Twin (green and digital) patents identification: an automated patent landscaping approach

Francesca Ghinami[1], Sandro Montresor[2], and Stefano Usai[1]

[1]University of Cagliari, Department of Economics and Business, Italy
[2]University of Trento, Department of Economics, Italy

1 Introduction

The identification of green, digital, and twin (green and digital) patents is of growing importance for understanding technological innovation in the context of environmental and digital transitions. However, there is significant variation in the methods used to classify such patents, each with its strengths and limitations. This paper aims to critically assess the existing approaches to patent identification, highlighting their limitations and exploring potential methodological innovations to improve accuracy and comprehensiveness. By addressing these gaps, this study contributes to the ongoing effort to develop more robust methods for patent landscaping, particularly for green and digital technologies. In response to the limitations of existing methods, we propose a fully automated hybrid-method (that combines code-based, keyword-based, and machine-learning identification techniques), with a rule-based seed selection to ensure precision and replicability, and a Transformer based pruning for deep relevance understanding. Moreover, by exploiting the most precise transformer for patents, PaECTER, pre-trained on patents texts and citation networks, we reduce the computational cost, and remove noise while maintaining broad patent coverage. Moreover, by reducing the human intervention to the sole keywords definition, we

Acknowledgements

This study was funded by the European Union - NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP F53C22000760007).

aim to obtain the most reproducible and scalable way to identify patents of a specific type, in our case, 'twin' ones.

2 Literature Review: existing methods, potential and limitations

The intersection of digital and green technologies has gained prominence as a key driver of innovation, particularly within the broader framework of the twin transition. Digital technologies, including those associated with Industry 4.0, enable enhanced efficiency through ICT integration, while green technologies focus on reducing environmental impact. The convergence of these domains highlights the crucial role of digitalization in fostering sustainability, yet identifying and classifying twin transition technologies remains a complex task. Traditional patent landscaping methods exhibit significant limitations, necessitating more sophisticated and integrated approaches.

One of the most widely used approaches relies on classification-based methods, which utilize predefined patent classification schemes such as the CPC and IPC systems. These provide a structured and replicable means of categorizing technological innovations. However, they are inherently rigid and updated infrequently, making them ill-suited for tracking emerging and interdisciplinary technologies.

In contrast, keyword-based approaches offer greater flexibility, allowing patents to be identified based on textual content in titles, abstracts, or descriptions. While seemingly adaptable, these methods present critical challenges. Linguistic variability, including the presence of synonyms and evolving technical jargon, can result in both false positives—where irrelevant patents are retrieved—and false negatives—where key patents are overlooked. Furthermore, these methods are prone to "trending effects," where popular keywords disproportionately highlight certain technologies while overlooking others, further distorting the landscape of twin patent identification.

Citation-based methods provide another alternative by leveraging backward and forward citation networks to trace technological linkages between patents. While these approaches can be useful in mapping knowledge diffusion and innovation pathways, they also come with notable drawbacks. Citations tend to favor well-established patents, leading to an overrepresentation of older technologies while recent and disruptive innovations remain underdetected. Strategic citation practices, where applicants selectively reference patents to strengthen claims or align with regulatory frameworks, introduce additional biases. Moreover, examiner-added citations, which are sometimes based on classification conventions rather than actual technological

relevance, add further noise to the dataset.

Another widely used approach is expert-driven classification, where domain specialists manually curate patent datasets. This method offers high accuracy and domain-specific relevance, particularly when identifying emerging technologies that might not yet be well-represented in structured databases. However, maintaining such expert-curated datasets over time also proves impractical, given the rapid evolution of technological landscapes.

Machine learning approaches are increasingly been explored as a means to overcome some of these limitations by automating patent classification. Supervised ML models, in particular, show promise in identifying relevant patents with high accuracy. However, they require large, high-quality labeled datasets for training—something that is often unavailable for twin patents due to their emerging status. The reliability of these models is further dependent on training data quality, meaning any biases present in the dataset can directly impact classification outcomes. Deep learning approaches, such as BERT-based classifiers, have demonstrated improvements in accuracy but come with high computational costs, limiting their practical application at scale.

Recognizing these challenges, recent studies have increasingly adopted mixed-methods approaches, combining elements of classification-based, keyword-based, citation-based, and machine learning techniques to enhance accuracy and coverage. Notable among these efforts is the work of Jindra and Leusin (2022), who integrate six distinct strategies, leveraging classification codes, keyword searches, and expert-driven selection to identify twin patents. Their framework represents a step forward in combining different methodologies, yet limitations remain.

Building upon these insights, this study proposes an integrated approach that addresses the shortcomings of previous methodologies. By combining rule-based classification, citation expansion, and machine learning-driven filtering, we introduce a fully automated, unsupervised patent landscaping technique designed to systematically identify digital-sustainable twin patents. In addition to improving accuracy and representativeness, this method aims to enhance replicability and scalability, offering a novel contribution to the field of patent analysis.

3 Methodology: Towards an Integrated Approach

The method proposed in this study builds upon and integrates these approaches, beginning with a rule-based seed selection process. This selection leverages a combination of keywords, CPC and IPC codes, and citation data, ensuring that identified

patents meet multiple criteria rather than relying on a single method. Unlike previous studies that utilize isolated approaches, our framework requires patents to be identified by at least two out of six different selection criteria, reducing sectoral or geographical biases and ensuring a more representative dataset. Following seed selection, an expansion phase is conducted to capture related patents through backward and forward citations. To enhance representativeness while avoiding the inclusion of irrelevant patents, a two-step expansion process is employed: (1) an initial expansion based on CPC class similarities and (2) a secondary expansion incorporating patents linked via backward and forward citations. This ensures that emerging technologies and interdisciplinary innovations are captured within the dataset. The pruning phase is executed through a transformer-based machine learning model, specifically the PaECTER BERT. This model has the benefit to be pre-trained specifically on patent texts and citation networks, and has been recently found to be the best performing BERT transformer for patents by ghos et al. (2024). By leveraging a pre-trained model rather than manually labeled training data, the method minimizes human intervention while maintaining high accuracy and scalability. The pruning step refines the expanded dataset by filtering out false positives, ensuring that only patents that genuinely integrate both digital and green technological elements are retained. Finally, given that we apply this method to a quite particular type of patents, *i.e.* those that include both a digital and a sustainability-related type of technology, we propose two unsupervised methods to test the adherence of the pruned expanded set to the seed set. This part was particularly challenging, given that all the information contained in patents was exploited either for the seed selection (IPCs, CPCs, keywords) of for the expansion (CPCs and backward-forward citations), or for the classification (text similarity). We opted to exploit a Semantic Component Similarity Analysis, where semantic embeddings from PaECTER, computed on sustainable and on digital (but not twin) patents only, are exploited to create two semantic centroids, to then measure cosine similarity of each pruned patent with both the sustainable and the digital centroids.

4 Data

The empirical analysis is conducted using the Patstat Autumn 2024 database, a comprehensive dataset of global patent records maintained by the European Patent Office (EPO). The database includes 12,692,387 patents applications, corresponding to 11,028,229 patent families. Given our focus on replicability, and that most available methods for patents landscaping are conducted on Google Patents, in our replication

code we also issue an R package (rpatstat_build) that automatically builds any version of the Patstat database, released between the 2012 and 2024.

5 Preliminary results

A first analysis of the patents identified using each method reveals minimal to no overlap between the patents identified by the different methods proposed by Jindra and Leusin (2022). This underscores the limitations of both the individual methods and the combined approach in capturing a cohesive set of relevant patents.

	Description of identification	**Identification strategy:**	***digital***	***green***
1	Patents tagged with both Y04 and Y02 codes	Specialists' opinion	*CPC*	*CPC*
2	Patents with at least one *digital* and one *sustainability keyword* in their title or abstract	Keyword-based	*keyword*	*keyword*
3	Patents with at least one *digital keyword* in the title or abstract and that are also classified under the *Y02 code*	Specialists' opinion & keyword-based	*keyword*	*CPC*
4	Patents that have at least one *AI-related keyword* in the title or abstract and are also classified under the *Y02 code*	Specialists' opinion & keyword-based	keyword *(AI)*	*CPC*
5	Patents classified with at least one of the considered *digital-related IPC groups* and also classified under the Y02 code	Classification-based & specialists' opinion	*IPC (group)*	*CPC*
6	Patents classified with at least one of the considered *digital-related IPC subclasses* and also classified under the *Y02 code*	Classification-based & specialists' opinion	*IPC (subclass)*	*CPC*

Table 1: Twin patents identificatiom methods proposed by Jindra and Leusin (2022)

Out of the 431,438 distinct patent applications (belonging to 117'920 patent families) identified as twin by any of the methods in Table 1, none are identified by all methods, and only 10'225 are identified by at least two methods. We proceed to select these 10'225 patents as seed, and to expand the seed by forward (3370) and backward (2409) patent families in level 1, and 335 patents in the 5 most relevant CPC subclasses [1]). We expand again this first level through bidirectional citations, obtaining a final expansion set of 5109 unique patent families related to - but not present in - the seed. Once selected the antiseed, constitued by 10'225 patent families randomly selected outside the seed and expansion set, we computed

[1]B60L – Electric Propulsion; Powerr Supply for Electrically Propelled Vehicles. H02J – Circuit Arrangements or Systems for Supplying or Distributing Electric Power Y02B – Climate Change Mitigation Technologies Related to Energy Efficiency in Buildings. G05 – Controlling; Regulating

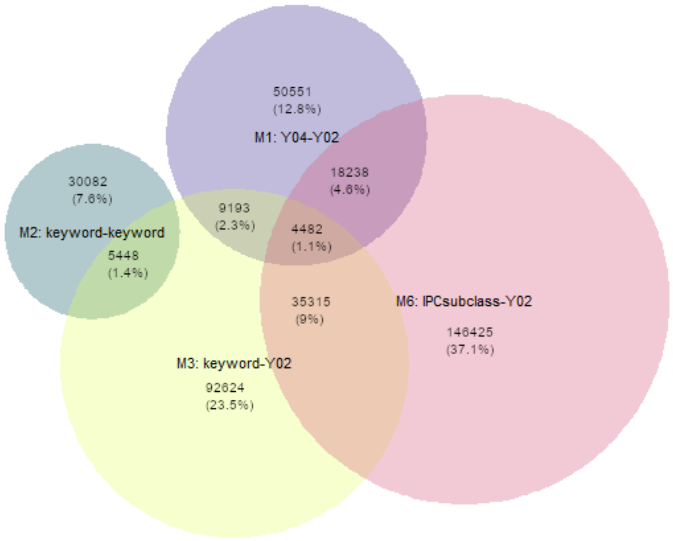

Figure 1: Overlap of patents identified by different modules.

the semantic emeddings for the three sets. Preliminary results show that our hybrid approach effectively improves both precision and comprehensiveness in twin patent identification. The comparative analysis with existing methodologies in terms of the precision, accuracy and robustness is currently underway. The method's sensitivity to different input configurations, including alternative keyword sets and classification criteria, is also being assessed to further refine its applicability.

6 Implications for Research and Policy

This research contributes to both the methodological literature on patent landscaping and the policy debate on the twin transition. By offering a reliable tool for identifying twin patents, the study enables policymakers and scholars to better assess technological trajectories, track innovation diffusion, and design targeted interventions that support the simultaneous advancement of digitalization and sustainability. The findings also have implications for technology forecasting, as the method allows for early detection of emerging twin transition trends, thereby aiding decision-making in both the public and private sectors.

7 Conclusion and Future Work

By integrating multiple methodologies into a unified framework, this study offers a novel, scalable, and replicable approach to identifying twin patents. Future research will focus on refining classification models by incorporating additional contextual features, testing alternative embeddings for text similarity measurements, and conducting large-scale validation across multiple datasets to ensure robustness and accuracy. Beyond patent classification, this work underscores the potential of machine learning-driven approaches for analyzing complex and interdisciplinary innovation landscapes.