). The standard ML procedure to first train and then test the mapping function is that of randomly splitting the data into a training set, over which the model is estimated and tuned, and a testing set, over which its predictive power is tested (Hastie et al., 2009).

The size of these two sets must be chosen considering the trade-off between the benefit of a large training set (i.e., more information for building the mapping in (1)) and the benefit of a sufficiently large testing set (i.e., larger information for a more precise estimation of the testing error). To account for this trade-off, we followed the usual compromise of randomly dividing the database into 70% of observations for training and the remaining observations as an out-of-sample test set (Boehmke and Greenwell, 2019).

To carry out our analysis, we used four different ML predicting algorithms:

- Naïve Bayes (NB): a discriminant classifier assuming no correlation among the features in each class (Gareth et al., 2013);
- Random Forest (RF): a family of randomised tree-based classifier decision trees which uses different random subsets of the features at each split in the tree (Breiman, 2001);
- Gradient Boosting Machines (GBM): an ensemble method which works in an iterative way where at each stage new learner tries to correct the pseudo-residual of its predecessors (Friedman, 2001);
- Neural Network (NN): a model that uses a set of connected input/output units in which each connection has an associated weight and learns by adjusting the weights to predict the correct class label of the given inputs (Ripley et al., 2016).

We also dealt with target variable imbalance. Indeed, firms classified as Cleantech are largely fewer than firms not classified as Cleantech (169 out of 8,501) and this generally produces under-performing predictions, as the unconditional probability to be Cleantech is highly skewed towards the absence of Cleantech, thus giving this category a larger advantage when classifying new instances. To address the imbalance issue, the four ML models were estimated using two different strategies:

- *Class weighting*: a method imposing a higher cost when errors are made in the minority class (based on the *inverse probability weighting* of the observations);
- *Random Over-Sampling Examples (ROSE)*: an algorithm where artificially balanced samples are generated according to a smoothed bootstrap approach (Lunardon et al., 2014).

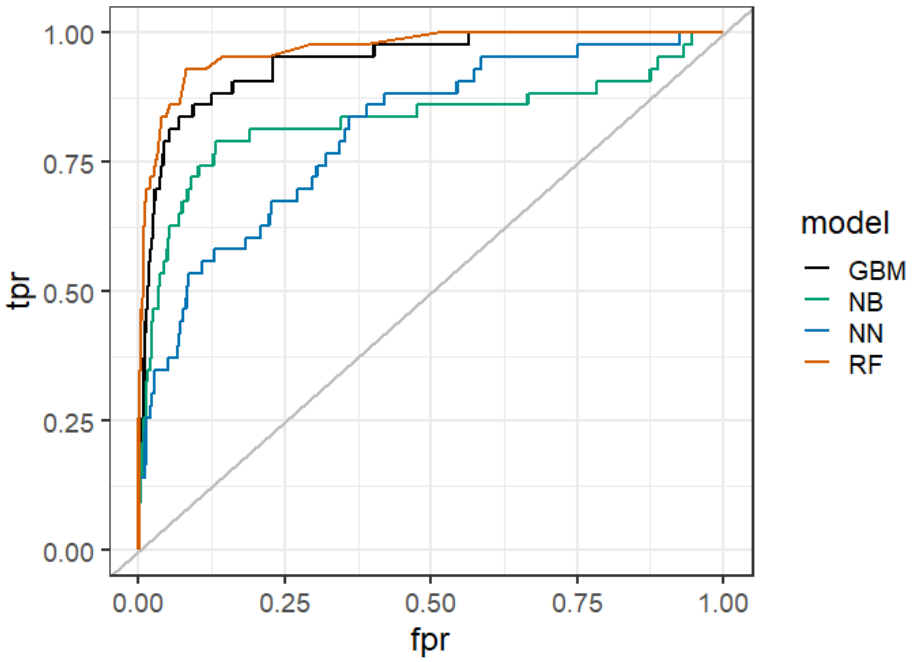
The hyper-parameters optimisation was carried out over the training set using 10-fold repeated cross-validation, with five repetitions in the case of Class weights and 50 repetitions in the case of ROSE. All models were implemented using the R software, trained with the optimisation algorithms available through the CARET package (Kuhn, 2020).

The performance of Cleantech classification prediction was assessed through the *Receiver Operating Characteristics (ROC) curve* (Fawcett, 2006). The ROC curve shows the classifier's diagnostic ability by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis since its discrimination threshold is varied (Antulov-Fantulin et al., 2021).

Results using the class weighting strategy

Figure B.1 shows the ROC curves for the four algorithms trained on 70% of the observations and tested on the remaining 30% using the class weighting strategy. The estimates are based on the cross-validation algorithm, which trains and tests the model by tuning the hyper-parameters to maximise the area under the ROC curve.

Figure B.1: Class weighting, ROC curves (on the test set)*



* Repeated Cross-validation: Number (number of folds or number of resampling iterations) = 10. Repeat (for repeated k-fold cross-validation only: the number of complete sets of folds to compute) = 5. The hyper-parameters best tune used in the final models with class weighting is shown in Table B.2.

Table B.2: Class weighting: best tuning parameter

NB	RF	GBM	NN
fL = 0	Mtry= 2	n.trees = 150	size = 1
Usekernel = TRUE		interaction.depth =1	decay = 0.1
Adjust = 1		shrinkage = 0.1	
		n.minobsinnode = 10	

Table B.3 shows the model performance for the class weights strategy; one class means the model classifies all as negative due to the unbalancing of the data. The only model predicting positives in the case of Class weights is the GBM. The performances of GBM are satisfactory (Accuracy=0.937; and AUC=0.944).

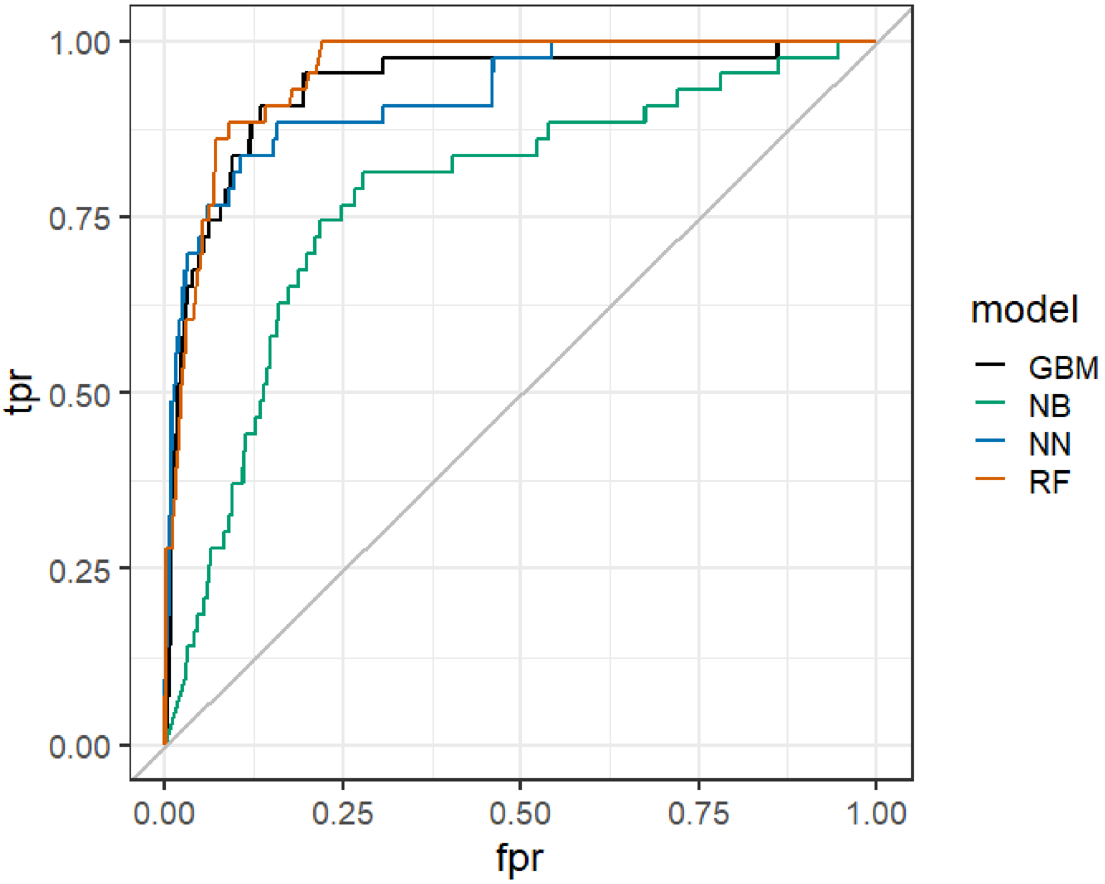
Table B.3: Class weighting: model performances (on the test set)

	GBM	RF	NN	NB
Accuracy	0.937	one class	one class	one class
Sensitivity	0.814	one class	one class	one class
Specificity	0.939	one class	one class	one class
Pos Pred Value	0.185	one class	one class	one class
Neg Pred Value	0.997	one class	one class	one class
Prevalence	0.017	one class	one class	one class
Detection Rate	0.014	one class	one class	one class
Detection Prevalence	0.074	one class	one class	one class
Balanced Accuracy	0.876	one class	one class	one class
Area under the curve	0.944	one class	one class	one class

Results using the ROSE strategy

Figure B.2 shows the ROC curves for the four algorithms trained on 70% of the observations and tested on the remaining 30% using the ROSE strategy. As before, the estimates are based on the cross-validation algorithm, which trains and tests the model tuning the hyper-parameters to maximise the area under the ROC curve.

Figure B.2: ROSE, ROC curves (on the test set)*



*Repeated Cross-validation: Number (number of folds or number of resampling iterations) = 10. Repeat (for repeated k-fold cross-validation only: the number of complete sets of folds to compute) = 50. Hyper-parameters best tune used in the final models with ROSE is shown in Table B.4.

Table B.4: ROSE, best tuning parameter

NB	RF	GBM	NN
fL = 0	Mtry= 2	n.trees = 100	size = 1
Usekernel = TRUE		interaction.depth =1	decay = 0.1
Adjust = 1		shrinkage = 0.1	
		n.minobsinnode = 10	

Table B.5 shows the model performance for the ROSE strategy. In this case, all models classify both categories (positives and negatives). The best model for accuracy and AUC is the GBM (Accuracy=0.922; and AUC=0.941).

Table B.5: ROSE: model performances (on the test set)

	GBM	RF	NN	NB
Accuracy	0.922	0.854	0.844	0.629
Sensitivity	0.744	0.907	0.860	0.814
Specificity	0.925	0.853	0.844	0.625
Pos Pred Value	0.146	0.096	0.086	0.036
Neg Pred Value	0.995	0.998	0.997	0.995
Prevalence	0.017	0.017	0.017	0.017
Detection Rate	0.013	0.015	0.015	0.014
Detection Prevalence	0.086	0.160	0.168	0.382
Balanced Accuracy	0.835	0.880	0.852	0.720
Area under the curve	0.941	0.939	0.942	0.779

Best model selection results

It is worth noticing that the results presented above refer to a 1-gram DTM, as results from using a 2-grams DTM are equivalent but require a more considerable computational burden.

In both the strategies used to deal with the unbalanced data (class weighting and ROSE), the algorithm with the best performance is GBM. Regarding accuracy and AUC, the GBM combined with class weighting outperforms the GBM combined with ROSE, so we selected the GBM combined with class weighting for final prediction. Table B.6 shows our main result, the confusion matrix for the best model on the entire database.

Table B.6: Confusion matrix with the best model (GBM combined with Class weights)

	Negative	Positive	Sum
Negative	8,538	1,005	9,543
Positive	58	399	457
Not Classified	454,486	72,643	527,129
Sum	463,082	74,047	537,129

The result of this phase was a sample reduced to 74,047 companies (out of the initial sample made of 537,129 companies).

Annex C: List of stop words

a
a's
able
about
above
according
accordingly
across
actually
after
afterwards
again
against
ain't
all
allow
allows
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
appreciate
appropriate
are
aren't
around
as
aside
ask
asking
associated
at
available
away
awfully
b
be
became
because
become
becomes
becoming
been
before
beforehand
behind
being
believe
below
beside
besides
best
better
between
beyond
both
brief
but
by
c
c'mon
c's
came
can
can't
cannot
cant
cause
causes
certain
certainly
changes
clearly
co
com
come
comes
concerning
consequently
consider
considering
contain
containing
contains
corresponding
could
couldn't
course
currently
d
definitely
described
despite
did
didn't
different
do
does
doesn't
doing
don't
done
down
downwards
during
e
each
edu
eg
eight
either
else
elsewhere
enough
entirely
especially
et
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
exactly
example

except
f
far
few
fifth
first
five
followed
following
follows
for
former
formerly
forth
four
from
further
furthermore
g
get
gets
getting
given
gives
go
goes
going
gone
got
gotten
h
had
hadn't
happens
hardly
has
hasn't
have
haven't
having
he
he's
hello
help
hence
her
here
here's
hereafter
hereby
herein
hereupon
hers
herself
hi
him
himself
his
hither
hopefully
how
howbeit
however
i
i'd
i'll
i'm
i've
ie
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
isn't
it
it'd
it'll
it's
its
itself
j
just
k
keep
keeps
kept
know
knows
known
l
last
lately
later
latter
latterly
least
less
lest
let
let's
like
liked
likely
little
look
looking
looks
ltd
m
mainly
many
may
maybe
me
mean
meanwhile
merely
might
more
moreover
most
mostly
much
must
my
myself
n
name
namely
nd
near
nearly
necessary
need
needs
neither
never
nevertheless
new
next
nine
no
nobody
non

none
noone
nor
normally
not
nothing
novel
now
nowhere
o
obviously
of
off
often
oh
ok
okay
old
on
once
one
ones
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
own
p
particular
particularly
per
perhaps
placed
please
plus
possible
presumably
probably
provides
q
que
quite
qv
r
rather
rd
re
really
reasonably
regarding
regardless
regards
relatively
respectively
right
s
said
same
saw
say
saying
says
second
secondly
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sensible
sent
serious
seriously
seven
several
shall
she
should
shouldn't
since
six
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
t's
take
taken
tell
tends
th
than
thank
thanks
thanx
that
that's
thats
the
their
theirs
them
themselves
then
thence
there
there's
thereafter
thereby
therefore
therein
theres
thereupon
these
they
they'd
they'll
they're
they've
think
third
this
thorough
thoroughly
those

though
three
through
throughout
thru
thus
to
together
too
took
toward
towards
tried
tries
truly
try
trying
twice
two
u
un
under
unfortunately
unless
unlikely
until
unto
up
upon
us
use
used
useful
uses
using
usually
uucp
v
value
various
very
via
viz
vs
w
want
wants
was
wasn't
way
we
we'd
we'll
we're
we've
welcome
well
went
were
weren't
what
what's
whatever
when
whence
whenever
where
where's
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
who's
whoever
whole
whom
whose
why
will
willing
wish
with
within
without
won't
wonder
would
would
wouldn't
x
y
yes
yet
you
you'd
you'll
you're
you've
your
yours
yourself
yourselves
z
zero

References

- Antulov-Fantulin, N., Lagravinese, R., & Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior & Organization*, 183, 681–699. https://doi.org/10.1016/j.jebo.2021.01.014
- Bellucci, A., Fatica, S., Georgakaki, A., Gucciardi, G., Letout, S., & Pasimeni, F. (2021). Venture capital financing and green patenting (No. 2021/11). JRC Working Papers in Economics and Finance.
- Boehmke, B., & Greenwell, B. M. (2019). Hands-on machine learning with R. CRC press.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Caprotti, F. (2011). The cultural economy of cleantech: Environmental discourse and the emergence of a new technology sector. Transactions of the Institute of British Geographers, 37(3), 370–385. https://doi.org/10.1111/j.1475-5661.2011.00485.x
- Christensen, J. L., & Hain, D. S. (2017). Knowing where to go: The knowledge foundation for investments in renewable energy. *Energy Research and Social Science*, 25, 124–133. https://doi.org/10.1016/j.erss.2016.12.025
- Cleantech Group. (2021). From Chaos to Transformation: The Companies and Themes Delivering Sustainable Innovation (Global Cleantech 100, p. 61). Cleantech Group. https://www.cleantech.com/release/cleantech-group-names-2021-global-cleantech-100-the-top-innovators-driving-sustainability-and-change/
- Criscuolo, C., & Menon, C. (2015). Environmental policies and risk finance in the green sector: Cross-country evidence. *Energy Policy*, 83, 38–56. https://doi.org/10.1016/j.enpol.2015.03.023
- Cumming, D., Henriques, I., & Sadorsky, P. (2016). 'Cleantech' venture capital around the world. *International Review of Financial Analysis*, 44, 86–97. https://doi.org/10.1016/j.irfa.2016.01.015
- Ezgi Bezirhan Arikan & Havva Duygu Ozsoy. (2015). A Review: Investigation of Bioplastics. *Journal of Civil Engineering and Architecture*, 9(2). https://doi.org/10.17265/1934-7359/2015.02.007
- De Haan Montes, G., Gvetadze, S., Lottner, F., Milander, H., Pan, X., Tian, C., & Torfs, W. (2023). Determinants of EU Greentech investments: the role of financial market conditions (No. 2022/86). EIF Working Paper
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
- Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5). https://doi.org/10.18637/jss.v025.i05
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232. JSTOR.
- Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning : with Applications in R, New York, Springer.
- Haščič, I., & Migotto, M. (2015). Measuring environmental innovation using patent data. OECD. https://doi.org/10.1787/5js009kf48xw-en
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

- IEA-EPO (2021). Patents and the energy transition: global trends in clean energy technology innovation. IEA-EPO Report, downloadable at *https://www.iea.org/reports/patents-and-the-energy-transition*
- Kraemer-Eis, H., Gvetadze, S., Lang, F., and Torfs, W. (2022). European Small Business Finance Outlook. EIF Working Paper 2022/84. EIF Research & Market Analysis. November 2022.
- Kraemer-Eis, H., Signore, S., & Prencipe, D. (2016). The European venture capital landscape: an EIF perspective. Volume I: The impact of EIF on the VC ecosystem (No. 2016/34). EIF Working Paper.
- Kuhn, M. (2020). caret: Classification and Regression Training (R package version 6.0-85). *https://CRAN.R-project.org/package=caret*. R package version 6.0-85
- Lovins, A. B. (1977). Soft energy paths: Toward a durable peace. Friends of the Earth International,San Francisco. *https://www.osti.gov/biblio/7088135*
- Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a package for binary imbalanced learning. *R journal*, 6(1).
- Mills, L. (2015). Global trends in clean energy investment (No. 5). Bloomberg New Energy Finance.
- Pitchbook (2023) Clean Energy Report. VC trends and emerging opportunities. Pitchbook Emerging Tech Research. Q1/2023
- Ripley, B., Venables, W., & Ripley, M. B. (2016). Package 'nnet'. R Package Version, 7(3–12), 700.
- Soete, L. (1987). The impact of technological innovation on international trade patterns: the evidence reconsidered. *Research policy*, 16(2-4), 101-130.
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. 42–49.

About

... the European Investment Fund

The European Investment Fund (EIF) is Europe’s leading risk finance provider for small and medium sized enterprises (SMEs) and mid-caps, with a central mission to facilitate their access to finance. As part of the European Investment Bank (EIB) Group, EIF designs, promotes and implements equity and debt financial instruments which specifically target the needs of these market segments.

In this role, EIF fosters EU objectives in support of innovation, research and development, entrepreneurship, growth, and employment. EIF manages resources on behalf of the EIB, the European Commission, national and regional authorities and other third parties. EIF support to enterprises is provided through a wide range of selected financial intermediaries across Europe. EIF is a public-private partnership whose tripartite shareholding structure includes the EIB, the European Union represented by the European Commission and various public and private financial institutions from European Union Member States and Turkey. For further information, please visit *www.eif.org*.

... EIF’s Research & Market Analysis

Research & Market Analysis (RMA) supports EIF’s strategic decision-making, product development and mandate management processes through applied research and market analyses. RMA works as internal advisor, participates in international fora and maintains liaison with many organisations and institutions.

... this Working Paper series

The EIF Working Papers are designed to make available to a wider readership selected topics and studies in relation to EIF´s business. The Working Papers are edited by EIF´s Research & Market Analysis and are typically authored or co-authored by EIF staff, or written in cooperation with EIF. The Working Papers are usually available only in English and distributed in electronic form (pdf).

EIF Working Papers

2023/091	*Using machine learning to map the European Cleantech sector.*
2023/090	*Recent trends in EU corporate demography and policy*
2023/089	*Entrepreneurial Finance and the Russian War against Ukraine: a Survey of European Venture Capital and Private Equity Investors*
2023/088	*Fairness in algorithmic decision systems: a microfinance perspective*
2023/087	*Female access to finance: a survey of literature.*
2023/086	*Determinants of EU Greentech investments: the role of financial market conditions*
2022/085	*EIF Private Equity Mid-Market Survey 2022: Market sentiment and impact of the current geopolitical & macroeconomic environment.*
2022/084	*European Small Business Finance Outlook 2022.*
2022/083	*The 2021 EIF SME Access to Finance Index. October 2022 update.*
2022/082	*EIF VC Survey 2022: Market sentiment and impact of the current geopolitical & macroeconomic environment.*
2022/081	*EIF Business Angel Survey 2021/22: Market sentiment.*
2022/080	*EIF Private Debt Survey 2021 - ESG considerations in the lending strategy of private debt funds.*
2022/079	*EIF Private Debt Survey 2021 - Private Debt for SMEs - Market overview.*
2021/078	*EIF Private Equity Mid-Market Survey 2021: Market sentiment.*
2021/077	*EIF Private Equity Mid-Market Survey: Fund managers' perception of EIF's value added.*
2021/076	*The 2020 EIF SME Access to Finance Index - October 2021 update.*
2021/075	*European Small Business Finance Outlook.*
2021/074	*EIF Venture Capital Survey 2021: Market sentiment.*
2021/073	*EIF VC Survey Autumn 2020 - Regional analysis.*
2021/072	*Financial and social inclusion in Europe.*
2021/071	*EIF Venture Capital, Private Equity Mid-Market & Business Angels Surveys 2020: Market sentiment – COVID-19 impact – Policy measures.*
2021/070	*The impact of VC on the exit and innovation outcomes of EIF-backed start-ups.*
2021/069	*Scale-Up Financing and IPOs: Evidence From Three Surveys.*
2020/068	*The EIF SME Access to Finance Index - September 2020 update.*
2020/067	*European Small Business Finance Outlook 2020. The impact of COVID-19 on SME Financing markets.*
2020/066	*Measuring microfinance impact: A practitioner perspective and working methodology.*
2020/065	*The social return on investment (SROI) of four microfinance projects.*
2020/064	*The market sentiment in European Private Equity and Venture Capital: Impact of COVID-19.*
2020/063	*ESG considerations in Venture Capital and Business Angel investment decisions:*

	Evidence from two pan-European surveys.
2020/062	*The Business Angel portfolio under the European Angels Fund: An empirical analysis.*
2019/061	*European Small Business Finance Outlook.*
2019/060	*EIF Business Angels Survey 2019 - Market sentiment, public intervention and EIF's value added.*
2019/059	*EIF VC Survey 2019 - Fund managers' market sentiment and policy recommendations.*
2019/058	*EIF SME Access to Finance Index – June 2019 update.*
2019/057	*European Small Business Finance Outlook.*
2019/056	*The real effects of EU loan guarantee schemes for SMEs: A pan-European assessment.*
2019/055	*The European Venture Capital Landscape: an EIF perspective. Volume V: The economic impact of VC investments supported by the EIF.*
2019/054	*Econometric study on the impact of EU loan guarantee financial instruments on growth and jobs of SMEs.*
2018/053	*European Small Business Finance Outlook.*
2018/052	*The effects of EU-funded guarantee instruments of the performance of Small and Medium Enterprises - Evidence from France.*
2018/051	*EIF VC Survey 2018 - Fund managers' perception of EIF's Value Added.*
2018/050	*European Small Business Finance Outlook.*
2018/049	*EIF SME Access to Finance Index – June 2018 update.*
2018/048	*EIF VC Survey 2018 – Fund managers' market sentiment and views on public intervention.*
2018/047	*EIF SME Access to Finance Index.*
2017/046	*European Small Business Finance Outlook.*
2017/045	*The European venture capital landscape: an EIF perspective. Volume IV: The value of innovation for EIF-backed startups.*
2017/044	*Financing Micro Firms in Europe: An Empirical Analysis.*
2017/043	*European Small Business Finance Outlook.*
2017/042	*Credit Guarantee Schemes for SME lending in Western Europe.*
2017/041	*The European Venture Capital landscape: an EIF perspective. Volume III: Liquidity events and returns of EIF-backed VC investments.*
2017/040	*Financing Patterns of European SMEs Revisited: An Updated Empirical Taxonomy and Determinants of SME Financing Clusters.*
2017/039	*Guaranteeing Social Enterprises – The EaSI way.*
2016/038	*The European Venture Capital Landscape: an EIF perspective. Volume II: Growth patterns of EIF-backed startups.*
2016/037	*European Small Business Finance Outlook.*
2016/036	*The role of cooperative banks and smaller institutions for the financing of SMEs and small midcaps in Europe.*
2016/035	*European Small Business Finance Outlook.*

2016/034	*The European Venture Capital Landscape: an EIF perspective. Volume I: the impact of EIF on the VC ecosystem.*
2016/033	*Evaluating the impact of European microfinance. The foundations.*
2015/032	*European Small Business Finance Outlook.*
2015/031	*SME Securitisation – at a crossroads?*
2015/030	*Financing patterns of European SMEs: An Empirical Taxonomy.*
2015/029	*The Economic Impact of EU Guarantees on Credit to SMEs - Evidence from CESEE Countries.*
2015/028	*European Small Business Finance Outlook.*
2015/027	*Bridging the university funding gap: determinants and consequences of university seed funds and proof-of-concept Programs in Europe.*
2014/026	*European Small Business Finance Outlook.*
2014/025	*Institutional non-bank lending and the role of debt funds.*
2014/024	*European Small Business Finance Outlook.*
2014/023	*Pricing Default Risk: the Good, the Bad, and the Anomaly.*
2014/022	*Guidelines for SME Access to Finance Market Assessments.*
2014/021	*Financing the mobility of students in European higher education.*
2013/020	*European Small Business Finance Outlook.*
2013/019	*SME loan securitisation 2.0 – Market assessment and policy options.*
2013/018	*European Small Business Finance Outlook.*
2013/017	*Forecasting distress in European SME portfolios.*
2012/016	*European Small Business Finance Outlook.*
2012/015	*The importance of leasing for SME finance.*
2012/014	*European Small Business Finance Outlook.*
2012/013	*Progress for microfinance in Europe.*
2011/012	*European Small Business Finance Outlook 2/2011.*
2011/011	*Business Angels in Germany. EIF's initiative to support the non-institutional financing market.*
2011/010	*European Small Business Finance Outlook.*
2011/009	*The performance and prospects of European Venture Capital.*
2010/008	*Impact of Legislation on Credit Risk – How different are the U.K. and Germany?*
2010/007	*SME Loan Securitisation – an important tool to support European SME lending.*
2010/006	*Drivers of Private Equity Investment activity. Are Buyout and Venture investors really so different?*
2010/005	*Private Equity Market Outlook.*
2010/004	*Private Equity and Venture Capital Indicators – A research of EU-27 Private Equity and Venture Capital Markets.*

#Believe InSmall

www.eif.org

Scan above to
obtain a PDF
version of this
working paper

Follow our stories and work on social media

European Investment Fund

@eif4smes

@EIF_EU

engage.eif.org