The determinants of science-based green patents

Nicoletta Corrocher[a], Andrea Morrison[a,b], Federico Nutarelli[a*]

[a] ICRIOS, Bocconi University, Milan, Italy
[b] University of Pavia, Pavia, Italy

Abstract

Assessing the determinants of the most significant green patents has emerged as a crucial goal not only among scholars but also for policymakers. However, a common definition of the importance based on citation patterns, as provided in Chai et al., 2020, has been limited to the light-emitting diode (LED) industry. Such a definition is constrained by unobservable factors that lead to patents being highly cited, such as self-citations, a bias towards older patents that have had more time to accumulate citations, variations in citation practices, and patent scope. To address these issues, we propose a measure of importance based on the distance from a paper-patent boundary, which represents the scientific frontier. Our study demonstrates that green patents grounded in scientific principles are not only important in their own right but also occupy a central position within the citation network. By utilizing this definition of importance, we investigate the factors that contribute to the scientific basis of green patents. Our analysis reveals a critical finding: the percentage of green citations is a key determinant of the scientific basis of patents. This empirically underscores the need to prioritize "pure green" innovations over "brown" innovations (Heal, 2007) in order to be recognized as science-based and consequently occupy a central position in the green citation network.

Keywords: Green innovation, green patents, science-based patents
JEL Codes: O30; O34; Q55

1 Introduction

A resurgence of interest in the scholarly community has emerged around the role of science in society and in particular on its link with technological innovation. This is not surprising if we consider that, on the one side, humanity is facing important challenges, which represent serious threats (e.g. mounting hazards of climate change; health hazards, and related diseases like Covid-19); on the other side, humanity seems also to produce fewer brilliant ideas to address such challenges, so the role of science (and its public support) could be put under discussion. If innovation can be the solution to address some or many of the most pressing societal challenges we are facing, then the question is whether science can feed technological innovation and to what extent. In this paper, we focus our attention on one of such challenges, i.e. climate change, and one of its possible technological fixes, that is the development of climate change mitigation technologies. In this context, there is a particular sense of urgency, as the catastrophic effects of climate change are tangible. Therefore, green technologies need to be deployed rather quickly and effectively. In addition, climate change is a global and pervasive phenomenon, whose public good features are unquestionable and call for strong public support. So, empirical evidence suggesting that public spending in science would generate the desired results in terms of deployable green innovation is certainly welcome. We address this topic by investigating the link between scientific advances and green patents. Understanding the scientific roots of climate change-mitigation technologies will contribute to the emerging debate on the role of science in society. In addition, we will explore the determinants of science-based green technologies. This analysis will provide insights to the existing science policy debate, as it will show the who, where and why of the science-based green invention. We build our empirical exercise on the work of Ahmadpoor and Jones (2017) (henceforth AJ), who first proposed a distance measure to capture the extent to which a patent builds on prior scientific advances. Using their methodology, we are able to identify the "green

*Corresponding author: federico.nutarelli@unibocconi.it. Invernizzi Center for Research on Innovation, Organization, Strategy and Entrepreneurship, Via Guglielmo Rontgen, 1 20136 Milano, Italy.

patent-paper boundary", i.e. green patents that are at the scientific frontier. We illustrate the main feature of these green patents, also in a comparative way, in terms of their performance, technological domain, geographical distribution, institutional origin and dynamics over time. Our descriptive findings suggest, in line with AJ, that a relatively large share of green patents are linked to some scientific output, even if indirectly. Furthermore, green patents that build on scientific discoveries are found to be widely cited, which is in line with AJ results. This clearly points to the relevance of scientific advances also for clean technologies. Green science-based patents, not surprisingly, are mainly produced by universities or governments, but also by companies with links to universities. Finally, the geographical distribution shows that most countries leading in green patents also lead in science-based green patents. Interestingly, this is the case of emerging green leaders, like China, and Korea. We test the predictive power of these findings by using a machine learning methodology which reports an increase in the weighted importance of China and Korea of about 37% over time.

2 The link between science and innovation in green technologies: a literature review

For centuries, the relationship between science and technology has been the subject of intense discussions. Policymakers have for a long time considered science as the main source of knowledge that would ultimately contribute to the emergence of new technical and organizational capabilities, improvements in quality of life, and economic growth. Scientific discoveries indeed play a crucial role in driving economic growth through innovations: as documented in Poege et al. (2019), the quality of scientific publications is a strong predictor of their impact on technological development and at the same time the value of patents that are directly based on scientific research increases in proportion to the scientific quality. Many of the most valuable innovations depend on scientific knowledge and for commercial inventors looking to capitalize on new technologies science can act as a source of inspiration for their own research and development activities (Gittelman and Kogut, 2003; Fleming and Sorenson, 2004).

The idea of a publicly funded science system that feeds into privately organized innovation channels has been for a long time the model for most national systems of innovation. However, this notion has recently faced scrutiny, as there is an increase in the demand for evidence of the benefits of science spending. The fact that science quality is defined within the realm of science itself contributes to a perception of science as being an independent upstream activity, at times detached from technological progress, with an indirect and delayed impact on society at best. The investigation of the science-based sources of technological development has become over time an increasingly relevant topic, due to changes in firms' internal R&D processes, market evolution and new norms and policies (Fleming and Sorenson, 2004; Arora et al., 2018; Marx and Fuegi, 2020).

If science plays a relevant role as a source of inspiration and as a possible way for firms to differentiate from competitors, it is important to trace the scientific base of R&D also to understand, from a policy perspective, what are the scientific domains that are most conducive to innovations and how public funding might be allocated. Therefore, it is crucial for policymakers and scientists to improve their understanding of the impact of science on technical progress and innovation. From a research perspective, a large number of works have looked at the magnitude of patents' citations to science, which constitutes an interesting indicator to identify the characteristics of the search process implemented by firms to develop innovations, the novelty of inventions, and the extent to which knowledge spills over from the universities to private companies (Katila and Ahuja, 2002; Gittelman and Kogut, 2003; Fleming and Sorenson, 2004; Ahmadpoor and Jones, 2017; Wang et al., 2017). The extent to which innovations rely on science varies among different sectors and technologies (Pavitt, 1984; Verbeek et al., 2002). Along this line of reasoning, Fleming and Sorenson (2004) argue that technologies with many interconnected components may benefit more from scientific knowledge. Moreover, the use of scientific literature may help inventors explore new fields and discover novel combinations of knowledge (Arts and Fleming, 2018).

The investigation of the role of (public) science in the production of (private) technological innovations is particularly relevant in the case of green technologies (OECD, 2010; Popp, 2017; Persoon et al., 2020). Green technologies have an interesting characteristic in that they utilize new and different combinations of knowledge compared to non-green technologies, making them novel (Barbieri et al., 2020). These innovations are expected to bring radical change due to the absence of established environmental best practices and technological trajectories (Verhoeven et al., 2016). A recent study by Persoon et al. (2020) compares innovations in renewable energy technologies with innovations in fossil fuel energy technologies and finds that innovations in renewable rely more on science because they are more radical and are based on novel knowledge combinations and technological breakthroughs, which are associated with scientific

breakthroughs. Moreover, green technologies are characterized by technological uncertainty and require skills outside of the firm's knowledge domain (De Marchi, 2012; Ghisetti et al., 2017). Popp (2017) for example shows that research funded and performed by the government plays an important role in linking basic and applied research.
Besides the intrinsic novelty of green technologies, previous studies have also shown that these technologies are more complex than non-green ones, as they involve a wider range of objectives and knowledge inputs (De Marchi, 2012). The increased complexity of green technologies is also evident in the multi-purpose and systemic nature of environmental innovations (Ghisetti et al., 2015). Environmental technologies are expected to achieve various joint objectives, such as production efficiency and product quality and involve several dimensions, including design, user involvement, product-service delivery, institutional requirements, and regulatory frameworks (Carrillo-Hermosilla et al., 2010; Mazzanti and Rizzo, 2017). Finally, green technologies generate extensive spillovers to subsequent technological developments and these spillovers apply to very different technological domains and sectors (Barbieri et al., 2020). Therefore, stimulating the development of these very pervasive technologies might also mean investing in science that supports the most valuable and impactful inventions. This represents a crucial issue in the current climate policy debate (Cárdenas Rodríguez et al., 2014; Popp, 2017), since public actors play a significant enabling role in the development of green technologies as they are more inclined than private investors to invest in projects with a higher degree of risk (Kapoor and Oksnes, 2011; IRENA and CPI, 2018; Mazzucato and Semeniuk, 2018).

3 Data

In order to explore the determinants of science-based green patents, we rely on the two main sources of data: the PATSTAT database and the patents' citations to science collected in Marx and Fuegi, 2020[1].
The PATSTAT dataset is a comprehensive collection of worldwide patents maintained by the European Patent Office (EPO). It contains detailed information on approximately 100 million patent documents worldwide, including both applications and granted patents. The data in PATSTAT cover a broad range of technological fields and geographic regions and are constantly updated. The PATSTAT database includes a vast range of variables, including information on patent publications (such as CPC technology class, patent title, and claims), patent citations (both forward and backward citations), patent family members, legal status events (such as patent grants, abandonments, and expirations), and patent applicant variables (such as applicant and inventor names, countries, and corporations).
In what follows we will focus on the CPC Y.
The full PATSTAT database contains information about almost 2 Million green patents from 1949 to 2021 [2].
Fig.1 presents the number of green patents' applications by year in PATSTAT.

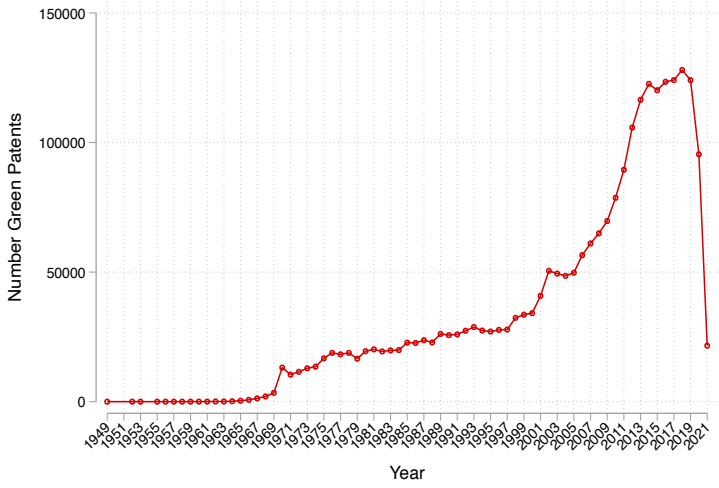


Figure 1. Number of green patents by year.

[1]Available at `http://relianceonscience.org`. Notice that we adopted an updated version of the database which adopts EMAKG rather than Microsoft Academic Graph (MAG) in order to form patent-papers linkages. EMAKG is an enhanced version of MAG.
[2]It should be noted that the information presented is valid as of the time of writing

The database provided by Marx and Fuegi, 2020 includes patent-to-paper linkages for worldwide patents, which are referred to as Patent Citations to Science (PCS). In addition to PCS, the database provides information on the papers themselves, such as the names of the authors, their affiliations, whether the paper has been presented at conferences, its title and abstract, and its field. However, the database does not contain details about the patents. In our work, we focus on the patents included in Marx and Fuegi, 2020's database, as they represent the frontier of our analysis, given that they are at a null distance to science by directly citing papers. Further details on the construction of the distance metric will be provided in the following sections.

The first key step in our analysis was to merge the two datasets mentioned above. To avoid counting the same patent multiple times based on the countries where it was registered, we performed the merge at the DOCDB simple family patent level. The DOCDB is a collection of patent documents that are considered to cover a single invention in several countries. Hence, DOCDBs gather several patents covering a single invention into a unique identifier. We used parallel computing tools to carry out the merge operation due to the enormous size of the databases.

After the creation of a unified database, we selected the green patents belonging to CPC technology sectors Y02 and Y04, as it is common practice in the green literature (see, for example, Angelucci et al., 2018; Altenburg et al., 2020; Li et al., 2021). We excluded patents whose application filing year was prior to 1975. This is because the average number of patents realized in the years before 1975 is far below the average number of patents realized in the following years (see Fig.1). Note that the exclusion operation was performed at the patent level and not at the DOCDB level. We also exclude patents after 2019, since the decline in patent activity can be readily explained by incompleteness or lack of up-to-date information.

The resulting merged database includes 1.526.224 patent families (from now on simply patents) and their characteristics. The database consists, moreover, of 165.312 companies, 655.180 individuals, 4461 universities, and 273 hospitals in 223 countries.

Patent-to-paper links are also included. A patent has been linked to a paper based on whether one or more patents within a patent family were part of a PCS.

Finally, we constructed a variable listing the patents that a specific patent cited. Based on the latter variable we further construct other covariates, namely the percentage of green patents cited and the number of citing patents.

To summarize, the final database includes patent-level information such as the country of the inventor, the institution where the patent has been conceived, and the technological domain as reported by the CPC classification[3]. Furthermore, the database contains some patent-level covariates among which is the list of cited patents. The latter is key in constructing the distance measure.

The main analyses in the paper are conducted with USPTO patents only [4] to make them comparable with related literature (see e.g., Ahmadpoor and Jones, 2017 among others).

4 Methodology

The following section describes the distance metric and the main methodological concepts behind the analysis.

4.1 Distance metric

The construction of the distance metric is based on Ahmadpoor and Jones (2017) work. In their paper, Ahmadpoor and Jones (2017) propose a new distance measure between a patent and prior scientific advances in a given field, using patent citations to academic papers. The distance measure is based on the idea that patents that cite papers are closer to prior scientific advances than those that do not[5].

In formal terms, in Ahmadpoor and Jones (2017) a distance metric $D_i \in \{1,2,3,\ldots\}$ is defined for each patent (or paper) i. This metric is determined by recursively finding the minimum citation distance to the "patent-paper boundary". If a patent directly cites a paper, both nodes are assigned $D_i = 1$, indicating the "patent-paper boundary". If papers or patents cannot be connected to this boundary at any distance, they are considered "unconnected". A paper i with $D_i = n+1$ is one that is cited by a paper j with $D_j = n$ and is not cited by any paper k with $D_k < n$. Similarly, a patent i with $D_i = n+1$ is one that cites a patent j with $D_j = n$ and does not cite any patent k with $D_k < n$. It is important to

[3] Those variables were adopted in an expanded version of the database once the distance measure has been constructed

[4] The analysis was also conducted on overall world patents and are available upon request.

[5] Ahmadpoor and Jones (2017) define a distance metric among both patents and papers. For the scope of the present work, we only refer to the distance at the patent level

note that the graph is directed, with citations traced backward in time using references in each patent and paper, and jumping from the patent to the paper domain where $D_i = n$ [6].
Since this work focuses only on patents, we took the definition of distance for patents and applied a Breadth First Search (BFS) algorithm to to measure the citation distance of patents to the "patent-paper boundary" following the idea of Ahmadpoor and Jones, 2017. The approach proposed in Ahmadpoor and Jones (2017), indeed – though not applying a BFS directly – builds on this idea by recursively defining the distance metric as the minimum citation distance to the "patent-paper boundary" for each node.
In short, a BFS algorithm is a graph traversal algorithm that starts at a given node (a patent in our case) and explores all of its neighboring nodes at the current depth level before moving on to the next depth level. BFS is regarded an efficient algorithm that can efficiently explore and traverse a graph in a systematic and breadth-first manner (for more details see the Appendix).

The distance measure lies, by construction, in the range $\{0, +\infty\}$. A patent j has an ∞ distance if it is never reached by any other patent in terms of citations. In that case, j is said to be an un-connected component.
Figure 2 presents the share of (un-)connected components in our sample.

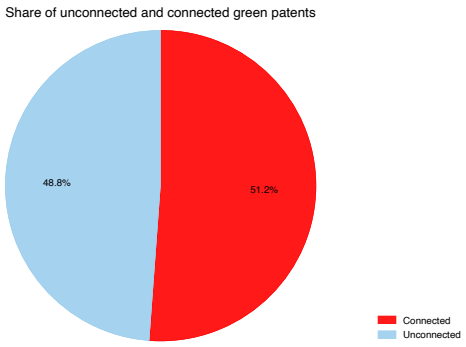


Figure 2. Share of connected components when only USPTO patents are kept on the frontier.

The share of connected and (un-)connected components can be also explored by the country of the inventor. This exercise is reproduced in Appendix D for only the countries where the majority of patents are produced according to Fig.6 below. The share of unconnected patents seems to be lower in Asian countries where, however, auto-referencing is strongly persistent as extensively discussed in the Appendix.
Following Ahmadpoor and Jones (2017), we checked the distribution of the distance variable in our sample as shown in Fig.3.

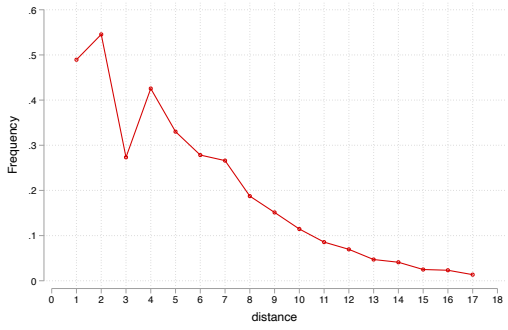


Figure 3. Distance distribution of connectivity. The graph shows a majority of green patents having $d = 2$.

As it is evident, the majority of patents have a distance of 2 from the patent frontier[7]. This is in line with Ahmad-

[6]Please refer to Ahmadpoor and Jones, 2017 for further details.
[7]This is the case also when considering the whole universe of green patents

poor and Jones (2017). Specifically, the percentage of green patents having a distance between 2 and 4 constitutes the 30% of the total sample of patents (connected and unconnected) and the 59% of the connected patents. In the same fashion, patents with a distance between 2 (excluded) and 4 constitutes the 16% of the total sample and the 32% of the connected components. Finally, patents with a distance between 5 and 7 constitute the 4.8% of the total sample and the 8% of the connected components.
Following Ahmadpoor and Jones (2017), we also compute the probability of a so-called "home run" defined as being in the upper 5% of citations received in that field and year. Consistently with Ahmadpoor and Jones (2017) we found that patents at a distance of 1 to the frontier are highly cited and appear as home-runs around 13% of the time, which is around the 62.5% of the background rate [8]. Other connected patents (i.e., those having $D_i \geq 2$) were home runs 6% of the time which is slightly lower than the background rate. Finally, disconnected components were found to be home-runs around 2.5% of the time. Fig.4 summarizes those results.

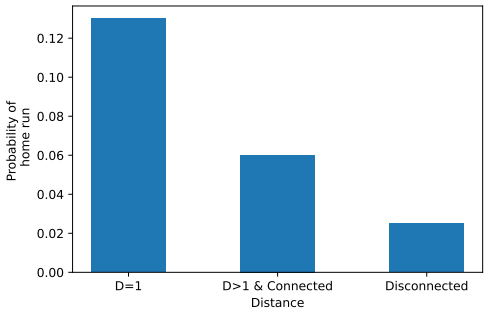


Figure 4. Probability of being home-runs by distance values.

The subsequent section will utilize the aforementioned distance measure to discuss the results.

5 Results

This section is organized into two distinct parts: the initial one presents descriptive outcomes concerning the determinants of science-based green patenting, employing the provided distance measure. Here, a patent's classification as "science-based" is determined by its proximity to the frontier. In the second part of the section, a more rigorous analysis utilizing machine learning tools (fully described in the Appendix) is provided, with the objective of identifying the specific features that characterize science-based patents.

5.1 Exploring the determinants of science-based green patents

One notable finding of our study is the use of a distance measure to evaluate the citation impact of green patents based on their proximity to the frontier. As demonstrated in Figure 5, which present the median and average number of citations for patents at different distances from the frontier, those closest to the frontier tend to receive more citations than those further away [9].

[8] here assumed to be the overall average

[9] Further analyses are available regarding the overall citation pattern (not limited to USPTO only).

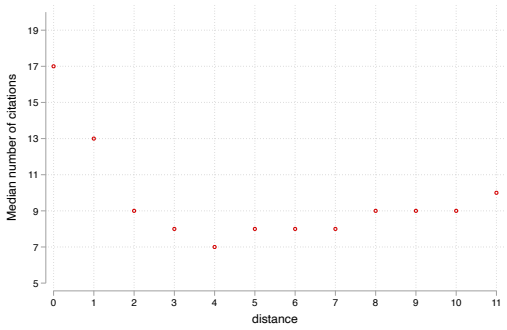


Figure 5. Citation pattern by distance (median).

An analysis of the relationship between the distance of green patents from the technological frontier and their citation impact is crucial to understand the determinants of scientific excellence and innovation in the field of environmental technology. By assessing this relationship, we can gain insights into the factors that contribute to being science-based – and hence closer to the frontier – and identify the mechanisms that drive the diffusion of knowledge in this domain. Yet, since we know that central patents in terms of citations are also those that are closer to the frontier, by grasping the features that lead a patent to be science-based we are characterizing also the most influential green patents. Such an understanding is essential in order to appreciate the importance of the descriptive studies that follow. Specifically, detailed statistics on the distance metric have been built starting from the inventor's country, the OECD fields and subfields, the technological class, and the institution owning the patent in an effort to grasp the features that identify science-based patents.

One such determinant is the inventor's country. As illustrated in Figure 6, the United States, Germany, Japan, France, the United Kingdom, China, and Korea exhibit higher numbers of green patents, with a majority of these patents being situated between a distance of 2 and 3 from the frontier[10]. While these findings confirm the existence of leaders in the field, they also highlight the presence of followers, such as Austria, Italy, and Belgium, who may strategically rely on the green patents of these leaders. The results are confirmed by the literature (see e.g., Li et al., 2021). As an example, Corrocher and Mancusi (2021) report that the United States, Japan, and Germany are the top three countries in terms of total green energy patents and also have the most collaborations with other countries in that technological domain. The paper also notes that China is rapidly catching up and has been increasing its green patenting activities in recent years.

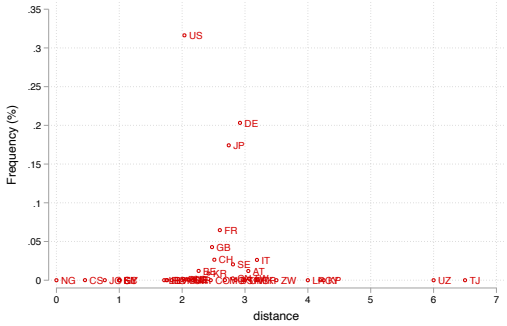


Figure 6. Distance distribution by inventor country. The graph shows a majority of green patents being patented in the US, Japan Germany, and Korea.

After establishing the leaders in green patenting, a natural question arises regarding the scientific basis of green patents across different areas. Figure 7 displays the median distance to the scientific frontier for green patents citing frontier patents (i.e. patents at a distance of 1), classified according to the so called Field of Study (FOS) of the paper

[10]The relative distances among these countries is lower when considering also non-USPTO patents with China catching-up the U.S. and Germany.

that they cite [11].
In order to achieve our objective, we grouped the frontier patents according to their FOS, hence moving the frontier of a step forward (indeed, grouping the frontier patents into the FOS(s) of the papers they cite moves the distance of green patents of a unit toward the frontier). Hence, the resulting distance of green patents citing frontier green patents lies now between 0 and 1.
We then grouped the patents that were originally at a distance 1 to the patent frontier according to their FOS and re-computed the distance measure at the FOS level.
By construction, *"fields of study (FOS) are organised in a multi-level hierarchy where parent research areas are fine-grained and multiple. The last MAKG version provides a descriptive classification of fields of study based on abstracts of publications"* (Pollacci, 2022). FOS therefore are referred to papers and are structured in macro areas (fields) and granular areas (sub-fields). According to Pollacci (2022) *"most of the FOS (macro areas) refer to the so-called STEM disciplines, thus science, technology, engineering and mathematics and any subjects that fall under these four disciplines as computer science (CS), biology, and chemistry. Conversely, humanities-related disciplines such as history, art, and philosophy seem characterised by fewer FOS"*.
The results indicate that Engineering and Natural Sciences are the predominant fields of green patents citing the frontier patents, with the latter exhibiting a relatively higher average distance than other fields. This finding may be attributed to the smaller number of green patents in the Humanities, Medical and Health Sciences, and Agricultural Sciences categories.

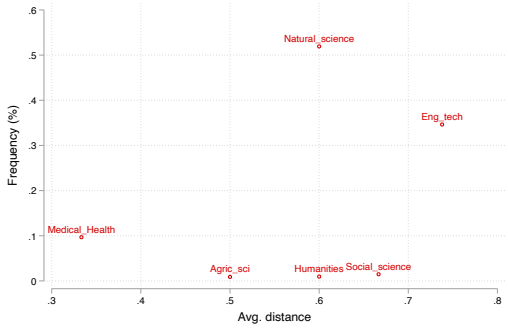


Figure 7. Distance distribution by FOS fields . The graph shows the average distance and frequency of the FOS fields among green patents.

The same exercise can be repeated with FOS sub-fields rather than FOS fields[12]. Again, the sub-fields refer to the paper cited by the frontier patent. Patents are then attributed such a sub-field. The mentioned procedure has been done in Figure 8. Similarly to Figure 7, Figure 8 reveals a prevalence of FOS sub-fields such as Chemical Sciences, Electronic Engineer and Materials Engineer confirming the scarcity of green patents with FOS in Arts and Health Sciences. It is worth noting that the absence of green patents in certain fields does not necessarily imply a lack of potential for innovation in those areas. There may be various reasons for the limited number of green patents, such as the level of public and private funding or the regulatory and policy environment.
The empirical evidence presented in Figure 8 underscores the variation in green patenting across different technological fields, suggesting the need for technology-specific sustainability policies, as advocated by Söderholm (2020). Our study extends the analysis of green patenting to specific technological sub-fields across a wide range of countries, making it the first of its kind. This provides a broader perspective on the role of various fields in green energy innovation. Notably, the dearth of green patents in the Arts and Health Sciences fields highlights potential research and innovation gaps in these areas pertaining to green energy technologies.

[11] Such fields are part of the macro-classification provided by Pollacci (2022).
[12] Such sub-fields are again provided in the EMAKG database

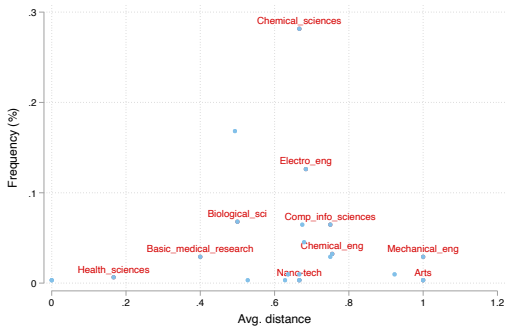


Figure 8. Distance distribution by FOS subfields . The graph shows the average distance and frequency of the FOS subfields among green patents.

Figure 9 puts the emphasis on CPC technological classes [13] with a specific focus on the Y CPC class.
Figure 9 displays the technological green classes according to their median distance and the percentage of connected components within a class among all the patents belonging to that class. CPC class Y represents "emerging cross-disciplinary technologies" and includes a broad range of subcategories such as "Y02" for technologies or applications for mitigation or adaptation against climate change, "Y04" for information or communication technologies having an impact on other technology areas, and "Y10" for emerging technologies not elsewhere classified.
Specifically, the technological classes located in the top-left quadrant of Figure 9 correspond to a higher concentration of connected components with a lower median distance from the frontier. Conversely, the technological classes located in the bottom-right quadrant include unconnected patents with a higher distance from the frontier. In other words, class $Y02P70/521$ includes connected green patents with a median distance of 1 to the frontier, whereas $Y10T16/466$ has almost 51% of connected components and a median distance of its green patents of 12. Notice that the median distance has been considered rather than the average distance due to the presence of unconnected components in the analysis.

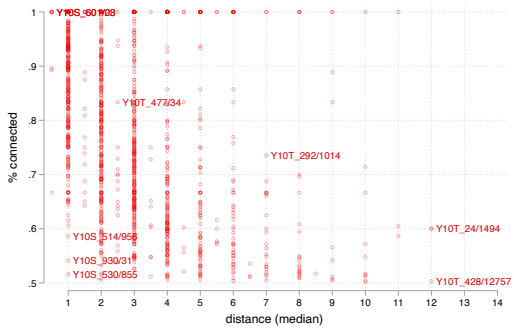


Figure 9. Median distance and percentage of connected components by technological class

The central CPC classes in green patenting are Y02P, Y10S, Y02T and Y02E which are, on average, closer to the frontier and have patents that are more connected to other green patents. These classes include technologies related to the production or processing of goods, as well as to transport. This is supported by evidence from several studies, including Nomaler, Verspagen, et al. (2021), which, in an effort to catch technological trajectories in the green landscape, show that these classes have remained consistently central to green patenting over the years. Despite the growing interest in the Y sub-classes as a means to assess the centrality of green patenting, the literature on this topic remains limited. To the best of our knowledge, few studies have provided empirical evidence on the predominance of one Y sub-class over another in terms of centrality. Moreover, the available empirical evidence has either led to contrasting conclusions (Barbieri et al., 2022) or has highlighted the challenges that the literature faces in fully capturing the centrality of green classes (Higham et al., 2022).

[13] The full set of technological classes is available at `https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html`

Figures 10 illustrates the distribution of technological green classes based on their median distance and the proportion of patents belonging to technological class *i* within the overall set of green patents. The results corroborate the findings of Figure 9 highlighting also the importance of class Y02 in terms of abundance with respect to other green classes at a low distance. It is important to note that the results presented in Figure 9 are not adjusted for the presence of unconnected components within technological classes. Specifically, Figure 10 do not take into account the proportion of unconnected components in both high-frequency and low-frequency classes[14].

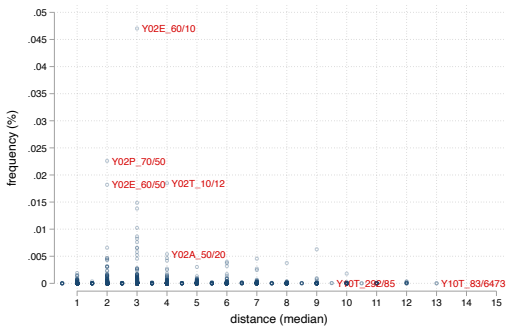


Figure 10. Median distance and percentage of patents by technological class

Figure 11 extends the previous analysis by examining the full range of density, as opposed to focusing solely on medians. To enhance the clarity of the plot, we selected a subset of technological classes for presentation purposes. The plots show a higher presence of low-distance patents within each technological class presented. This is plausible as the selected classes belong to either engineering or natural science fields.

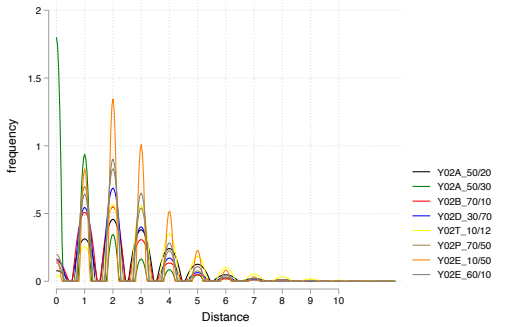


Figure 11. Distribution of distances among the connected patents .

The final part of the descriptive analysis examines the institutional determinants of science-based green patents. We consider five types of institutions, namely private organizations, individuals, public institutions, hospitals, and universities. Figure 12 presents the median distance of green patents for each organization type, accounting for patents with infinite distances in the calculations. Figure 13 repeats the same analysis without including patents with infinite distances.

[14]It should be noted that, since we compute the median, the proportion of unconnected components in low-distance classes is assumed to be less than 50%.

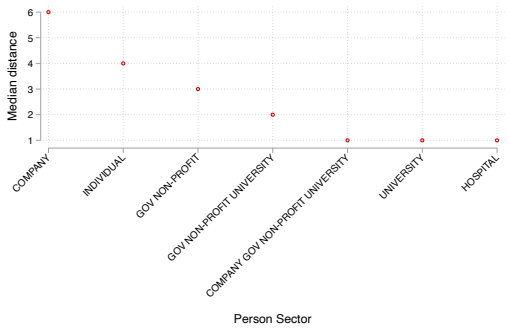


Figure 12. Median distance by institution when considering infinite distances .

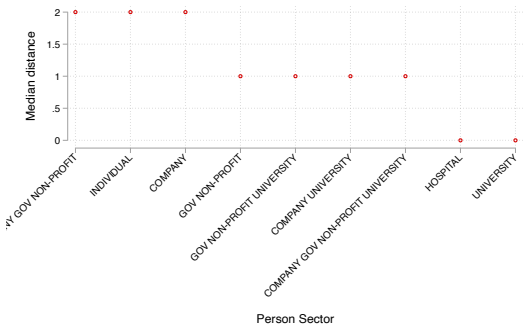


Figure 13. Median distance by institution when not considering infinite distances .

Our findings indicate that patents originating from hospitals, universities, and companies affiliated with universities tend to exhibit a greater degree of science-based patents. In contrast, green patents obtained by individuals and companies exhibit a larger deviation from the frontier (in line with Popp, 2019). These outcomes remain consistent even when company-related and university-related organizations are categorized respectively as "COMPANY" and "UNIVERSITY"[15].

The distribution of patents across institutions is presented in Figure 14. The pie chart demonstrates that firms hold the largest proportion of green patents at 64.8%, followed by private individuals at 31.7%, universities at 1.6%, and government institutions at 1.9%. It is interesting to notice that when including non-USPTO patents the difference in the share of patents held by firms and individual narrows with a percentage of green patents for firms at 49.2% and a percentage detained by individuals at 45.5%. This evidence can be explained in several manners. For instance, there might be entry barriers for individuals willing to submit their patents to the USPTO (e.g. need for external financing or strict rules for patenting). Or, simply, individuals from non-US countries tend to patent in the office of their country of origin, which, however, marks the U.S. individuals as outliers due to the disproportion among individuals green patenting and firm patenting in the country.

Overall the above results suggest that organizations with greater financial resources are more likely to generate green patents compared to private individuals. However, the production of such patents depends on the presence of science-based green patents held by universities and other entities.

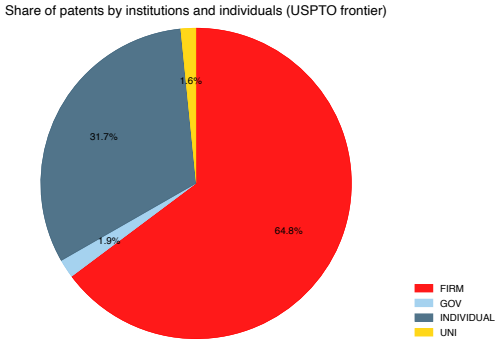


Figure 14. Share of green patents among institutions . The graph shows a higher presence of green patents among companies, followed by individuals.

Finally, Figure 15 illustrates the distribution of distances between green patents across institutions. Despite producing fewer green patents, universities tend to generate more science-based patents than other institutions, as expected. Universities tend to produce more science-based patents. Notably, firms tend to adopt green technologies that build upon existing knowledge, as consistently observed across both analyses.

[15]The results are available upon request.

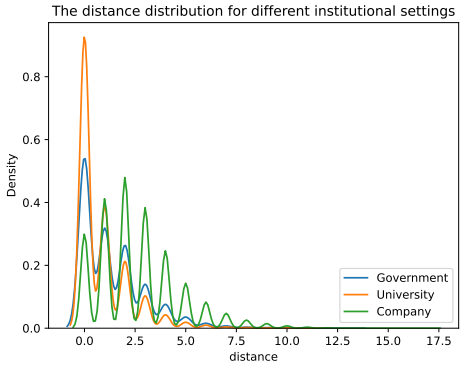


Figure 15. Density of distance by institution.

Prior to conducting the Machine-Learning analyses, it is important to provide a summary of the main descriptive findings thus far. This section emphasizes the significance of examining the determinants of science-based green patents, which serve as central nodes in the citation pattern. Our analysis reveals that the leading countries in green patenting are the United States, Germany, Japan, France, the United Kingdom, China, and Korea, with a median distance to the frontier of their patents ranging between 2 and 3. In terms of paper's FOS (macro)fields, the most prolific areas in green patents are Natural Science and Engineering, while the least prolific areas are the Arts and Humanities (as described by Fig.7). With regards to technological classes, our findings show a predominance of Y02P, Y02T, Y10S and Y02E classes in terms of both frequency and percentage of connected components within patents classified under these classes. This evidence is partially supported by the presence of start up clusters in these CPC technological field which, according to Marra et al. (2017), foster science-based green production. Further evidence supporting the findings lie on the known complexity of the fields of the production chain of the Y02P, Y02T, Y10S and Y02E classes (as shown in Balland and Boschma, 2022).
In conclusion, science-based green patents are primarily generated within universities, public institutions, and companies associated with universities. However, private companies and individuals possess a greater share of green patents in terms of quantity. We believe that this is due to the fact that the former can rely on substantial financing from stakeholders, while the latter benefit from greater flexibility and encounter less bureaucracy compared to organizations.

5.2 Assessing the determinants of science-based green patents

The current study employs machine learning techniques to carry out an analysis on five databases, i.e. the original database denoted as Overall, which contains Y and X variables and its partitions into four decades (1975-1985, 1986-1996, 1997-2007, and 2008-2020) to observe changes in relevant feature selection over time. Additionally, to address potential issues with independence between the databases, we conducted a robustness check by including an analysis of the Overall dataset. Regarding the dataset used for the machine learning analysis – and with respect to the dataset employed for the descriptive analyses – it included technology, country, and year dummies. These variables were included to account for possible fixed effects and to allow for their selection by the machine learning models as determinants X. The outcome variable Y is represented by a binary indicator that measures the degree of science-based attributes of a DOCDB family. Specifically, Y takes on a value of 1 if the majority of green patents within a DOCDB family have a distance metric less than or equal to 2. Additionally, we performed robustness checks that yielded similar results when the distance metric was relaxed to be less than or equal to 3, 4, and 5, respectively.
As aforementioned, Section 5.1 provided a descriptive summary of the key features indicative of a science-based green patent, which included inventor countries such as the United States, Korea, Japan, France, the UK, and Germany; a high number of citations by other patents; belonging to specific classes (Y02A 50/30, Y02P 70/521, Y02D 30/70, Y02T 30/00 and Y02E 20/30); and being a patent from a University, hospital or government-affiliated non-profit organization linked to a university.
To formally test whether these features are sufficient to describe a science-based green patent, we first examined which machine-learning classifiers in Tab. Appx.4 (see Appendix) performed better in predicting science-based patents. We then listed and interpreted the features selected by the best-performing classifiers each decade and overall, which are considered more important for the prediction task. A complete list of the variables selected in each decade and for the

overall sample is provided in Tab. Appx.5 of the Appendix. Specifically, since RF displays a better performance as compared to other machine-learning models (see Table Appx.4), the features are selected using RF. In particular, the Boruta algorithm (described in the Appendix) is utilized to compute the importance metric that contributes the most to the prediction of science-based features.

Each feature's importance measures its contribution to a machine learning model's predictive performance. In RF, the feature importance is estimated by the decrease in impurity measure resulting from splitting on a feature. The algorithm builds decision trees using different random subsets of features, and the feature importance scores are aggregated over all trees in the forest. These scores can be used for feature selection and gaining insights into the data-generating process[16].

A ranking of the variables according to the RF classifier is then made (see Fig. Appx.2) accounting at the same time for the number of times that the variable is selected by a model and the average importance attributed to such a variable (i.e. the information contained in Fig. Appx.2). Specifically, as a pre-processing step, we just considered the variables that have been selected by at least 3 models out of 5. Secondly, an index, I_p is constructed by multiplying the number of models that select the remaining variables from step one by their average importance. The resulting rank is reported in Tab.1 and represents the $\tilde{X} \subset X$ variables.

Ranking	Feature	Number of models	Average Importance	I_p
1^{st}	% of green cited	5	21.39	106.95
2^{nd}	Number citing	5	20.54	102.70
3^{rd}	US	5	17.44	87.20
4^{th}	University	5	16.73	83.65
5^{th}	JP	5	14.92	74.60
6^{th}	Y02E 70	5	14.40	72.00
7^{th}	DE	5	13.10	65.50
8^{th}	GB	5	11.72	58.60
9^{th}	Y02E 10	4	13.26	53.04
10^{th}	Y02B 10	4	12.20	48.80
11^{th}	FR	4	11.51	46.04
12^{th}	KR	5	8.46	42.30
13^{th}	Y02E 50	4	9.86	39.44
14^{th}	Gov.	3	12.34	37.02
15^{th}	Y02B 90	3	11.14	33.42
16^{th}	Y02T 30	4	8.15	32.60
17^{th}	Y02D 30	3	10.53	31.59
18^{th}	Y02P 80	3	9.65	28.95
19^{th}	Y04S 30	5	5.45	27.25
20^{th}	Y02B 40	3	8.57	25.71
21^{st}	Y02P 70	4	6.32	25.28
22^{nd}	Y04S 10	3	7.93	23.79
23^{rd}	NL	3	7.93	23.79
24^{th}	CA	3	7.30	21.90
25^{th}	Y02B 50	3	6.90	20.70

Table 1. First 25 features ranked by importance in determining how much a patent is science-based according to the index.

The countries identified through machine learning are observed to be the same as those identified by descriptive statistics. Notably, being affiliated with a university or government institution emerges as a common factor. In terms of sectors, the descriptive analysis identifies Y02A 50/30, Y02P 70/521, and Y02D 30/70 as significant. However, among these, only Y02D 30 and Y02P 70 are highlighted as relatively important through the machine learning analysis, but not to the same extent as indicated by the descriptive analysis.

[16]The reader is referred to the Appendix for a more detailed explanation

The machine learning analysis reveals a surprising and crucial finding, which differs from the descriptive analysis: the percentage of green patents cited by the patent plays a significant role in determining whether a green patent is science-based. This finding has important implications, as it suggests that the citation behavior of green patents can provide insights into their scientific value. Specifically, examining the percentage of green and non-green patents cited by a green patent could help identify the science-based patents and distinguish them from other green patents.
This result's significance is rooted in the literature, as the role of citation analysis in assessing the scientific and technical quality of patents has been extensively studied. Several studies, indeed, have shown that patent citations can be used as a measure of technological importance and that highly cited patents tend to be more valuable and more likely to be licensed or litigated (Jaffe et al., 2005). In the context of green patents, however, the role of citations in determining the scientific basis of a patent has not been fully explored. Recent studies have shown that green patents are more likely to cite non-patent literature, such as scientific articles and technical reports, than traditional patents (see e.g., Chai et al., 2020). Building on these findings, our results empirically demonstrate that the number of green patents cited by a green patent is a critical factor in determining its scientific basis. Specifically, we found that green patents that cite a higher number of other green patents are more likely to be science-based. This finding has important implications for the evaluation and management of green technology portfolios.
To ensure the robustness of our findings and examine the dynamics of the selected variables over time, Figure 16 has been constructed as follows. The importance of each selected variable ranked in Tab.1 was averaged over the number of iterations of the Boruta algorithm in each decade. A normalization step followed in order to characterize the weights of each variable for each decade. The weighted average for each decade is then reported in Figure 16.

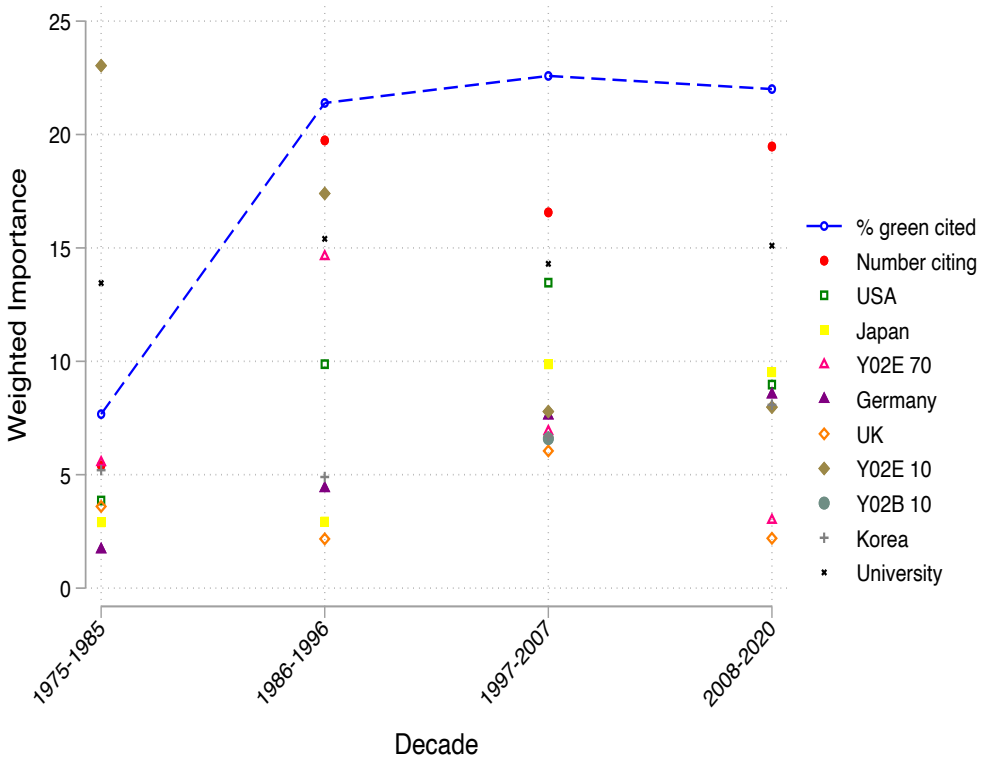


Figure 16. Dynamics of the first 11 selected variables from Tab.1

Based on Figure 16, it is clear that certain top-ranking features, such as the classes Y02E 10, Y02E 70, and the country UK, have experienced a decrease in their average importance over time, whereas the majority have either shown constant growth over time, as seen in the country-variables Korea, Germany, and Japan or remained stable at their initial level of importance, as observed in the variable University. Notably, the percentage of green citations has consistently outperformed other features in terms of importance, which confirms the findings of Table 1, even when

considering the temporal dimension. This observation is critical as it sheds light on the significance of the expanding literature that compares brown technologies to pure green technologies, as noted in works such as Heal (2007) and Skinner and Valentine (2023), among others. According to our results, while being classified as the former category does not guarantee any level of scientific basis or citation centrality, being part of the latter category is strongly correlated with scientific validity. In other words, being part of the "green network" guarantees a scientific basis for the patent.
Other notable results concern emerging countries in green patenting such as Korea and Japan which seemed to have almost caught up with traditionally prolific countries in terms of green patenting such as the USA.

6 Discussion and Conclusions

The present article makes two main contributions to the literature on green innovation. First, it offers a definition of being "science-based" in the context of green patents. Second, having established that science-based green patents are central in the network of patent citations, it provides both a descriptive and a formal approach to identifying the features that characterize science-based patents.
Starting from a unique database obtained by merging the PATSTAT database and Marx and Fuegi (2020)'s database, we have operationally defined a green patent as being science-based using a distance measure from a "patent-paper" frontier, as described by Ahmadpoor and Jones (2017). Specifically, we consider a patent i to be science-based if its distance from the frontier, denoted as D_i, is equal to $n+1$, where n is the distance of a patent j cited by i from the frontier and there are no other patents cited by i that are closer to the frontier than j. This distance measure allows for the possibility of unconnected components having a distance of infinity. We have calculated this distance measure using the breadth-first search (BFS) algorithm.
By defining this distance measure, we were able to establish a relationship between the distance of a green patent from the frontier and its citation path, finding an inverse relationship between the two.
Using the previously defined distance measure, we employed descriptive statistics to characterize science-based green patents based on the country of the inventor (i.e., the birthplace of knowledge), the OECD (sub)-fields, the technological classes, and the institution responsible for inventing the green patent. Our findings indicate a preponderance of green patents originating from the United States, Germany, Japan, France, the United Kingdom, China, and Korea. On average, the distance of these patents from the frontier ranged from 2 to 3, indicating a continued reliance on past innovations. The top OECD fields in terms of green patenting are, as expected, Natural Science and Engineering. We have identified the central CPC classes in green patenting as Y02P, Y10S, and Y02T. These classes are closer to the frontier and have patents that are more closely connected to other green patents.
In addition, our descriptive analysis reveals that green patents belonging to firms and individuals tend to be part of larger connected components, whereas those belonging to universities and public institutions tend to be more science-based. These findings suggest that companies, having greater financial resources, are able to produce more green patents than individuals. However, the production of such patents ultimately depends on the presence of science-based green patents held by universities and other entities.
Subsequently, we conducted a formal analysis utilizing machine learning techniques to provide statistical support for the descriptive findings. The machine learning classifiers identified similar features to those observed in the descriptive analysis, with a few notable exceptions. Specifically, the machine learning analysis identified the percentage of green patents cited as a top-ranked and robust determinant for a patent to be considered science-based. This result is a significant finding, as it suggests that the citation behavior of green patents can offer insights into their scientific worth. Motivated by this discovery, we investigated the dynamics of the top-ranked features selected by the machine-learning algorithm. Our findings indicate that the percentage of green patents cited is a more stable and robust determinant of scientific value than the other features we considered. This underscores the importance of pursuing "pure green" innovations, as opposed to "brown" innovations (Heal, 2007), in order to be considered science-based and therefore central in the green citation network. Importantly, the dynamics plot highlights the emergence of Korea and Japan among the key green patent producers.
In conclusion, our study has provided a novel definition of being "science-based" in the context of green patents and identified the key features that characterize science-based green patents. Our findings underscore the importance of pursuing pure green innovations, as opposed to brown innovations, in order to be considered science-based and central in the green citation network.
Our study is situated in the broader literature on green innovation, which has shown that green innovation is critical

for achieving sustainability goals and addressing climate change (e.g., Neufeldt et al., 2021; Kivimaa and Kern, 2016; Tukker et al., 2016). Moreover, past research has found that green patents are more likely to be cited and to have a broader impact than non-green patents (e.g., Kemp et al., 2019; Li and van't Veld, 2015) and tried to find out more about the determinants of the most cited green patents (Chai et al., 2020). However, our study is unique in its focus on science-based green patents and its use of a distance measure to operationalize this concept. To the best of our knowledge, ours is the first empirical attempt to provide a definition of science-based green patents and to determine the features of such green patents.

The novelty and contribution of our findings are further underscored by their relevance for policymakers and industry practitioners. For instance, our finding that the percentage of green patents cited is a top-ranked determinant of scientific value can inform patent examination practices, as well as research and development efforts aimed at developing science-based green technologies. Moreover, our identification of the central CPC classes in green patenting (Y02P, Y10S, and Y02T) can inform technology transfer policies and investment decisions aimed at promoting green innovation in these areas. Ultimately, in a more general way, our results are in line with past studies that have shown the positive impact of green innovation on economic growth, environmental sustainability, and social welfare (Kemp et al., 2019)

Appendix

A Auto-citation patterns

One of the main reasons why we conducted the analyses on USPTO patents only is the presence of biases arising from auto-citation patterns. Tab.Appx.1 presents the percentage of patents that belong to the inventor and/or applicant's country, denoted as P_c, and cite at least one patent within the same country, denoted as CP_c. The second column repeats this exercise by considering the application authority instead of the country of the inventor and/or applicant, with P_{app} and CP_{app} representing P_c and CP_c, respectively when the country of the inventor and/or applicant is taken into account.

Table Appx.1. Share of patents citing at least one patent within the same country for a sample of countries.

Country	P_c/CP_c	P_{app}/CP_{app}
AU	.248	.0512
BR	.1655	.0315
CA	.3125	.0604
CN	**.446**	**.2466**
DE	.5089	.1457
DK	.146	.0332
EP	.423	.0920
ES	.193	.0427
FR	.282	.0754
GB	.278	.0810
IT	.1635	.0293
JP	.4902	.2528
KR	.32	.1364
NL	.147	.0427
NO	.139	.0293
NZ	.096	.0194
SE	.184	.0333
US	.95	.2444
WO	.339	.0709

The evidence from Table Appx.1 clearly indicates that China stands out with a much higher share of patent self-citation when both P_c/CP_c and P_{app}/CP_{app} are considered together. A similar analysis is conducted for patents on the frontier to examine the proportion of domestically authored papers cited by science-based patents worldwide. Notably, 45% of China's patents on the frontier cited papers written by Chinese authors, surpassing the average self-citation rate of approximately 32%. These findings overall enforce the choice of USPTO patents only over the entire sample of world patents.

B The BFS algorithm

Given a graph $G = (V,E)$, where V is the set of vertices (nodes) and E is the set of edges, the BFS algorithm starts at a given source vertex $s \in V$ and explores all its neighboring vertices at the current depth level before moving on to the next depth level. The algorithm maintains a queue Q of vertices to be visited, and a set S of visited vertices to avoid visiting them again. Initially, Q contains only the source vertex s, and S is empty. At each iteration, the algorithm dequeues a vertex v from Q, marks it as visited by adding it to S, and explores its neighboring vertices that have not been visited yet. For each neighboring vertex w of v that is not in S, the algorithm adds it to Q and marks its distance $d(w)$ from the source vertex as $d(w) = d(v) + 1$, where $d(v)$ is the distance of v from the source vertex. The algorithm stops when all vertices reachable from the source vertex have been visited, or when Q becomes empty.

The BFS algorithm can be easily visualized as a tree rooted at the source vertex, where each vertex is connected to its parent vertex by an edge of the tree, and its children's vertices are the neighboring vertices discovered by the algorithm at the current depth level. Figure Appx.1 [17] shows an example of a BFS tree starting from vertex A in a directed graph, where the edges are labeled with their weights. The vertices are visited in the order A, B, C, D, E, F, G, and H, and their distances from A are 0, 1, 1, 2, 2, 2, 3, and 3, respectively. The BFS algorithm has the property that the distances computed by it are the minimum distances from the source vertex to each reachable vertex in the graph, which makes it useful for many applications such as shortest path, connected components, and network analysis.

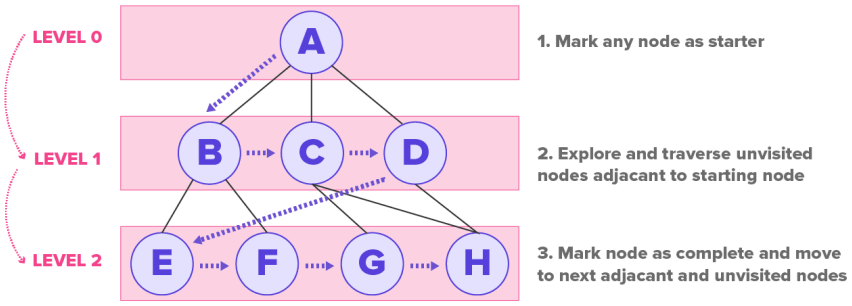


Figure Appx.1. An example of a BFS tree starting from the vertex A in a directed graph, where the edges are labeled with their weights. The vertices are visited in the order A, B, C, D, E, F, G, and H, and their distances from A are 0, 1, 1, 2, 2, 2, 3, and 3, respectively.

C Average distance by CPC class

The following table shows the average distance of CPC classes. It displays how the green classes classified in Y have, on average, a lower distance to the frontier than other CPC classes.

[17] *The BFS graph image used in this paper is adapted from the Hackr.io website, licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license.*

CPC	Super-class	Average CPC class	Average CPC super-class
A01B	A	2.145	
A01C	A	2.484	
A01D	A	2.004	
A01F	A	1.730	
A01G	A	2.877	
A01H	A	3.966	
A01J	A	1.705	
A01K	A	2.729	
A01L	A	1.781	
A01M	A	2.495	
A01N	A	2.202	
A21B	A	1.943	
A21C	A	1.996	
A21D	A	2.474	
A22B	A	1.941	
A22C	A	2.010	
A23B	A	2.399	
A23C	A	2.292	
A23D	A	2.266	
A23F	A	2.563	
A23G	A	2.090	
A23J	A	2.123	
A23K	A	2.390	
A23L	A	2.726	
A23N	A	2.523	
A23P	A	2.394	
A23V	A	3.729	
A23Y	A	3.004	
A24B	A	2.614	
A24C	A	1.708	
A24D	A	2.090	
A24F	A	2.729	
A41B	A	2.977	
A41C	A	2.751	
A41D	A	2.616	
A41F	A	2.444	
A41G	A	3.115	
A41H	A	2.028	
A42B	A	2.318	
A42C	A	2.269	
A43B	A	2.106	
A43C	A	2.039	
A43D	A	2.113	**2.350**
A44B	A	2.386	
A44C	A	2.233	
A44D	A	2.261	
A45B	A	2.449	
A45C	A	2.654	
A45D	A	2.327	
A45F	A	2.719	
A46B	A	2.044	
A46D	A	1.773	
A47B	A	2.271	
A47C	A	2.328	
A47D	A	2.516	
A47F	A	2.272	
A47G	A	2.503	
A47H	A	2.095	
A47J	A	2.261	
A47K	A	2.302	
A47L	A	2.179	
A61B	A	2.202	
A61C	A	2.076	
A61D	A	2.205	
A61F	A	2.047	
A61G	A	2.281	
A61H	A	2.800	
A61J	A	2.438	
A61K	A	2.258	
A61L	A	2.329	
A61M	A	2.099	
A61N	A	2.192	
A61P	A	2.197	
A61Q	A	2.056	
A62B	A	2.251	
A62C	A	2.559	
A62D	A	2.452	
A63B	A	2.497	
A63C	A	1.672	
A63D	A	2.091	
A63F	A	2.741	
A63G	A	2.975	
A63H	A	2.409	
A63J	A	2.513	
A63K	A	1.960	
B01B	B	1.898	
B01D	B	2.264	
B01F	B	2.199	
B01J	B	2.403	
B01L	B	2.224	**2.208**
B02B	B	3.486	
B02C	B	2.681	
B03B	B	2.130	
B03C	B	2.303	

CPC	Super-class	Average CPC class	Average CPC super-class
B03D	B	2.503	
B04B	B	1.802	
B04C	B	1.824	
B05B	B	2.218	
B05C	B	2.503	
B05D	B	2.639	
B06B	B	2.463	
B07B	B	2.361	
B07C	B	2.236	
B08B	B	3.012	
B09B	B	2.337	
B09C	B	3.026	
B21B	B	2.082	
B21C	B	2.032	
B21D	B	2.324	
B21F	B	2.268	
B21G	B	1.220	
B21H	B	1.849	
B21J	B	2.264	
B21K	B	2.421	
B21L	B	1.893	
B22C	B	2.454	
B22D	B	2.184	
B22F	B	2.882	
B23B	B	2.163	
B23C	B	2.305	
B23D	B	2.045	
B23F	B	1.942	
B23G	B	2.176	
B23H	B	2.510	
B23K	B	2.547	
B23P	B	2.800	
B23Q	B	2.333	
B24B	B	2.682	
B24C	B	2.281	
B24D	B	2.386	
B25B	B	2.379	
B25C	B	1.920	
B25D	B	1.697	
B25F	B	2.218	
B25G	B	2.458	
B25H	B	2.552	
B25J	B	2.862	
B26B	B	2.006	
B26D	B	2.251	
B26F	B	2.121	
B27B	B	2.085	**2.208**
B27C	B	2.493	
B27D	B	1.952	
B27F	B	2.004	
B27G	B	1.892	
B27H	B	1.191	
B27J	B	2.097	
B27K	B	2.768	
B27L	B	2.187	
B27M	B	2.008	
B27N	B	2.071	
B28B	B	2.383	
B28C	B	2.444	
B28D	B	2.544	
B29B	B	2.160	
B29C	B	2.219	
B29D	B	2.337	
B29K	B	2.375	
B29L	B	2.396	
B30B	B	1.987	
B31B	B	2.052	
B31C	B	1.420	
B31D	B	2.140	
B31F	B	1.939	
B32B	B	2.711	
B33Y	B	2.801	
B41B	B	0.687	
B41C	B	2.067	
B41D	B	1.500	
B41F	B	1.681	
B41G	B	0.806	
B41J	B	2.629	
B41K	B	2.246	
B41L	B	2.426	
B41M	B	2.588	
B41N	B	1.880	
B41P	B	1.776	
B42B	B	2.296	
B42C	B	2.191	
B42D	B	1.998	
B42F	B	1.840	
B42P	B	1.400	
B43K	B	2.348	
B43L	B	2.364	
B43M	B	1.929	
B44B	B	1.749	
B44C	B	2.112	
B44D	B	2.054	

CPC	Super-class	Average CPC class	Average CPC super-class
B44F	B	1.708	
B60B	B	2.102	
B60C	B	2.079	
B60D	B	2.233	
B60F	B	2.592	
B60G	B	1.896	
B60H	B	2.137	
B60J	B	1.863	
B60K	B	2.294	
B60L	B	2.793	
B60M	B	2.021	
B60N	B	1.895	
B60P	B	2.347	
B60Q	B	2.261	
B60R	B	2.104	
B60S	B	1.852	
B60T	B	1.823	
B60V	B	1.574	
B60W	B	2.521	
B60Y	B	3.318	
B61B	B	2.156	
B61C	B	2.326	
B61D	B	2.048	
B61F	B	2.031	
B61G	B	2.122	
B61H	B	2.300	
B61J	B	0.989	
B61K	B	1.988	
B61L	B	2.355	
B62B	B	2.288	
B62C	B	0.930	
B62D	B	2.160	
B62H	B	2.586	
B62J	B	2.788	**2.208**
B62K	B	2.473	
B62L	B	2.435	
B62M	B	2.298	
B63B	B	2.462	
B63C	B	2.453	
B63G	B	2.415	
B63H	B	2.406	
B63J	B	2.613	
B64B	B	2.136	
B64C	B	2.389	
B64D	B	2.418	
B64F	B	2.607	
B64G	B	2.339	
B65B	B	2.142	
B65C	B	2.372	
B65D	B	2.020	
B65F	B	2.377	
B65G	B	2.054	
B65H	B	2.055	
B66B	B	2.326	
B66C	B	2.105	
B66D	B	2.265	
B66F	B	2.257	
B67B	B	2.012	
B67C	B	1.970	
B67D	B	2.096	
B68B	B	1.788	
B68C	B	1.852	
B68F	B	1.750	
B68G	B	1.943	
B81B	B	3.075	
B81C	B	3.116	
B82B	B	2.820	
B82Y	B	3.042	
C01B	C	2.647	
C01C	C	1.983	
C01D	C	2.769	
C01F	C	2.341	
C01G	C	2.913	
C01P	C	2.609	
C02F	C	2.766	
C03B	C	2.196	
C03C	C	2.325	
C04B	C	2.720	
C05B	C	3.795	
C05C	C	2.472	
C05D	C	2.542	
C05F	C	2.407	
C05G	C	3.629	**2.380**
C06B	C	2.194	
C06C	C	2.290	
C06D	C	2.105	
C06F	C	1.516	
C07B	C	2.122	
C07C	C	2.104	
C07D	C	2.178	
C07F	C	2.407	
C07G	C	2.291	
C07H	C	2.234	
C07J	C	2.102	
C07K	C	2.485	
C08B	C	2.312	
C08C	C	2.462	

CPC	Super-class	Average CPC class	Average CPC super-class
C08F	C	2.278	
C08G	C	2.379	
C08H	C	2.574	
C08J	C	2.888	
C08K	C	2.949	
C08L	C	2.846	
C09B	C	1.916	
C09C	C	2.083	
C09D	C	2.755	
C09F	C	2.494	
C09G	C	3.611	
C09H	C	1.240	
C09J	C	2.856	
C09K	C	2.621	
C10B	C	2.229	
C10C	C	2.743	
C10F	C	0.821	
C10G	C	2.425	
C10H	C	0.667	
C10J	C	2.186	
C10K	C	2.363	
C10L	C	2.357	
C10M	C	2.381	
C10N	C	2.415	
C11B	C	2.011	
C11C	C	2.317	
C11D	C	1.866	
C12C	C	1.865	
C12F	C	1.620	
C12G	C	2.296	**2.380**
C12H	C	1.937	
C12J	C	2.228	
C12L	C	0.571	
C12M	C	2.497	
C12N	C	2.571	
C12P	C	2.683	
C12Q	C	2.652	
C12R	C	3.048	
C12Y	C	3.189	
C13B	C	1.764	
C13K	C	2.233	
C14B	C	1.797	
C14C	C	2.045	
C21B	C	2.056	
C21C	C	2.099	
C21D	C	3.061	
C22B	C	2.454	
C22C	C	3.130	
C22F	C	3.360	
C23C	C	2.610	
C23D	C	2.012	
C23F	C	2.651	
C23G	C	2.383	
C25B	C	2.710	
C25C	C	2.154	
C25D	C	2.713	
C25F	C	2.564	
C30B	C	2.758	
C40B	C	2.534	
D01B	D	1.954	
D01C	D	1.652	
D01D	D	3.146	
D01F	D	3.101	
D01G	D	1.545	
D01H	D	1.554	
D02G	D	2.562	
D02H	D	1.459	
D02J	D	1.984	
D03C	D	1.365	
D03D	D	2.190	
D03J	D	1.616	
D04B	D	1.959	
D04C	D	2.216	
D04D	D	2.232	
D04G	D	2.370	
D04H	D	2.294	
D05B	D	2.234	
D05C	D	2.743	
D05D	D	1.952	**2.124**
D06B	D	1.978	
D06C	D	2.022	
D06F	D	2.207	
D06G	D	2.241	
D06H	D	1.992	
D06J	D	0.935	
D06L	D	1.692	
D06M	D	2.786	
D06N	D	2.711	
D06P	D	2.370	
D06Q	D	2.145	
D07B	D	2.032	
D10B	D	2.740	
D21B	D	2.118	
D21C	D	3.175	
D21D	D	1.852	
D21F	D	1.577	
D21G	D	1.553	

CPC	Super-class	Average CPC class	Average CPC super-class
D21H	D	2.150	**2.124**
D21J	D	2.562	
E01B	E	1.837	
E01C	E	2.327	
E01D	E	2.924	
E01F	E	2.122	
E01H	E	2.452	
E02B	E	2.377	
E02C	E	1.989	
E02D	E	2.591	
E02F	E	2.566	
E03B	E	2.272	
E03C	E	2.166	
E03D	E	2.415	
E03F	E	2.347	
E04B	E	2.130	
E04C	E	2.071	
E04D	E	1.981	**2.211**
E04F	E	2.122	
E04G	E	2.170	
E04H	E	2.443	
E05B	E	1.952	
E05C	E	1.883	
E05D	E	1.920	
E05F	E	2.061	
E05G	E	1.842	
E05Y	E	1.934	
E06B	E	1.974	
E06C	E	2.201	
E21B	E	2.244	
E21C	E	2.072	
E21D	E	2.285	
E21F	E	2.860	
F01B	F	1.983	
F01C	F	2.395	
F01D	F	2.069	
F01K	F	2.337	
F01L	F	1.952	
F01M	F	2.087	
F01N	F	2.057	
F01P	F	2.134	
F02B	F	2.061	
F02C	F	2.320	
F02D	F	2.105	
F02F	F	1.946	
F02G	F	2.283	
F02K	F	2.215	
F02M	F	2.019	
F02N	F	2.371	
F02P	F	2.123	
F03B	F	2.355	
F03C	F	1.685	
F03D	F	2.209	
F03G	F	2.763	
F03H	F	2.351	
F04B	F	2.298	
F04C	F	2.524	
F04D	F	2.312	
F04F	F	2.353	
F05B	F	2.166	
F05C	F	1.776	
F05D	F	2.521	
F15B	F	2.153	
F15C	F	1.518	
F15D	F	2.261	
F16B	F	1.959	**2.118**
F16C	F	2.050	
F16D	F	1.871	
F16F	F	1.967	
F16G	F	2.031	
F16H	F	2.093	
F16J	F	1.914	
F16K	F	2.209	
F16L	F	1.965	
F16M	F	2.700	
F16N	F	2.206	
F16P	F	1.734	
F16S	F	0.0833	
F16T	F	1.794	
F17B	F	1.150	
F17C	F	2.229	
F17D	F	2.751	
F21H	F	0.667	
F21K	F	3.109	
F21L	F	2.807	
F21S	F	2.576	
F21V	F	2.765	
F21W	F	2.900	
F21Y	F	2.787	
F22B	F	2.103	
F22D	F	1.410	
F22G	F	1.594	
F23B	F	2.170	
F23C	F	2.073	
F23D	F	2.001	
F23G	F	2.357	
F23H	F	1.347	

CPC	Super-class	Average CPC class	Average CPC super-class
F23J	F	2.174	
F23K	F	1.969	
F23L	F	2.051	
F23M	F	1.330	
F23N	F	2.054	
F23Q	F	2.000	
F23R	F	1.937	
F24B	F	2.040	
F24C	F	2.176	
F24D	F	2.051	
F24F	F	3.273	
F24H	F	2.198	
F24S	F	2.238	
F24T	F	2.638	
F24V	F	2.358	
F25B	F	2.763	
F25C	F	2.950	
F25D	F	2.722	**2.118**
F25J	F	1.823	
F26B	F	2.191	
F27B	F	1.904	
F27D	F	1.923	
F27M	F	0.291	
F28B	F	1.853	
F28C	F	1.940	
F28D	F	2.244	
F28F	F	2.218	
F28G	F	2.022	
F41A	F	2.053	
F41B	F	2.546	
F41C	F	2.754	
F41F	F	1.918	
F41G	F	2.077	
F41H	F	2.258	
F41J	F	2.158	
F42B	F	1.826	
F42C	F	1.598	
F42D	F	2.432	
G01B	G	2.436	
G01C	G	2.664	
G01D	G	2.170	
G01F	G	1.955	
G01G	G	2.208	
G01H	G	2.668	
G01J	G	2.571	
G01K	G	2.337	
G01L	G	2.287	
G01M	G	2.731	
G01N	G	2.480	
G01P	G	2.088	
G01Q	G	2.332	
G01R	G	2.615	
G01S	G	2.490	
G01T	G	2.432	
G01V	G	2.284	
G01W	G	3.987	
G02B	G	2.556	
G02C	G	2.294	
G02F	G	2.863	
G03B	G	3.308	
G03C	G	2.145	
G03D	G	1.763	
G03F	G	2.598	
G03G	G	3.255	
G03H	G	2.488	
G04B	G	1.901	
G04C	G	1.999	
G04D	G	1.931	**2.469**
G04F	G	3.069	
G04G	G	2.426	
G04R	G	2.340	
G05B	G	2.876	
G05D	G	2.675	
G05F	G	2.827	
G05G	G	2.006	
G06C	G	0.250	
G06E	G	2.377	
G06F	G	2.702	
G06G	G	2.346	
G06J	G	0.547	
G06K	G	3.132	
G06M	G	1.489	
G06N	G	4.399	
G06Q	G	2.984	
G06T	G	3.020	
G07B	G	2.345	
G07C	G	2.438	
G07D	G	2.137	
G07F	G	2.208	
G07G	G	2.848	
G08B	G	2.685	
G08C	G	2.929	
G08G	G	2.902	
G09B	G	2.991	
G09C	G	2.941	
G09D	G	1.025	
G09F	G	2.606	

CPC	Super-class	Average CPC class	Average CPC super-class
G09G	G	2.807	
G10B	G	0.857	
G10C	G	2.530	
G10D	G	2.587	
G10F	G	2.529	
G10G	G	2.880	
G10H	G	2.913	
G10K	G	2.101	
G10L	G	2.524	
G11B	G	2.546	
G11C	G	2.689	
G12B	G	0.426	
G16B	G	8.228	**2.469**
G16C	G	2.632	
G16H	G	3.090	
G16Z	G	2.113	
G21B	G	2.879	
G21C	G	1.973	
G21D	G	3.065	
G21F	G	1.965	
G21G	G	2.396	
G21H	G	3.761	
G21J	G	2.24	
G21K	G	2.588	
G21Y	G	2.28	
H01B	H	2.883	
H01C	H	2.355	
H01F	H	2.607	
H01G	H	3.142	
H01H	H	2.043	
H01J	H	2.333	
H01K	H	2.029	
H01L	H	2.644	
H01M	H	2.768	
H01P	H	2.603	
H01Q	H	2.282	
H01R	H	2.295	
H01S	H	2.445	
H01T	H	2.258	
H02B	H	1.935	
H02G	H	2.207	
H02H	H	2.554	
H02J	H	2.812	
H02K	H	2.265	
H02M	H	2.577	
H02N	H	2.914	
H02P	H	2.586	
H02S	H	3.142	
H03B	H	2.704	
H03C	H	1.989	**2.480**
H03D	H	2.136	
H03F	H	2.413	
H03G	H	2.505	
H03H	H	2.553	
H03J	H	2.186	
H03K	H	2.609	
H03L	H	2.483	
H03M	H	2.630	
H04B	H	2.572	
H04H	H	2.650	
H04J	H	2.664	
H04K	H	2.919	
H04L	H	2.599	
H04M	H	2.533	
H04N	H	2.640	
H04Q	H	2.439	
H04R	H	2.411	
H04S	H	2.508	
H04W	H	2.565	
H05B	H	2.461	
H05C	H	2.016	
H05F	H	1.673	
H05G	H	2.110	
H05H	H	2.665	
H05K	H	2.705	
Y02A	Y	2.907	
Y02B	Y	2.839	
Y02C	Y	0.733	
Y02D	Y	2.447	
Y02E	Y	2.848	
Y02P	Y	2.020	**2.188**
Y02T	Y	2.491	
Y02W	Y	2.741	
Y04S	Y	2.253	
Y10S	Y	0.449	
Y10T	Y	2.344	

For the sake of clarity, Fig.Appx.2 has been included below representing the average distances for a subset of CPC classes of the Table above. As it is evident from the picture, the green classes included in the CPC super-class Y have a lower average distance.

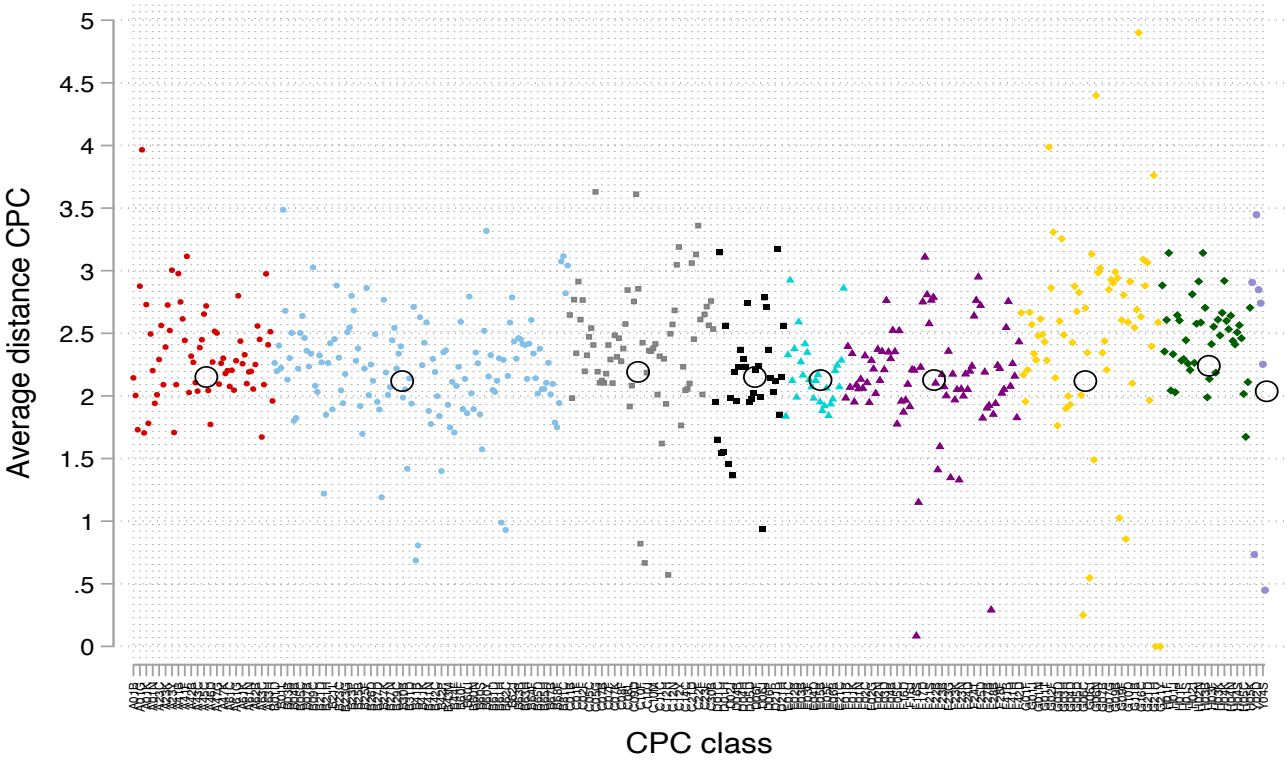
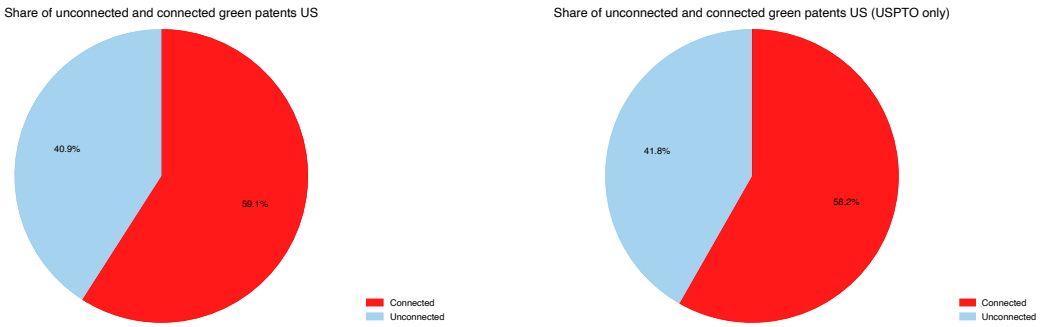


Figure Appx.2. Sub-sample of average distance by CPC class and super-class (big circles) The figure shows the average distance of a subset of CPC sub-classes present in the dataset. The empty circle represent the overall average of the CPC super-classes.

D (Un-)Connected components by inventor country

The present section displays the share of connected and unconnected components by the inventor country. The inventor country is chosen rather than the applicant one as we wanted to focus on the place where the knowledge developed.



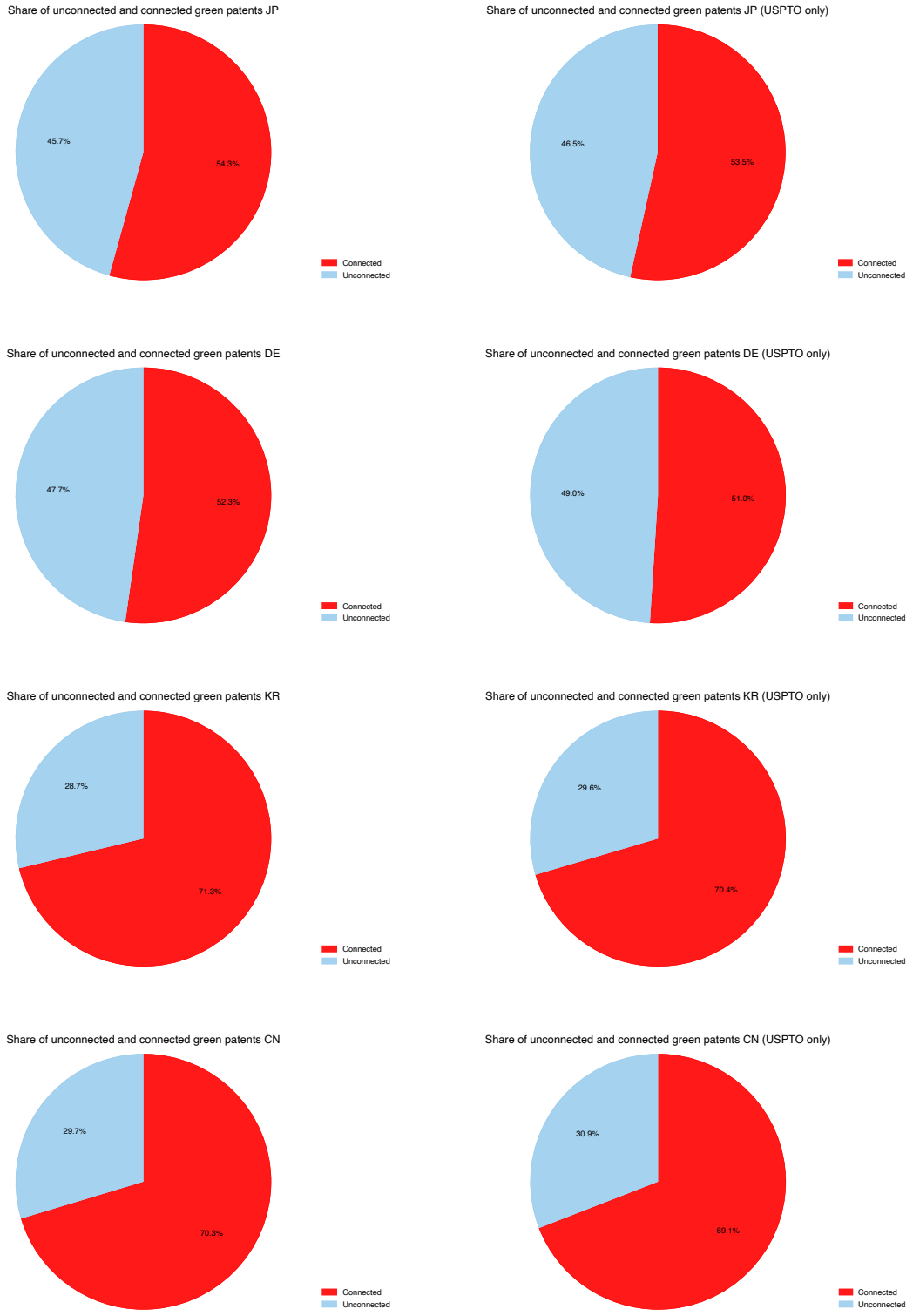


Figure Appx.1. Connected and unconnected components by inventor country considering the whole universe of patents and only those in the USPTO.

E CPC sectors legend

The CPC classification is a classification of patents that improves the older version IPC with a novel class Y. The first letter of the classification is the section symbol (e.g. the letter Y stems from emerging cross-sectional technologies). This is followed by a two-digit number to give a "class symbol" (e.g., "A01" represents "Agriculture; forestry; animal husbandry; trapping; fishing"). The final letter makes up the "subclass". The subclass is then followed by a 1- to 3-digit "group" number, an oblique stroke, and a number of at least two digits representing a "main group" ("00") or "subgroup". In the present paper, we adopted the classification until the main group.
Specifically, the analysis comprises the CPC classes Y02 and Y04 and their sub-categories. A complete and detailed overview of the CPC classes included is available at `https://www.uspto.gov/web/patents/classification/cpc/html/cpc-Y10S.html#Y10S`. Below we report two tables for Y02 (Tab. Appx.2) and Y04 Tab. Appx.3) classes respectively with the adopted classification symbols:

CPC class	CPC subclass	Description
Y02A **TECHNOLOGIES FOR ADAPTATION TO CLIMATE CHANGE**	Y02A 10	at coastal zones; at river basins
	Y02A 20	Water conservation ;Efficient water supply; Efficient water use
	Y02A 30	Adapting or protecting infrastructure or their operation
	Y02A 40	Adaptation technologies in agriculture, forestry, livestock or agroalimentary production
	Y02A 50	in human health protection, e.g. against extreme weather
	Y02A 90	Technologies having an indirect contribution to adaptation to climate change
Y02B **CLIMATE CHANGE MITIGATION TECHNOLOGIES RELATED TO BUILDINGS, e.g. HOUSING, HOUSE APPLIANCES OR RELATED END-USER APPLICATIONS**	Y02B 10	Integration of renewable energy sources in buildings
	Y02B 20	Energy efficient lighting technologies, e.g. halogen lamps or gas discharge lamps
	Y02B 30	Energy efficient heating, ventilation or air conditioning
	Y02B 40	Technologies aiming at improving the efficiency of home appliances, e.g. induction cooking or efficient technologies for refrigerators, freezers or dish washers
	Y02B 50	Energy efficient technologies in elevators, escalators and moving walkways, e.g. energy saving or recuperation technologies
	Y02B 70	Technologies for an efficient end-user side electric power management and consumption
	Y02B 80	Architectural or constructional elements improving the thermal performance of buildings
	Y02B 90	Enabling technologies or technologies with a potential or indirect contribution to GHG emissions mitigation
Y02C	Y02C 20	Capture or disposal of greenhouse gases
Y02D **CLIMATE CHANGE MITIGATION TECHNOLOGIES IN INFORMATION AND COMMUNICATION TECHNOLOGIES [ICT]**	Y02D 10	Energy efficient computing, e.g. low power processors, power management or thermal management
	Y02D 30	Energy efficient computing, e.g. low power processors, power management or thermal management
Y02E **REDUCTION OF GREENHOUSE GAS [GHG] EMISSIONS, RELATED TO ENERGY GENERATION, TRANSMISSION OR DISTRIBUTION**	Y02E 10	Energy generation through renewable energy sources
	Y02E 20	Combustion technologies with mitigation potential
	Y02E 30	Energy generation of nuclear origin
	Y02E 40	Technologies for an efficient electrical power generation, transmission or distribution
	Y02E 50	Technologies for the production of fuel of non-fossil origin
	Y02E 60	Enabling technologies; Technologies with a potential or indirect contribution to GHG emissions mitigation
	Y02E 70	Other energy conversion or management systems reducing GHG emissions
Y02P **CLIMATE CHANGE MITIGATION TECHNOLOGIES IN THE PRODUCTION OR PROCESSING OF GOODS**	Y02P 10	Technologies related to metal processing
	Y02P 20	Technologies relating to chemical industry
	Y02P 30	Technologies relating to oil refining and petrochemical industry
	Y02P 40	Technologies relating to the processing of minerals
	Y02P 60	Technologies relating to agriculture, livestock or agroalimentary industries
	Y02P 70	Climate change mitigation technologies in the production process for final industrial or consumer products
	Y02P 80	Climate change mitigation technologies for sector-wide applications
	Y02P 90	Enabling technologies with a potential contribution to greenhouse gas [GHG] emissions mitigation
Y02T **CLIMATE CHANGE MITIGATION TECHNOLOGIES RELATED TO TRANSPORTATION**	Y02T 10	Road transport of goods or passengers
	Y02T 30	Transportation of goods or passengers via railways, e.g. energy recovery or reducing air resistance
	Y02T 50	Aeronautics or air transport
	Y02T 70	Maritime or waterways transport
	Y02T 90	Enabling technologies or technologies with a potential or indirect contribution to GHG emissions mitigation
Y02W **CLIMATE CHANGE MITIGATION TECHNOLOGIES RELATED TO WASTEWATER TREATMENT OR WASTE MANAGEMENT**	Y02W 10	Technologies for wastewater treatment
	Y02W 30	Technologies for solid waste management
	Y02W 90	Enabling technologies or technologies with a potential or indirect contribution to greenhouse gas [GHG] emissions mitigation

Table Appx.2. Description of Y02 CPC class.

CPC class	CPC subclass	Description
Y04S **SYSTEMS INTEGRATING TECHNOLOGIES RELATED TO POWER NETWORK OPERATION, COMMUNICATION OR INFORMATION TECHNOLOGIES FOR IMPROVING THE ELECTRICAL POWER GENERATION, TRANSMISSION, DISTRIBUTION, MANAGEMENT OR USAGE, i.e. SMART GRIDS**	Y04S 10	Systems supporting electrical power generation, transmission or distribution
	Y04S 20	Management or operation of end-user stationary applications or the last stages of power distribution; Controlling, monitoring or operating thereof
	Y04S 30	Systems supporting specific end-user applications in the sector of transportation
	Y04S 40	Systems for electrical power generation, transmission, distribution or end-user application management characterised by the use of communication or information technologies, or communication or information technology specific aspects supporting them
	Y04S 50	Market activities related to the operation of systems integrating technologies related to power network operation or related to communication or information technologies

Table Appx.3. Description of Y04 CPC classes.

F Boruta Algorithm

Let X be a feature matrix with n observations and p features, and Y be the target vector. The Boruta algorithm works as follows:

Algorithm 1 Boruta with RF

1- Initialize the set of important features $Z = \{\}$.
2- Create m shadow features for each feature X_j, $(j = 1, 2, ..., p)$, where m is a user-specified parameter.
3- Train a Random Forest (RF) model on the original and shadow features. Let the feature importance score for feature X_j be given by the Gini impurity reduction, denoted by I_j.
4- For each feature X_j, calculate the mean importance score of its shadow features, denoted by S_j.
5- If $I_j > S_j$, add X_j to the set of important features Z.
6- Repeat steps 3-5 until all important features have been identified or all features have been considered.

The equivalent algorithm for XGBoost is as follows:

Algorithm 2 Boruta with XGBoost

1- Initialize the set of important features $Z = \{\}$. 2- Create m shadow features for each feature X_j, $(j = 1, 2, ..., p)$, where m is a user-specified parameter.
3- Train an XGBoost model on the original and shadow features. Let the feature importance score for feature X_j be given by the gain, denoted by G_j.
4-For each feature X_j, calculate the mean gain of its shadow features, denoted by T_j.
5- If $G_j > T_j$, add X_j to the set of important features Z.
6- Repeat steps 3-5 until all important features have been identified or all features have been considered.

In both algorithms, the feature importance score is used to rank the features, and the shadow features are used as a reference to determine if a feature is truly important or if its importance score is due to chance. The algorithm continues until either all important features have been identified or all features have been considered.

G Machine Learning algorithms

In this section we apply the machine-learning regularization techniques in order to select the characteristics that make a green patent science-based (i.e. with a distance below 2 from the frontier). In other words, we explore which are the features that mostly characterize science-based green patents. The latter exercise provides a statistical validation to the descriptive results obtained in the previous sections.

Regularization techniques are commonly used in machine learning to prevent overfitting, improve the generalization performance, and make the models more interpretable by selecting relevant features. In prediction problems, regularization can help in selecting features that are most relevant to the target variable. The two most commonly used regularization techniques are L1 (Lasso), L2 (Ridge) regularization (Hastie et al., 2009) and ElasticNet which mixes the two mentioned forms of regularization.

L1 regularization adds a penalty term to the loss function, which is proportional to the absolute value of the coefficients, and is given by the following equation (Tibshirani, 1996):

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

where β_j is the j-th coefficient, p is the number of features, and λ is the regularization parameter. This encourages the coefficients to be sparse, i.e., only a few features are selected as important.

L2 regularization, on the other hand, adds a penalty term proportional to the square of the coefficients, and is given by the following equation (Hoerl and Kennard, 1970):

$$\lambda \sum_{j=1}^{p} \beta_j^2$$

This encourages the coefficients to be small and discourages multicollinearity.

The use of regularization techniques in machine learning has several advantages. Firstly, it helps to prevent overfitting, which is a common problem in machine learning. Overfitting occurs when the model is too complex and fits the training data too well, resulting in poor performance on new, unseen data (Bishop and Nasrabadi, 2006). Regularization techniques help to simplify the model by penalizing complex models, encouraging the model to fit the training data in a more general way.

Secondly, regularization can also help in feature selection. In many prediction problems, there may be a large number of features that are not relevant to the target variable. Regularization techniques can automatically identify the most important features and exclude the others, resulting in a more interpretable model (Hastie et al., 2009). Finally, regularization can also improve the generalization performance of the model. This means that the model is better able to make accurate predictions on new, unseen data (Boyd & Vandenberghe, 2004).

In the present work we exploit the last two mentioned properties of regularization techniques. Specifically, our problem is one of classification where the dependent variable, $Y_i \in \{0,1\}$, is dicotomic. Y_i represents the fact that a patent i is science-based or not, i.e. whether the patent is at a distance lower than 2 to the frontier. Following Athey and Imbens (2019) we performed an horse-ride among various machine-learning technique represented in Tab.Appx.4. We then confront their performances among each other and with a classical econometric method, namely a Logit Model. The techniques comprise LASSO, Ridge, ElasticNet, Random Forest (RF) and Extreme Gradient Boosting (XGBoost). To validate our analysis we included a McNemar test, which provides the presence of statistical differences between two machine-learning classifiers.

Notice that RF (as well as XGBoost) is not typically considered as a regularization technique in the traditional sense, but it can have some regularization-like properties. Its use is recommended in classification prediction problems (see, eg. Athey and Imbens, 2019 among others).

Random forest is an ensemble learning method that creates multiple decision trees and combines their predictions to obtain a final prediction (Breiman, 2001). By randomly selecting a subset of features at each split, RF can prevent overfitting to a certain extent, as the trees are less likely to become too complex. This can be seen as a form of implicit feature selection, as less important features are less likely to be selected at each split.

However, RF does not directly penalize the coefficients or impose any restrictions on the size of the coefficients, as traditional regularization methods such as L1 and L2 regularization do (see above). Instead, it relies on combining the predictions of many trees to reduce the variance and increase the stability of the predictions. Thus, while RF does not directly perform regularization, it can help to prevent overfitting and feature selection.

Similarly, XGBoost (Extreme Gradient Boosting) is a tree-based ensemble learning method that can be used for both regression and classification problems. Like RF, XGBoost creates multiple decision trees and combines their predictions to make a final prediction. In XGBoost, regularization can be achieved by adding a penalty term to the loss function during the training process. This penalty term is proportional to the magnitude of the coefficients, similar to L1 and L2 regularization. By adjusting the magnitude of the penalty term, the user can control the level of regularization in the model. For example, L1 regularization can be achieved in XGBoost by adding an L1 penalty term to the objective function, while L2 regularization can be achieved by adding an L2 penalty term.
In this way, XGBoost can be used as a regularization technique to prevent overfitting, improve the generalization performance, and select relevant features in prediction problems.
Another important property of RF an XGBoost is that they are immune to sparsity of the data. This is key in our prediction task as most of the features included (X) comprise dummy variable, and, specifically, yearly dummies and sectoral dummies [18]. The remaining features comprise the percentage of green cited by a patent, the average distance of the patents cited by the patent to the frontier and other patent-specific characteristics.
In order to prevent that the other methodologies adopted are biased by sparsity –and as a robustness check in the cases of RF and XGBoost– we performed a grid-search of the training and test sets. The latter consists of a cross-validation of the training and test sets operated several (1000 in our case) times in parallel. The performance indices of Tab. Appx.4 are then produced as a median, on the test sets, of the mentioned repetitions. In this way, we ensure that the features selected in the test set do not consist of all zeros (or ones).
As a further check against sparsity, we performed a dedicated analysis with pre-processed datasets using Principal Component Analysis[19].
We evaluated the described machine-learning classifiers according to the following metrics: the Area Under the Curve (AUC), the Balanced Accuracy (ACC), Matthews Correlation Coefficient (MCC) and the Precision-Recall AUC (PR-AUC). The choice of such metrics is motivated by the substantial imbalance of the class labels of the Y variable in which the zeros represent a 69% of the occurrences.
The F1-Score is the harmonic mean of precision and recall and is a commonly used performance metric in binary classification tasks. It is defined as:

$$F1 - Score = 2 \cdot \frac{(Precision * Recall)}{(Precision + Recall)}$$

where precision is the fraction of true positive predictions made out of all positive predictions and recall is the fraction of true positive predictions made out of all actual positive instances.
The Precision-Recall AUC (PR-AUC) is another commonly used performance metric, which focuses specifically on the precision and recall of a model. It is defined as the area under the curve of the precision-recall curve. Unlike the traditional receiver operating characteristic (ROC) curve, which is used to evaluate the false positive rate and true positive rate, the PR-AUC takes into account both the precision and recall of a model (Davis and Goadrich, 2006).
The Area Under the ROC Curve (AUC) is calculated as the area under the curve of the receiver operating characteristic (ROC) curve and is used to evaluate binary classification models. A higher AUC score indicates that the model has a higher ability to distinguish between positive and negative classes.
The Matthews Correlation Coefficient (MCC) is a measure of the balance between true positive, false positive, true negative, and false negative predictions and is particularly useful when the distribution of positive and negative classes is imbalanced. It is defined as follows:

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}}$$

where TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) are the number of instances in each classification. The MCC ranges from -1 to 1, with a value of 1 indicating perfect performance and a negative value indicating that the classifier is doing worse than random Powers (2020).
According to Powers (2020) among others in the literature, the F1-Score, PR-AUC, AUC, MCC are among the most commonly used and fair performance metrics for binary classification tasks, particularly in situations where there is class imbalance.

[18] Refer to Appendix E for a description of the sectors included.
[19] With the PCA we basically confirmed the results reached in terms of importance ranking for the first 10 positions of Tab. 1

H Assessing the performance of the machine-learning algorithms

The first step is referred to as a "*horse-ride*" between models, a widely used practice in adopting diverse machine-learning techniques (see e.g., Rácz et al., 2019 and Bargagli Stoffi, 2020 for a thorough review). Specifically, each database has been partitioned into a training set and a test set. The training set is employed to determine the optimal hyperparameters of each model, while the test set is made up of newly observed units that did not participate in the selection of hyperparameters. The test set is used to perform out-of-bag predictions on the dependent variable Y, and the models are then compared in terms of performance indices obtained on the test set (see e.g. Athey and Imbens, 2019, for a more detailed explanation). The performance indices are fully described in the Appendix and listed below. With respect to the models used to evaluate the predictions of the outcome variable, we selected a baseline model, Logit, and some traditional classifiers that employ regularization and internal cross-validation, specifically, Cross-Validated (CV) Least Absolute Shrinkage and Selection Operator (LASSO) and CV-Ridge. We also chose two non-linear classifiers that account for possible interactions among variables, Random Forest (RF), and Gradient Boosting, which is known as XGBoost, the most popular algorithm for implementing Gradient Boosting.
The idea, again, is to select a subset $\tilde{X} \subset X$ that mostly contributed to the prediction of Y (i.e. whether a green patent is science-based or not). Such a selection is performed either via regularization in the case of CV-LASSO and CV-Ridge or via feature-importance indices in the cases of RF and XGBoost. Regularization aims to minimize the sum of squared errors while constraining the sum of absolute values of coefficients to be less than a given constant, L1 norm for LASSO, and L2 norm for Ridge. LASSO forces some coefficients to be zero, performing feature selection. Ridge shrinks coefficients towards zero but does not force them to be exactly zero, making it useful for feature selection.

Table Appx.4 provides an evaluation of the performance of machine-learning classifiers based on four indicators: Area Under the Curve (AUC), Balanced Accuracy (ACC), Matthews Correlation Coefficient (MCC), and Precision-Recall AUC (PR-AUC). These indicators are explained in detail in the Appendix, and the Logit model serves as the benchmark for comparison. A McNeamar test is conducted to examine the similarity between pairs of models. RF exhibits superior performance compared to other models, as reflected in most of the performance indices presented in Table Appx.4, except for F1-score. These results significantly outperform the benchmark model, CV-LASSO, and CV-Ridge, which are highlighted in bold in the table. Therefore, the features are selected using RF and XGBoost, and the Boruta algorithm is utilized to compute the importance metric that contributes the most to the prediction of science-based features. Boruta compares the importance of each feature with its shadow feature and repeatedly calculates the feature importance score until all important features have been identified. A shadow feature consists of a random permutation of the values of the original feature and is used as a reference for determining the feature's importance. The Boruta algorithm repeatedly calculates the feature importance score for each feature and its shadow features until all the important features have been identified. More in detail, the algorithm starts by creating a set of shadow features for each original feature. Then, it trains a RF (XGBoost) model on the original and shadow features and calculates the feature importance score for each feature based on the out-of-bag (OOB) error of the model.
The algorithm accepts a feature if its importance score is higher than its shadow features, and it continues until all important features have been identified. A more formal description of the Boruta features selector can be found in Algorithms 1 and 2 of the Appendix. Boruta is applied to both RF and XGBoost. The selected covariates are presented in Tab. Appx.5 for each model.
Fig. Appx.2 displays a bar chart with the number of models that select the variable according to Tab. Appx.5 and the average importance attributed to the variable in such models.

	Logit					CV-LASSO					CV-Ridge					RF					XGBoost				
Period	**F1**	**AUC**	**ACC.**	**MCC**	**PR-AUC**	**F1**	**AUC**	**ACC.**	**MCC**	**PR-AUC**	**F1**	**AUC**	**ACC.**	**MCC**	**PR-AUC**	**F1**	**AUC**	**ACC.**	**MCC**	**PR-AUC**	**F1**	**AUC**	**ACC.**	**MCC**	**PR-AUC**
1975-1985	0.68	0.65	0.70	0.40	0.89	0.70	0.70	0.71	0.41	0.91	0.69	0.70	0.70	0.41	0.89	**0.77**	**0.87**	**0.79**	**0.56**	**0.92**	0.74	0.73	0.73	0.47	0.82
1986-1996	0.49	0.62	0.65	0.25	0.80	0.50	0.70	0.65	0.27	0.88	0.48	0.63	0.66	0.26	0.88	0.69	**0.84**	**0.77**	**0.51**	**0.91**	**0.77**	0.71	0.72	0.41	0.83
1997-2007	0.32	0.58	0.78	0.26	0.91	0.30	0.75	0.78	0.27	0.92	0.26	0.74	0.78	0.24	0.91	0.61	**0.87**	**0.84**	**0.52**	**0..93**	**0.88**	0.69	0.81	0.44	**0.93**
2008-2020	0.17	0.54	0.86	0.21	0.94	0.14	0.77	0.86	0.19	0.95	0.13	0.60	0.85	0.18	0.94	0.52	**0.84**	**0.88**	**0.41**	**0.95**	**0.90**	0.66	**0.88**	**0.43**	**0.95**
Overall sample	0.56	0.55	0.89	0.22	0.41	0.35	0.59	0.62	0.23	0.83	0.35	0.57	0.60	0.25	0.84	0.44	**0.82**	**0.82**	**0.40**	**0.92**	**0.67**	0.73	0.70	0.35	0.89
Year Dummies	Yes					Yes					Yes					Yes					Yes				
McNeamar Test	Logit-CV-LASSO	Passed				.	.				.	.				.	.				.	.			
	Logit-CV-Ridge	Not Passed				CV-LASSO-CV-Ridge	Not Passed				.	.				.	.				.	.			
	Logit-RF	Passed				CV-LASSO-RF	Passed				CV-Ridge-RF	Passed				.	.				.	.			
	Logit-XGBoost	Passed				CV-LASSO-XGBoost	Passed				CV-Ridge-XGBoost	Passed				RF-XGBoost	Not Passed				.	.			

Table Appx.4. The Table shows several precision indicators for Logit, CV-Lasso, CV-Ridge Classifier, Random Forest (500 trees), and XGBoost. The latter two have been inserted as they better deal with sparse data. In addition, a CV-ElasticNet has been tested but has been not included since it did not add much more than CV-LASSO and CV-Ridge. The metrics analyzed comprise F1-score, AUC, Balanced Accuracy, Matthew's coefficient, and Precision-Recall AUC. The latter performance indices are obtained as medians of the performance indices obtained by extracting $N = 1000$ different training and test sets. A Mcnemar's Test is added indicating no statistical difference between the (statistically) best models, i.e. Random Forest and XGBoost.

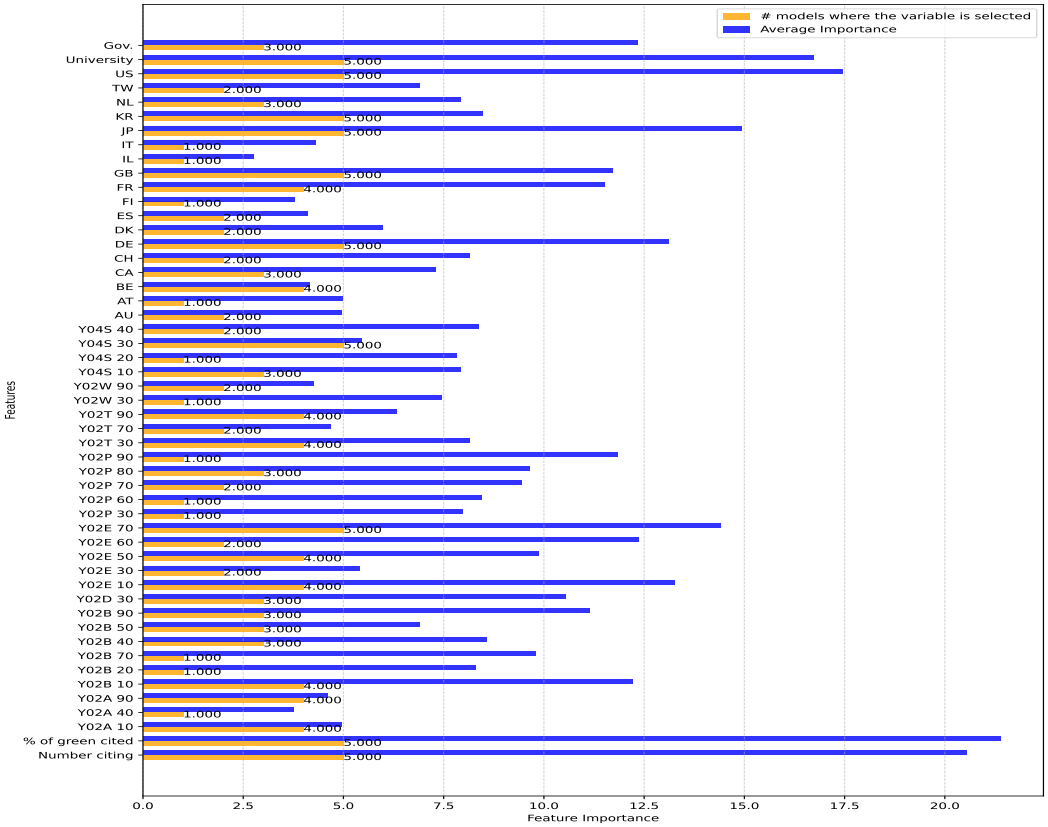


Figure Appx.2. Average importance of the selected features among the models which select them (x-axis and orange bars) and the number of models that select feature *i* according to Tab. Appx.5 (blue bars).

I Variable selection by model

The following table presents the variables selected by each model:

Model list				
1975-1985	1986-1996	1997-2007	2008-2020	Overall
Number citing	Number citing	Number citing	Number citing	Number citing
% of green cited	% of green cited	% of green cited	% of green cited	% of green cited
Y02A 10	Y02A 10	Y02A 10	Y02A 10	Y02A 40
Y02A 90	Y02B 10	Y02A 90	Y02A 90	Y02A 90
Y02B 10	Y02B 40	Y02B 10	Y02B 50	Y02B 10
Y02B 20	Y02B 50	Y02B 40	Y02E 10	Y02B 40
Y02B 70	Y02B 90	Y02B 50	Y02E 30	Y02B 90
Y02B 90	Y02D 30	Y02E 10	Y02D 30	Y02E 50
Y02D 30	Y02E 10	Y02E 60	Y02E 60	Y02E 60
Y02E 10	Y02E 50	Y02E 70	Y02E 70	Y02E 70
Y02E 30	Y02E 60	Y02P 30	Y02P 70	Y02P 60
Y02E 70	Y02E 70	Y02P 70	Y02T 90	Y02P 80
Y02P 80	Y02P 80	Y02T 30	Y04S 30	Y02T 30
Y02P 90	Y02T 30	Y02T 90	AU	Y02T 90
Y02T 30	Y02T 70	Y04S 30	BE	Y04S 10
Y02T 70	Y02T 90	AU	CA	Y04S 30
Y02W 90	Y04S 10	BE	DE	Y02W 30
Y04S 10	Y04S 20	CA	DK	Y02W 90
Y04S 30	Y04S 30	CH	ES	DE
Y04S 40	Y04S 40	DE	FR	GB
BE	AT	DK	GB	JP
CH	BE	ES	JP	KR
DE	CA	FI	KR	US
FR	CH	FR	NL	University
GB	DE	GB	US	Gov.
JP	FR	IL	University	.
KR	GB	IT	Gov.	.
NL	JP	JP	.	.
US	KR	KR	.	.
University	TW	LU	.	.
Gov.	US	NL	.	.
.	University	NZ		.
.	.	PT	.	.
.	.	TW	.	.
.	.	US	.	.
.	.	University	.	.

Table Appx.5. The above table collects the variables selected by the best machine-learning model as per the performance indices listed in Tab.Appx.4. In the Table we report the CPC indices whose references are listed in Appendix A.

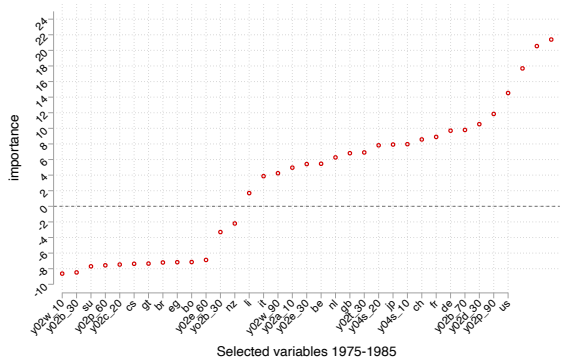


Figure Appx.3. Importance of a sub-sample of features in the period 1975-1985. Features with importance greater than 0 are selected.

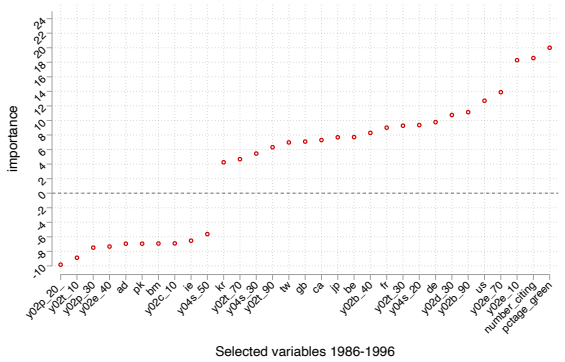


Figure Appx.4. Importance of a sub-sample of features in the period 1986-1996. Features with importance greater than 0 are selected.

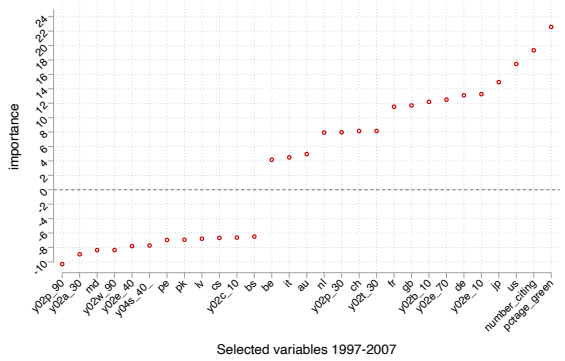


Figure Appx.5. Importance of a sub-sample of features in the period 1997-2007. Features with importance greater than 0 are selected.

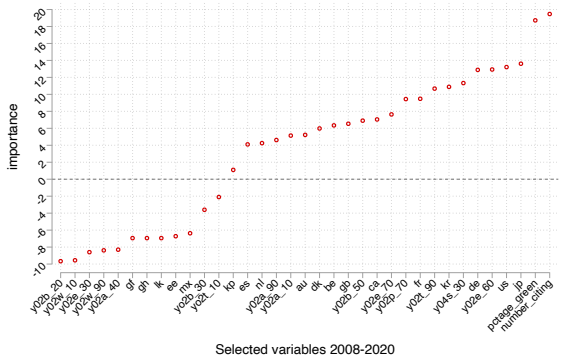


Figure Appx.6. Importance of a sub-sample of features in the period 2008-2020. Features with importance greater than 0 are selected.

Figure Appx.7. Features importance according to Random Forest and/or XGBoost.

References

Ahmadpoor, Mohammad and Benjamin F Jones (2017). "The dual frontier: Patented inventions and prior scientific advance". In: *Science* 357.6351, pp. 583–587.

Athey, Susan and Guido W Imbens (2019). "Machine learning methods that economists should know about". In: *Annual Review of Economics* 11, pp. 685–725.

Balland, Pierre-Alexandre and Ron Boschma (2022). "Do scientific capabilities in specific domains matter for technological diversification in European regions?" In: *Research Policy* 51.10, p. 104594.

Barbieri, Nicolò, Davide Consoli, Lorenzo Napolitano, François Perruchas, Emanuele Pugliese, and Angelica Sbardella (2022). "Regional technological capabilities and green opportunities in Europe". In: *The Journal of Technology Transfer*, pp. 1–30.

Bargagli Stoffi, Falco Johannes (2020). "Essays on applied machine learning". In.

Bishop, Christopher M and Nasser M Nasrabadi (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Chai, Kuang-Cheng, Yang Yang, Zhiyong Sui, and Ke-Chiun Chang (2020). "Determinants of highly-cited green patents: The perspective of network characteristics". In: *Plos one* 15.10, e0240679.

Corrocher, Nicoletta and Maria Luisa Mancusi (2021). "International collaborations in green energy technologies: What is the role of distance in environmental policy stringency?" In: *Energy Policy* 156, p. 112470.

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Heal, Geoffrey (2007). "Environmental accounting for ecosystems". In: *Ecological economics* 61.4, pp. 693–694.

Higham, Kyle, Martina Contisciani, and Caterina De Bacco (2022). "Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships". In: *Technological Forecasting and Social Change* 179, p. 121628.

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge regression: applications to nonorthogonal problems". In: *Technometrics* 12.1, pp. 69–82.

Jaffe, Adam B, Richard G Newell, and Robert N Stavins (2005). "A tale of two market failures: Technology and environmental policy". In: *Ecological economics* 54.2-3, pp. 164–174.

Kemp, René, Anthony Arundel, Christian Rammer, Michal Miedzinski, Carlos Tapia, Nicolò Barbieri, Serdar Türkeli, Andrea M Bassi, Massimiliano Mazzanti, Donald Chapman, et al. (2019). "Measuring eco-innovation for a Green economy". In: *Wirtsch Blätter, Special Issue on Nachhaltigkeit/Sustainability* 66.4, pp. 391–404.

Kivimaa, Paula and Florian Kern (2016). "Creative destruction or mere niche support? Innovation policy mixes for sustainability transitions". In: *Research policy* 45.1, pp. 205–217.

Li, Yaya, Yuru Zhang, Chien-Chiang Lee, and Jing Li (2021). "Structural characteristics and determinants of an international green technological collaboration network". In: *Journal of Cleaner Production* 324, p. 129258.

Li, Yuanhao and Klaas van't Veld (2015). "Green, greener, greenest: Eco-label gradation and competition". In: *Journal of environmental economics and management* 72, pp. 164–176.

Marra, Alessandro, Paola Antonelli, and Cesare Pozzi (2017). "Emerging green-tech specializations and clusters–A network analysis on technological innovation at the metropolitan level". In: *Renewable and Sustainable Energy Reviews* 67, pp. 1037–1046.

Marx, Matt and Aaron Fuegi (2020). "Reliance on science: Worldwide front-page patent citations to scientific articles". In: *Strategic Management Journal* 41.9, pp. 1572–1594.

Neufeldt, Henry, Lars Christiansen, and Thomas William Dale (2021). "Adaptation Gap Report 2020". In: *United Nations Environment Programme.*

Nomaler, Önder, Bart Verspagen, et al. (2021). *Patent Landscaping Using" green" Technological Trajectories*. Maastricht Economic and Social Research Institute on Innovation and ...

Pollacci, Laura (2022). "EMAKG: An Enhanced Version Of The Microsoft Academic Knowledge Graph". In: *arXiv preprint arXiv:2203.09159*.

Popp, David (2019). "Environmental policy and innovation: a decade of research". In.

Powers, David MW (2020). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061*.

Rácz, Anita, Dávid Bajusz, and Károly Héberger (2019). "Multi-level comparison of machine learning classifiers and their performance metrics". In: *Molecules* 24.15, p. 2811.

Skinner, A Nicole and Kristen Valentine (2023). "Green Patenting and Voluntary Innovation Disclosure". In: *Available at SSRN 4321932*.

Söderholm, Patrik (2020). "The green economy transition: the challenges of technological change for sustainability". In: *Sustainable Earth* 3.1, pp. 1–11.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tukker, Arnold, Tanya Bulavskaya, Stefan Giljum, Arjan de Koning, Stephan Lutter, Moana Simas, Konstantin Stadler, and Richard Wood (2016). "Environmental and resource footprints in a global context: Europe's structural deficit in resource endowments". In: *Global Environmental Change* 40, pp. 171–181.