Title: **Rethinking Patent Domains: An Application to Wind Energy**

Authors: Kaitlin Fung, Anna Goldstein, Erin Baker, Yiwen Wang

Abstract: Wind energy patents are conventionally defined using Cooperative Patent Classification (CPC) and International Patent Classification (IPC) codes that represent wind motors (F03D) and wind energy (Y02E 10/70). This study examines whether these codes sufficiently represent the wind energy patent domain. Using a combination of keywords and classification codes identified through expert input and manual review, we construct an expanded domain with a 7.5% increase in patent count, a 6% increase in recall, and only a 1% decrease in precision for wind energy patents. We also found that the conventional domain is not unbiased; it underrepresents patenting from China and patents published since 2010. This is important because an improved patent domain allows for a holistic patent data set, which is necessary to complete a thorough analysis of the wind industry. Through this wind energy application, we extend our methodology to form a generalized patent search process that can be used to target technology domains within a patent data set.

Keywords: Patent searching, Wind Energy, Search Method, Cooperative Patent Classification, International Patent Classification

1. Introduction and Background[1]

With the considerable dependence of world economies on fossil fuels, a global shift towards decarbonizing electricity sources is required to mitigate climate change. Currently, wind energy is one of the fastest growing renewable energy technologies and is poised to make a major contribution to the low-carbon energy transition [1]. However, despite its high potential, concerns about high initial costs and regulatory uncertainty around wind energy may stagnate its overall progress [2].

Technology innovation in wind energy is expected to result in cost reductions and accelerated deployment. Therefore, examining innovation trends in wind energy is important to support the overall development of wind energy technologies. Patent data can be employed to better understand this progress. Previous studies have used patent applications as indicators of technological development in the wind energy industry [1,3-6]. In this paper we examine the conventional method for defining the domain of a technology such as wind energy, and investigate an alternate scheme, which has potential to lead to a more expansive domain while preserving a high degree of precision.

Studies of wind energy patenting activity rely on a narrow set of search strategies to establish the domain of patents relevant to wind energy [7,8]. The most common search strategy employs the use of patent classification schemes, including the Cooperative Patent Classification (CPC) system and the International Patent Classification (IPC) system; the subclass F03D represents wind motors in both CPC and IPC schemes. The European Patent Office (EPO) and United States Patent and Trademark Office (USPTO) developed the CPC scheme in 2013 [9]. CPC was based heavily on the IPC scheme and retained much of the same hierarchies, titles, and expandability, although CPC is more granular, with more than 250,000 classification entries compared to IPC's 70,000 entries. CPC and IPC classifications are typically assigned by patent examiners.

CPC also includes a Y section for tagging emerging technologies, in parallel with the main sections of CPC classification entries [7]. In particular, the Y02 section of the CPC scheme tags patents related to the reduction of greenhouse gas emissions, energy generation, transmission, or distribution [10]. The CPC code that represents wind energy patents is Y02E 10/70. Y02 classifications were developed in consultation with a broad group of stakeholders [9]. They are assigned algorithmically by EPO and were applied retroactively to patents issued before 2013 [9]. Unfortunately, the absence of public documentation of this algorithm presents a challenge for patent searchers who would like to assess the reliability of Y-section tags, relative to human classification. A previous study has shown that patent applications filed through the EPO, USPTO, and Chinese Patent Office (SIPO) showed a 97.1% overlap between Y02E 10/70 and F03D classification codes [3].

Most studies on wind energy patents use only conventional classification codes for patent retrieval. For example, a number of studies used IPC code F03D to retrieve wind energy patents [11-17]. Besides the F03D code family, other studies use CPC code Y02E 10/70 (defined as wind energy) in a similar manner [18,19].

The union of F03D and Y02E 10/70 can be thought of as the conventional method for searching for wind energy patents. There is some question, however, whether this is definitive. In

[1] Nonstandard abbreviations used in this paper include:
DD = Domain Definition = A selection of patent criteria that establishes the boundaries used to retrieve patents related to a certain technology
WEDD = Wind Energy Domain Definition a selection of these patent criteria that establishes the boundaries used to retrieve patents related to wind energy

this vein, a previous study by Malhotra et al. used 19 additional classification codes to broaden their search and extract wind turbine system patents from the 2016 Spring EPO PATSTAT database [20]. In their study, they iteratively developed search criteria, using keywords to expand the classification and sub-classification codes assigned to retrieved patents [20]. While this indicates that the conventional methods can be improved upon, this paper did not evaluate their dataset in terms of precision. Furthermore, they did not report on the set of keywords used to retrieve their code set along with total number of iterations that were required to get to their final code set and the conditions that determined them to terminate the iteration.

In a related paper, Tsai et al. (2016), worked to establish a domain specifically for offshore wind, which is not directly represented by codes. In their method, they retrieved patents using a target list of wind energy keywords found in a patent's title or abstract and filtered out land-based patents with a list of offshore specific keywords. Following this, they completed a manual review of every patent within this subset to remove incorrectly identified offshore wind energy patents [21]. Although this manual review process was completed on a small subset of patents, this approach would be highly time-intensive for a larger set of patents.

A few other studies combined classification codes and keywords too [22, 23]. Lee and Lee (2013) searched "wind power/energy" in abstract; then narrowed the results by excluding patents with irrelevant IPC codes [22]. Odam and de Vries (2020) searched patents by keywords to find high presence classes, which were then used for further searches [23]. However, these studies did not validate their searching results through expert validation, thus cannot be evaluated by precisions and recall.

The purpose of this paper is to develop these methods for expanding the domain further, by introducing a rigorous methodology using both classification codes and keywords. Our proposed methodology allows patent searchers to retrieve and review technology related patents that are not tagged with an explicitly defined classification code. This method offers an approach to compare the performance of each retrieved patent set by reviewing a random sample of patents, which offers a less manually intensive approach than manual review of a complete patent dataset. Using this method, we arrive at a new domain definition for wind energy patents that improves on the conventional one in terms of size, efficiency, and performance. We first establish a general methodology for deriving and comparing three alternative domain definitions for any technology of interest in section 2, then we show the results of applying it to wind energy in section 3.

2. Domain Definition Methodology

Here we present our generalized approach to patent domain review for the establishment of three different domain definitions. The methodology to establish and assess all three domain definitions is visually summarized in figure 1. Through this process, a patent searcher can compare between each domain definition to choose the data set that is most suitable for their analysis. This generalized approach can be applied broadly to other emerging technologies. In section 3, we provide more detail on the application of this method to the wind energy domain.

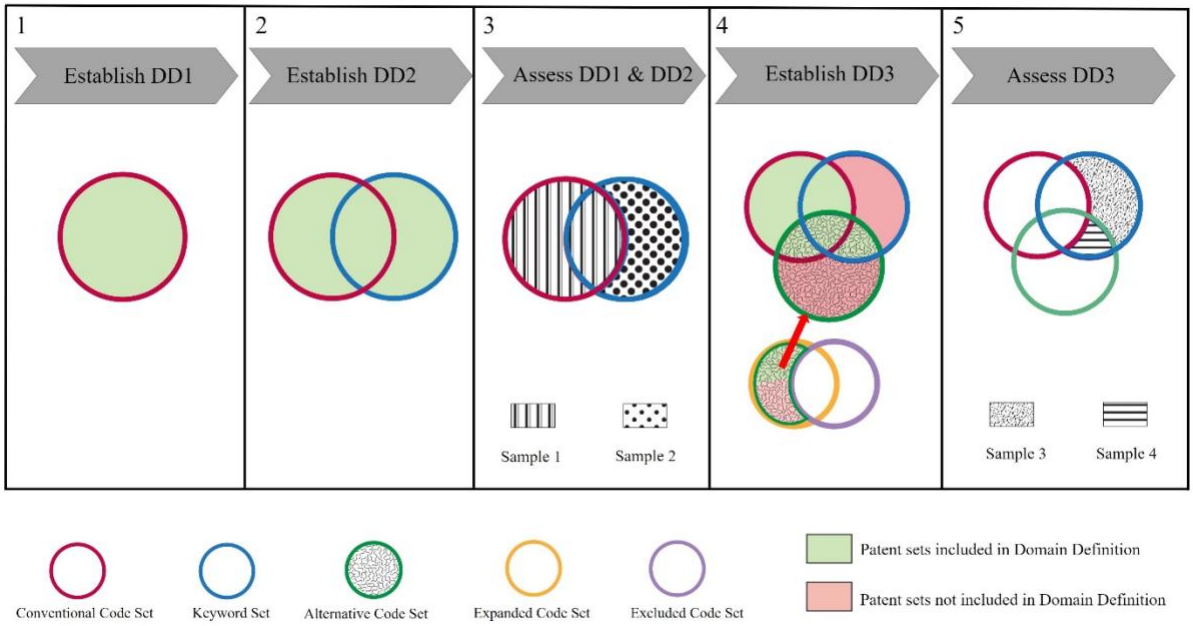


Figure 1. Patent Domain Definition Methodology. The methodology is completed in five steps followed from left to right. Within step 4, there is a pre-step of establishing the Alternative Code Set before establishing Doman Definition 3 (DD3). Each established Domain Definition is indicated as green shading for patent sets included in the domain definition and red shading for patents that are excluded from the domain definition. Cross hatching textures are used to indicates regions to sample within the methodology. Colored outline circles represent defined patent sets.

To begin, a patent set is defined as a collection of patents. Furthermore, a domain is a patent set that is retrieved by a set of patent criteria. A set of patent criteria, for example, could include a list of classification codes and/or keywords which, if present, qualify a patent for membership in the patent set. A domain definition (DD) is then a selection of these patent criteria that establishes the boundaries used to retrieve patents related to a certain technology. Table 1 outlines the set names used in this methodology and their associated definitions and set logic.

Table 1. Set Names and Definitions

Patent Set Name	Definition	Set Logic
DD1 Set	All granted patents tagged with a conventional classification code	--

Keyword Set	All granted patents with one or more targeted keywords in the title or abstract	--
DD2 Set	All granted patents in the union of DD1 set and Keyword Set	DD1 Set ∪ Keyword Set
Expanded Code Set	All granted patents with one or more codes that are determined to occur more frequently among confirmed technology-related patents	--
Excluded Code Set	All granted patents with one or more codes that are determined to occur more frequently among confirmed non-technology-related patents	--
Alternative Code Set	All granted patents that are members of the expanded code set AND not a member of the excluded code set	Expanded Code Set – Excluded Code Set
DD3 Set	All patents that are present in the union of DD1 Set and the intersection of the Keyword set and the Alternative Code Set	DD1 Set ∪ (Keyword Set ∩ Alternative Code Set)
New Additions Set	All granted patents within DD3 set that are not in DD1 Set	DD3 Set – DD1 Set

The first step of the process (figure 1) is to review the classification code schema including CPC and IPC to identify codes that explicitly state their relevance to the technical field of interest. These codes are referred to as conventional codes, and the inclusion of any of these codes are the patent criteria for the first domain definition (DD1). Granted patents that are retrieved by the DD1 patent criteria are referred to as DD1 set. This methodology is limited to granted patents to consider patents recognized under their corresponding patent authority agency. This would include granted patents only, and exclude ungranted patents, which are patents that were filed, yet never approved for ensuring patent protection rights.

The second step is to identify a list of relevant keywords that target the technical field; patents with a keyword in either the title or abstract are referred to as belonging to the Keyword Set (figure 1). Although patents are limited to the keyword search by its title or abstract for the wind energy application, the use of claims or full text for searching could be explored. The union of DD1 Set and Keyword Set is the second domain definition (DD2). In other words, DD2 is defined by the inclusion of all patents with a relevant keyword within its title or abstract, or it is tagged by a conventional classification code. All patents that are retrieved by the DD2 patent criteria are referred to as DD2 set.

Third, we draw two random samples of patents from DD1 set and DD2-DD1 set (samples 1 and 2 in figure 1). A sufficient size of the random sample can be determined with Equation C1 in Appendix C.

The purpose of this step is to compare how effective each domain definition is at capturing relevant patents. A patent that is retrieved by a domain definition is considered a *predicted positive*; while a patent that is not retrieved by a domain definition is a *predicted negative*. Furthermore, a patent that is confirmed by experts as being related to the technology in question is considered an *actual positive*. Conversely, a patent that is confirmed as being not related to the technology in question is an *actual negative*. These classifications can be used to assess each domain definition through a precision rate. Predicted positive patents that are also actual positives are considered to be "True Positives" or TP. Predicted positive patents that are actual negatives are referred to as "False Positives" or FP.

Table 2. Domain Definition Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	Retrieved by DD and validated by experts as technology related True Positive (TP)	Retrieved by DD and excluded by experts as not technology related False Positive (FP)
Predicted Negative	Not retrieved by DD and validated by experts as technology related False Negative (FN)	Not retrieved by DD and excluded by experts as not technology related True Negative (TN)

$$\text{Precision Rate} = \frac{\text{Retrieved by DD and validated by experts as technology related}}{\text{All patents retrieved by DD}} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{(Eq. 1)}$$

The precision rate (Eq. 1) is the estimated percentage of patents that are retrieved by a domain definition (DD) and validated by experts as being technology related among all retrieved patents. All patent that are retrieved by the DD are predicted positive patents (TP+FP). Furthermore, patents that were not retrieved by the DD, but were validated by experts as being technology related are referred to as "False Negative" or FN. Table 2 includes a confusion matrix that is used to define and compare the predicted retrieval of a domain definition to how it is assessed through manual review.

$$\text{Recall Rate} = \frac{\text{Retrieved by DD and validated by experts as technology related}}{\text{All technology related patents}} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{(Eq. 2)}$$

Another way to assess a domain is through its recall rate. Recall rate is the ratio of patents retrieved by the domain definition and validated by experts as being technology related among all technology related patents in the entire patent data set (Eq. 2). To assess the recall of DD1, we use DD2 as our standard by assuming that it encompasses all possible wind energy patents, that is, there are no predicted negative patents (FN+TN) associated with DD2. Therefore, both the precision and recall rates of DD1 and DD2 can be estimated, and patent examiners can assess whether DD1 or DD2 is better suited for their analysis.

However, if further improvement is still required after assessing the precision and recall rates of DD1 and DD2, then a third domain definition (DD3) can be established through the addition of an Alternative Code Set. The addition of the Alternative Code Set is used to eliminate non-technology related patent that are retrieved from the Keyword Set.

The Alternative Code Set can be generated by the patents randomly selected in sample 2 (figure 1), or the segment of DD2 that is excluded from DD1 (DD2 Set-DD1 Set). Classification codes with a high presence among patents that are identified as technology-related can be defined as the Expanded Code Set. Furthermore, classification codes with a high presence among patents that are identified as not technology-related can be defined as the Excluded Code Set. The Alternative Code Set is defined as the exclusion of the Excluded Code Set from the Expanded Code Set, as shown in figure 1.

With the construction of the Alternative Code Set, DD3 is then defined by the union of DD1 and the intersection of the Keyword Set and the Alternative Code Set (figure 1). The precision and recall of DD3 can be estimated using the results from sample 2 and verified through the results of samples 3 and 4, again using DD2 as the standard. Finally, after completing the methodology, the performance of DD1, DD2, and DD3, and New Additions can be compared to select the most acceptable domain definition for a patent searcher's analysis.

3. Application of the Domain Definition Methodology for Wind Energy

In this section, we apply our patent domain definition methodology for Wind Energy. The Spring 2020 version of PATSTAT was purchased for this study. Patent data was queried through SQL (Structured Query Language) statements. We identified the set of wind energy relevant patents using the above methodology, based on three key attributes of patents: title, abstract, and classification codes. Like the generalized methodology (DD1, DD2, DD3), three versions of the wind energy domain definition (WEDD) are explored (WEDD1, WEDD2, WEDD3). Figure 2 shows the graphical representation of results for all three developed wind energy domain definition. Additionally, Table 3 summarizes the statistical results for all three domain definitions.

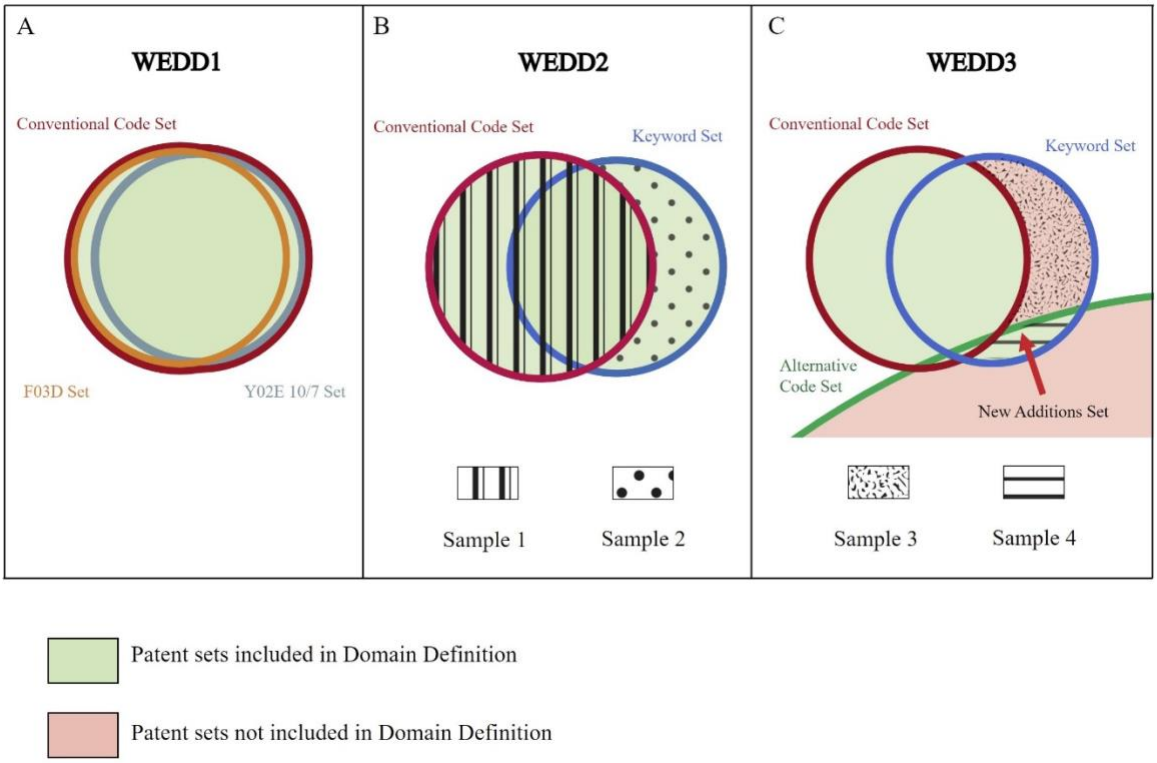


Figure 2. Graphical representation of the expansion of the wind energy domain definition.

Textured shaded regions correspond to manually sampled patent sets. Green shading represents patent sets that are included in the domain definition; red shading represents excluded patent sets. Panel A depicts WEDD1 as the union of conventional classification codes (red outline): F03D set (orange outline) and Y02E 10/70 set (grey outline). In this figure, Conventional Code Set and WEDD1 are used interchangeably. Panel B represents WEDD2 as the union of WEDD1 Set (red outline) and the Keyword Set (blue outline). Panel C depicts WEDD3 as the union of WEDD1 with the intersection of the Keyword Set (blue outline) and the Alternative Code Set (green outline). The Alternative Code Set is much larger than the WEDD sets and is only shown in part. The New Additions set in panel C represents the patents that are captured by WEDD3 but were not present in WEDD1.

Table 3. Descriptive statistics for the alternative Wind Energy Domain Definitions (WEDDs) considered.

	WEDD1	**WEDD2**	**WEDD3**
Patent Criteria	Any of the conventional classification codes	Any of the conventional classification codes or targeted keywords	Any of the conventional classification codes, or a combination of targeted keywords and additional classifications
Patent Criteria Set Logic	F03D Set ∪ Y02E 10/70 Set	WEDD1 ∪ Keyword Set	WEDD1 ∪ (Keyword Set ∩ Alternative Code Set)
Number of Granted Patents	86,410	118,135	92,933
Average Year Granted	2004	2006	2004
Number of Granted Patents since 2010	58,767	84,617	64,615
Percent of Granted Patents since 2010	68%	72%	70%
Number of Granted Patents from China	29,267	55,071	34,759
Percent of Granted Patents from China	34%	47%	37%
Estimated Precision Rate	98%	83%	97%
Estimated Recall Rate	86%	100%	92%

Table 3 shows that the expansion of WEDD1 (the conventional classification codes) to WEDD3 (our final wind energy domain) resulted in an increase in global patent count from 86,410 patents to 92,933 patents within PATSTAT. The precision rate from WEDD1 to WEDD3 reduced slightly from 98% to 97%, indicating that a few non-wind energy patents are being included in the domain definition, but the estimated recall rate increased from 86% to 92%, indicating that we are missing fewer wind energy related patents. The following subsections go through this process in detail.

3.1 Establishing Wind Energy Domain Definitions 1 and 2 (WEDD1 and WEDD2)

From the methodology, we first established the conventional wind energy domain definition (WEDD1), which consisted of all granted patents with at least one classification within F03D or Y02E 10/70, which are meant to capture wind motors and wind energy, respectively [1]. WEDD1, therefore, is the union of the granted patent sets that contain Y02E 10/70 and F03D (F03D Set ∪ Y02E 10/70 Set). We accounted for F03D codes in both IPC and CPC schemes; Y02E 10/70 is exclusive to the CPC scheme.

Next, we examined an expanded definition for the wind energy domain, WEDD2. Its patent criteria included the union of WEDD1 and the Keyword Set, i.e. the set of patents containing one or more wind energy keywords in its title or abstract. Keywords were obtained from Tsai et. al (2016) and suggestions from expert volunteers through the UMass Amherst Wind Energy Fellows program. The list of keywords is as follows: wind power, wind turbine, wind energy, wind generator, wind farm, windmill, energy of wind, energy from wind, wind rotor, wind axis, wind blade, wind generating set, wind array, wind plant, wind park, wind platform, wind base, wind hub, wind control, and wind installation. The use of these keywords were retrieved through a like operator in SQL where clause, to extract proximity matches through a combination of wildcard characters which are listed in the appendix within table A1. The use of the wildcard characters were chosen to capture plurality in the keyword and additional strings before and after the keyword. For instance, "%wind_ax%s%" captures both wind axis and wind axes.

Next, we assessed the two domains. After establishing both WEDD1 and WEDD2, we sampled 100 random patents from WEDD1 (sample 1 in figure 2) to review for wind energy relevance. A patent was identified as wind energy related if it is related to grid-connected stationary electricity generation powered by wind, either offshore or onshore. From this manual review, we estimate the precision of WEDD1 to be 98%.

We also conducted manual review of a random selection of patents from the Keyword Set - WEDD1 Set (shown as sample 2 in figure 2). Four volunteers participated in reviewing a set of 257 patents with these criteria. Each participant reviewed a sample of patents for wind energy relevance, as defined above. Determinations were made based on the patents' title and abstract (and full text if more information was needed). If no determination could be made based on the available information, the patent was excluded from the sample. All four participants reviewed a common set of 25 patents within their sample; inter-rater reliability for this common set was 100%. From this review process, we estimated the precision of the Keyword Set to be 42%.

Based on the results from these two samples, we estimate the precision of WEDD2 to be 83% -- significantly lower than WEDD1 (98%). As in the generalized methodology, we use WEDD2 as our most expansive domain definition. If there are no false negatives associated with WEDD2 (i.e. if recall of WEDD2 is 100%), then we can estimate the recall of WEDD1 to be 86%.

3.2 Establishing and Assessing Wind Energy Domain Definition 3 (WEDD3).

At this stage in our generic methodology, the searcher may choose to stop, satisfied with either DD1 or DD2. While WEDD2 identified new patents not in WEDD1, it had a relatively low precision. We therefore use an additional set of classification codes aimed at increasing

precision while maintaining recall. This set of codes, referred to as Alternative Code Set, is used to eliminate some of the non-relevant patents introduced by the Keyword Set, so that WEDD3 can be larger than WEDD1 with greater recall, but smaller than WEDD2 with greater precision.

The Alternative Code Set was constructed as follows. We selected a random sample of patents that were within Keyword Set but not in WEDD1 (sample 2 in figure 2). Experts then evaluated these patents for wind energy relevance through manual review. We then examined the classifications (both CPC and IPC) of all patents in sample 2, both relevant and not relevant.

After this review, we furthered our assessment with the following analysis. Let P represent patents within random sample 2; and let x represent the patents within P that are confirmed by experts as being wind energy relevant. Therefore, P-x represent patents that are excluded by experts, i.e. rated as not relevant to wind energy. Table 4 summarizes these patents set groupings.

Table 4. Sample 2 (Keyword Set – WEDD1) Patent Set Definitions

P	Patents within sample 2 reviewed for wind energy relevance
x	Expert validated patents within sample 2
P-x	Expert excluded patents within sample 2

We then examine the rate at which all codes are present among patents in x and P-x. The goal of this analysis is to populate the patent criteria for the Alternative Code Set with codes that appear frequently in x, while excluding those codes that appear frequently in P-x.

We perform this Alternative Code Set analysis at the subclass level, i.e. the first four characters of the CPC or IPC code, which captures the broadest function of a patent. As more characters are appended to the code, the described function becomes more specific. Consider, for example, the descriptions of a group and subgroup within the F16H subclass:

F16H: Gearing
F16H 57: General details of gearing
F16H 57/023: …Mounting of installation of gears or shafts in gearboxes

We then examine the presence of a code within the x set and P-x set. A presence of a code is defined as the number of times a code is tagged within a patent set. We use equations 3 and 4 to give every classification code i in P a score to represents its presence among x and its presence among P-x. We then rank each code in descending order by its Expanded Presence (Eq. 3) and Excluded Presence (Eq. 4).

$$\text{Expanded Presence}_i = \frac{Number\ of\ patents\ with\ code_i\ in\ x}{Number\ of\ patents\ in\ x} \quad \text{(Eq. 3)}$$

$$\text{Excluded Presence}_i = \frac{Number\ of\ patents\ with\ code_i\ in\ P-x}{Number\ of\ patents\ in\ P-x} \quad \text{(Eq. 4)}$$

After reviewing the data on Expanded Presence and Excluded Presence for codes in sample 2, we developed the following conditions: if a code had Expanded Presence >=4% and its Expanded Presence > Excluded Presence, then it was considered as a patent criterion for the Expanded Code Set. Conversely, if a code had an Excluded Presence >=4% and its Excluded Presence > Expanded Presence, then it was considered a patent criterion for the Excluded Code

Set. A 4% cutoff was selected due to a natural break in the data for the Expanded Presence and Excluded Presence around this percentage point.

As an example, the subclass H02P had an 8% Expanded Presence and only 1% Excluded Presence. Therefore, patents tagged with H02P were included in the Expanded Code Set. On the contrary, the subclass Y02B had an Excluded Presence of 11% and an Expanded Presence of 1%. Therefore, patents tagged with Y02B were an element of the Excluded Code Set, and therefore excluded from the Alternative Code Set. The characterization of these codes can then be used to generate an Alternative Code Set. If a code had the same Included Presence and Exclude Presence, then the code was not considered as a patent criterion for the Alternative Code Set.

We then evaluated three alternative formulations of the Alternative Code Set, to compare previous literature with our analytical results.

Alternative Code Set 1: 19 codes from Malhotra et al. as the Expanded Code Set, with no Excluded Code Set

Alternative Code Set 2: 15 codes from our analysis as the Expanded Code Set, with no Excluded Code Set

Alternative Code Set 3: 15 codes from our analysis as the Expanded Code Set from our analysis, and 10 codes from our analysis as the Excluded Code Set

To determine which Alternative Code Set to select, we compared its effect on the precision and recall of WEDD3. The precision and recall for WEDD3 with each Alternative Code Set was estimated by using the precision rate of the Alternative Code Set on the 257 volunteer-reviewed patents in sample 2, which generates conservative estimates for both metrics.

For simplicity, each code set will be referred to as Alternative Code Set X to describe the process for estimating the precision and recall of WEDD3 with each Alternative Code Set options. Equations 5 and 6 represents the precision (Eq. 5) and recall rate (Eq.6) for WEDD3. Pr represents the precision rate, Re represents the recall rate, and C represents the patent count. The subscript of each component describes the associated patent sets. The numerator is the same for both equations. The numerator represents the estimated true positive count, or the number of patents retrieved by WEDD3 that are validated by experts as being wind energy related. This is estimated by summing the estimated number of true positives in the WEDD1 Set and New Additions generated with Alternative Code Set X. However, the denominator differs in both equations. For the precision of WEDD3, the denominator represents all patent retrieved by WEDD3 (TP+FP). For the recall of WEDD3, the denominator represents all possible real cases of wind energy related patents (TP+FN), which assumes that WEDD2 has no false negative patents.

$$\mathrm{Pr}_{\mathrm{WEDD3}} = \frac{(\mathrm{Pr}_{\mathrm{WEDD1}} * \mathrm{C}_{\mathrm{WEDD1}}) + (\mathrm{Pr}_{\mathrm{Alt\ Code\ X\ on\ sample\ 2\ Set}} * \mathrm{C}_{\mathrm{New\ Additions\_x}})}{(\mathrm{C}_{\mathrm{WEDD1}} + \mathrm{C}_{\mathrm{New\ Additions}})} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad \text{(Eq. 5)}$$

$$\mathrm{Re}_{\mathrm{WEDD3}} = \frac{(\mathrm{Pr}_{\mathrm{WEDD1}} * \mathrm{C}_{\mathrm{WEDD1}}) + (\mathrm{Pr}_{\mathrm{Alt\ Code\ X\ on\ sample\ 2\ Set}} * \mathrm{C}_{\mathrm{New\ Additions\_x}})}{(\mathrm{Pr}_{\mathrm{WEDD2}} * \mathrm{C}_{\mathrm{WEDD2}})} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad \text{(Eq. 6)}$$

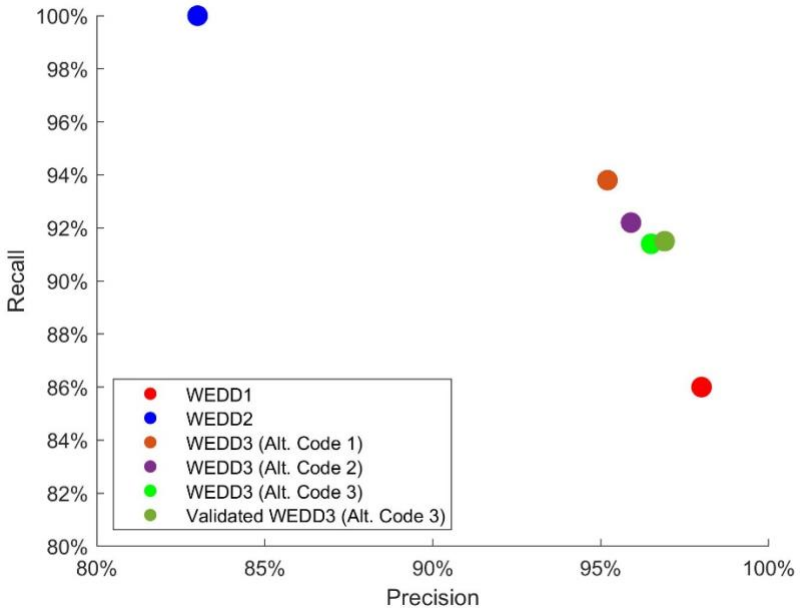


Figure 3. Comparing the precision and recall estimates of WEDD1, WEDD2, and WEDD3 generated by the three Alternative Code Sets

After assessing the precision and recall of WEDD3 with each Alternative Code X (figure 3), we ultimately selected Alternative Code Set 3 (15 codes used to generate the Expanded Code Set and 10 codes used to generate the Excluded Code Set) for use in the final wind energy domain definition (WEDD3). This is due to its higher precision rate relative to the Alternative Code Set 1 and 2 . As seen in figure 3, the use of the Alternative Code Set 3 resulted in a higher precision for WEDD3 (96.5%) when compared to the use of Alternative Code Set 1 and 2 in WEDD3 (95.2% and 95.9%).

To ensure that Alternative Code Set 3 was not over-fitted to the volunteer review sample, we conducted additional manual review. We reviewed 100 excluded patents (sample 3) and 100 retrieved patents (sample 4) for wind energy relevance, and then re-calculated the precision and recall of the Alternative Code Set again to compare with precision and recall estimates with the results from sample 2. Our results were similar, with slightly higher estimates for recall and precision (shown as Validated WEDD3 (Alt. Code 3)in figure 3).

Our final wind energy domain definition (WEDD3) is the union of WEDD1, which was derived from conventional codes, with the intersection of the Keyword Set and the Alternative Code Set 3 (figure 2). From this methodology, WEDD1 was expanded to WEDD3 through inclusion of patents that are referred to as the New Additions Set. The New Additions Set includes all granted patents within the intersection of Keyword Set and the Alternative Code Set excluding patents in WEDD1 Set.

3.3 Analysis of New Additions Set

The expansion of WEDD1 to WEDD3 resulted in an addition of 6,523 patents, which we term the New Additions Set. We find that the New Additions Set is qualitatively different from WEDD1 patents in several ways.

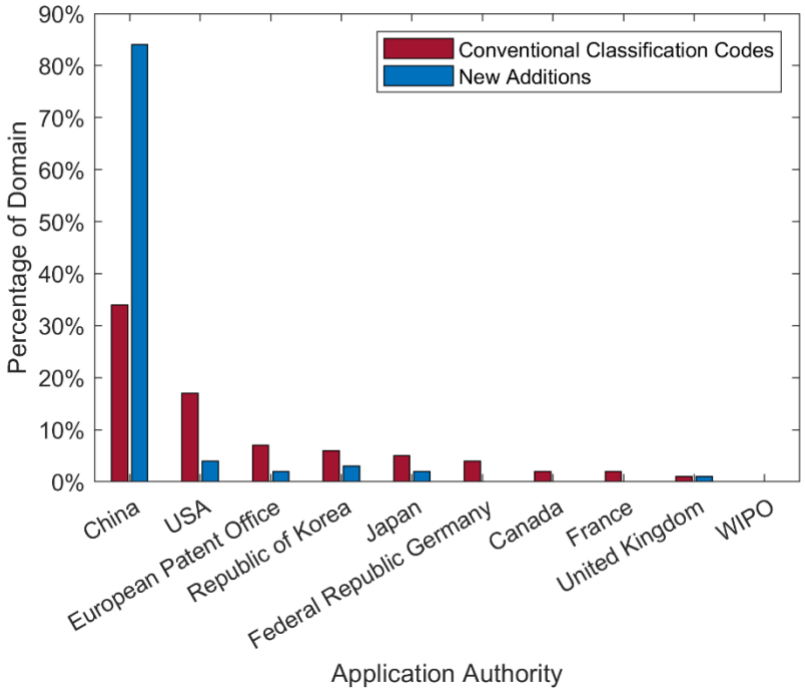


Figure 4. Comparing conventional classification codes with new additions by top 10 application authorities in PATSTAT data set. Each bar shows the percentage of patents in the domain that were granted by the given application authority.

First, patents in the New Additions Set are more likely to have been filed in China, relative to the conventional wind energy domain. Figure 4 shows the top 10 application authorities represented in our wind energy domains. 84% of New Additions Set were from China compared to only 34% of WEDD1 patents. As a reference for comparison, WEDD2 also observed a greater domain percentage of patents from China at 47% in comparison to WEDD1 at 34%.

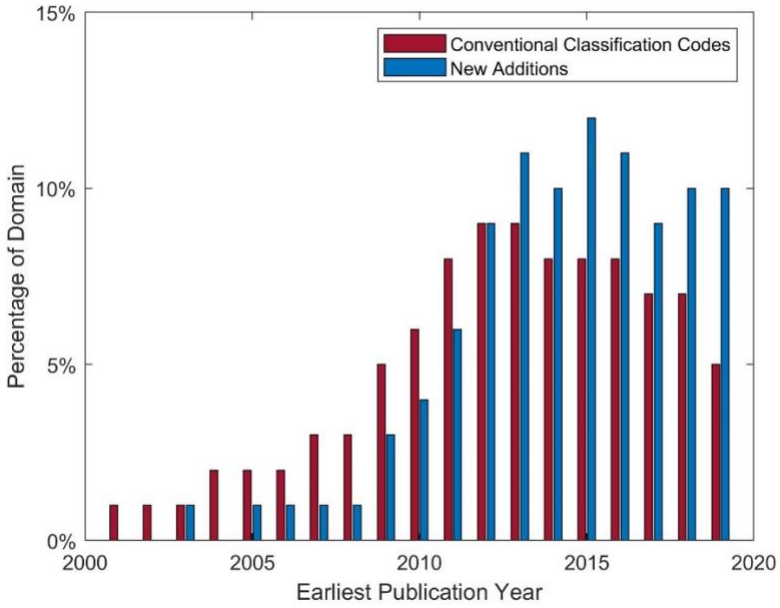


Figure 5. Comparing conventional classification codes with New Additions by earliest publication years from 2000-2020. Each bar shows the percentage of that domain that was published in the given year. This figure does not represent the entire WEDD1 (Conventional Classification Codes) and New Additions Set since it excludes patents published before 2000.

Second, the New Additions were more likely to have been granted in recent years (figure 5): 90% of New Additions were published post-2010, compared to 68% of WEDD1. We investigated whether this result is driven by the high percentage of New Additions filed in China, since nearly 90% of wind energy patents in China have been filed since 2010. Table 5 shows the percentage of patents since 2010, for both WEDD1 and New Additions, for the top three patent authorities: China, US, and EPO. We see a bias toward newer patents in the New Additions in all three patent authorities, indicating this is a general trend and not unique to China.

Table 5. Percentage of Patents Since 2010 for the Top Three Patent Authorities

Patent Authority	CN	US	EP
% of patents since 2010 in WEDD1	88%	50%	76%
% of patents since 2010 in New Additions	94%	65%	83%

We also compared the concentration of targeted keywords in a patent's title or abstract between WEDD1 set and New Additions set. It was found that "wind power", "wind turbine", "wind energy", and "wind generator" have a greater presence within the New Additions set, relative to WEDD1 (see table A1). Through a manual review of 100 patents in this set, we found that patents in New Additions Set were often related to certain topics. It was found that 10% were related to manufacturing, load testing, or material advancements in wind turbine blade technologies, 8% of patents were related to vibration and fault detection for diagnostic testing, 5% of patents were related to hybrid charging stations with battery storage, and 4% were related to offshore wind construction. Through establishment of WEDD3 from the methodology, we see

that these patents would have been missed if WEDD1 was used for wind energy patent analysis, based solely on conventionally defined classification codes.

4. Discussion

From our results, we find deficiencies in the conventional domain for wind energy patents. We initially hypothesized that the conventional classification codes in WEDD1 set excluded some wind energy related patents. After manually reviewing patents that were excluded from the WEDD1 set and included in Keyword set (sample 2), we found that 42% of patents with these criteria were still wind energy related. This confirmed our original hypothesis that the conventional codes (WEDD1 set) excluded relevant patents. Although the inclusion of keywords captures additional wind energy related patents, it also captured non-wind energy related patents. These patents were then further excluded by the use of our Alternative Code Set. The use of both target keywords and the Alternative Code Set increased the recall and expanded the wind energy domain relative to WEDD1 (conventional classification codes), without causing a major reduction in precision. This allowed us to uncover a significant set of patents that are not identified using conventional methods.

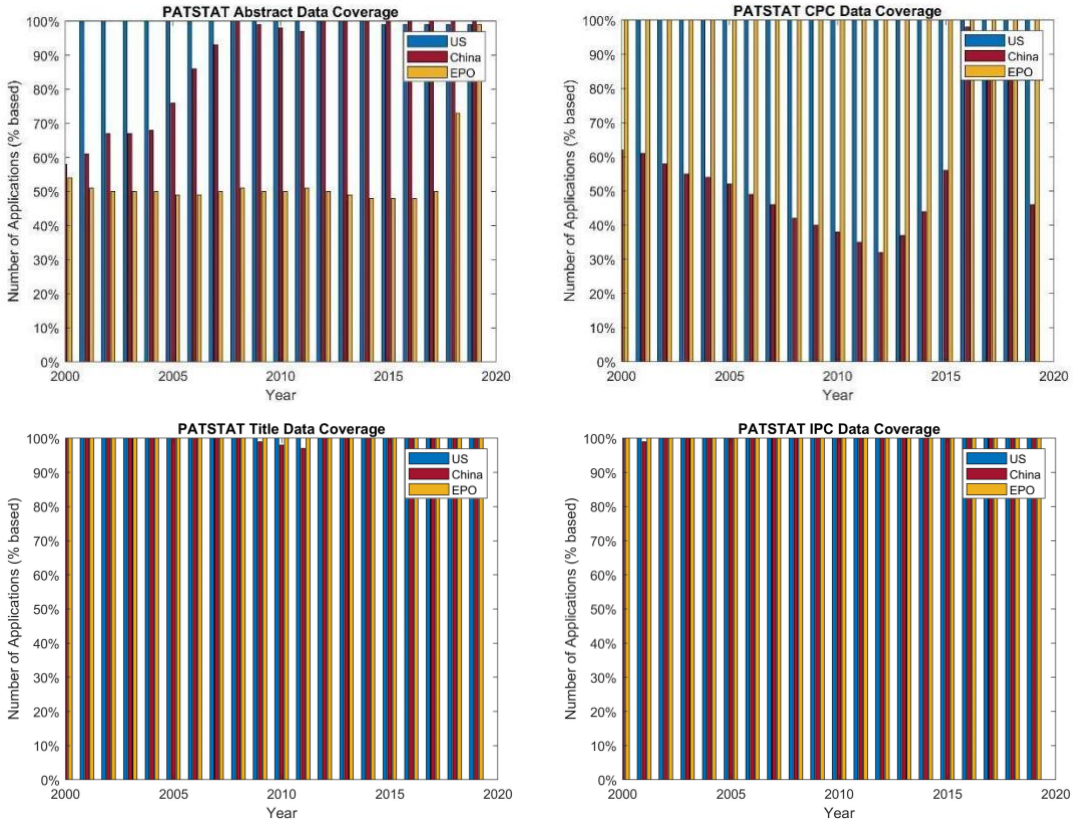


Figure 6. PATSTAT Data Completeness of Abstract, CPC, Title, and IPC Data Fields [24]

Differences between the domain of the conventional WEDD1 domain and the improved WEDD3 domain may be due to discrepancies in the completeness of targeted data fields within the data set, which is a general problem existed within PATSTAT [25-27]. The discrepancies can be seen when comparing the completeness of the title, abstract, IPC, and CPC data fields within the PATSTAT 2020 data set for US, China, and EPO (figure 6). It was found that between these three patent authorities, the completeness for title and IPC data fields stayed consistently high with >95% data completeness between 2000-2020. However, PATSTAT data coverage for

abstracts (figure 6) is greater among US and Chinese patents over this time period, in comparison to EPO.

Data incompleteness may be one reason that a greater number of new additions are from China. On the one hand if wind energy patents are regularly mis-coded when CPC codes are assigned, our strategy can identify them based on the use of IPC codes in the Alternative Code List. On the other hand, if patents are missing abstracts, our strategy might continue to miss them when searching for keywords. These two aspects together might explain the large number of new additions in China – China has high abstract coverage, but low CPC coverage. In addition, the low abstract coverage in EPO may imply that there are still wind related patents that have not been identified there.

Another possible reason we extract more patents from China could be the way that CPC codes are assigned in China, in particular for the Y section. The application of the Y section in the CPC scheme is different across patent authorities. The purpose of the Y section is to tag new technological developments or cross-sectional technologies that span over several CPC sections, with Y02 specifically tagging climate change mitigation technologies. The original implementation of Y02 was by algorithmic search [28], but the ongoing implementation of these tags in practice varies across patent authorities[2]. To date, no studies or public information has shed light on the mechanism by which each patent authority assigns Y section classifications to patents, whether by manual review, algorithmic search, or some other method.

Another potential explanation for the large number of new additions from China is related to innovation in the wind industry – it is possible that the IPC and CPC schemes are not keeping up with the latest advances in wind technologies, thus recent innovations in the wind energy industry may not be represented appropriately by the conventional codes. This is consistent with our findings that new additions are concentrated in recent years.

In order to briefly investigate these questions further, we applied a limited version of our approach (without expert verification) to geothermal energy, a relatively more mature energy technology. We found that this increased the size of the geothermal domain by 2.2% with a higher percentage from China and the US (Appendix F). Thus, while our method uncovered new patents, the percentage increase is less, perhaps adding weight to the hypothesis around innovation – geothermal has not seen the same burst of innovation since 2010 as wind energy has, and therefore, its codes may have better coverage of the type of innovations. On the other hand, the lack of EPO patents in the new additions may reflect the lack of abstract coverage there. Thus, more work remains to understand why some patents are not identified through codes.

One limitation of our study is that we only used keywords in the English language. Patents that were originally written in a non-English language were only searchable when English translations were available on PATSTAT. This is a strength of search criteria based on classification codes, which are not subject to language translation barriers. However, if patents are not tagged appropriately, they may be excluded when searching by only the conventional codes.

The increase in recall rate for WEDD3 over WEDD1 suggests that patent searches using conventional codes are missing wind energy relevant patents. The resulting increase in recall rate from the expanded domain can be useful for analyses that require the total number of wind energy patents, such as comparing the innovation activity in wind energy to other areas of technology. Using the expanded domain that includes additional keywords and classification

[2] Personal communication with EPO staff.

codes can result in a more thorough approach. Additionally, analysis of wind energy patent characteristics with the conventional domain can lead to biases when examining innovation activity in wind energy over time or across different countries.

5. Conclusion

The use of patent data to understand the technological domain of wind energy has become an important tool to understand the progression and growth of this industry. We find that the conventional method for searching for wind energy patents through the F03D and Y02E 10/70 classification codes excludes relevant patents that are not tagged with these wind energy specific codes. In this study, we expand upon the conventional classification codes by appending a set of target keywords (Keyword Set) and additional classification codes (Alternative Code Set) to the patent search criteria. The expanded wind energy domain resulted in an increase in patent count by 7.5%. The expanded wind energy domain observed an increase in recall of 6 percentage points with a 1 percent decrease in precision from the conventional domain. Patents from the New Addition Set, that were present in the expanded domain but not the conventional domain, contained a greater proportion of patents from China and granted patents since 2010. The expanded domain can be utilized in studies that require a robust data set for analysis of innovation activity in wind energy. The methodology used to derive and compare different domain definitions can be generalized for retrieving patents from any technology domain within other patent databases.

5. References

[1] E. Dubarić, D. Giannoccaro, R. Bengtsson, and T. Ackermann, "Patent data as indicators of wind power technology development," *World Patent Information*, vol. 33, no. 2, pp. 144–149, Jun. 2011, doi: 10.1016/j.wpi.2010.12.005.

[2] B. Fox, "The Offshore Grid: The Future of America's Offshore Wind Energy Potential," 2015. [Online]. Available: http://www.renewableuk.

[3] R. Kapoor, M. Karvonen, S. Ranaei, and T. Kässi, "Patent portfolios of European wind industry: New insights using citation categories," *World Patent Information*, vol. 41, pp. 4–10, Jun. 2015, doi: 10.1016/j.wpi.2015.02.002.

[4] Castrejon-Campos, O., Aye, L., & Hui, F. K. P. (2022). Effects of learning curve models on onshore wind and solar PV cost developments in the USA. *Renewable and Sustainable Energy Reviews*, *160*. https://doi.org/10.1016/j.rser.2022.112278

[5] Bento, N., Fontes, M., & Barbosa, J. (2021). Inter-sectoral relations to accelerate the formation of technological innovation systems: Determinants of actors' entry into marine renewable energy technologies. *Technological Forecasting and Social Change*, *173*. https://doi.org/10.1016/j.techfore.2021.121136

[6] Lewis, J. I., & Nemet, G. F. (2021). Assessing learning in low carbon technologies: Toward a more comprehensive approach. In *Wiley Interdisciplinary Reviews: Climate Change* (Vol. 12, Issue 5). John Wiley and Sons Inc. https://doi.org/10.1002/wcc.730

[7] Angelucci, S., Hurtado-albir, F. J., & Volpe, A. (2018). Supporting global initiatives on climate change: The EPO's "Y02-Y04S" tagging scheme. *World Patent Information*, *54*, S85–S92. https://doi.org/10.1016/j.wpi.2017.04.006

[8] Altuntas, F., & Gök, M. Ş. (2020). Technological evolution of wind energy with social network analysis. *Kybernetes*, *July*. https://doi.org/10.1108/K-11-2019-0761

[9] V. Veefkind, J. Hurtado-Albir, S. Angelucci, K. Karachalios, and N. Thumm, "A new EPO classification scheme for climate change mitigation technologies," *World Patent Information*, vol. 34, no. 2, pp. 106–111, Jun. 2012, doi: 10.1016/j.wpi.2011.12.004.

[10] Epo and Uspto, "Guide to the CPC (Cooperative Patent Classification) Guide to the CPC." [Online]. Available: www.cpcinfo.org

[11] Miyamoto, M., & Takeuchi, K. (2019). Climate agreement and technology diffusion: Impact of the Kyoto Protocol on international patent applications for renewable energy technologies. *Energy Policy*, *129*, 1331-1338. https://doi.org/10.1016/j.enpol.2019.02.053

[12] Lindman, Å., & Söderholm, P. (2016). Wind energy and green economy in Europe: Measuring policy-induced innovation using patent data. *Applied Energy*, *179*, 1351-1359. https://doi.org/10.1016/j.apenergy.2015.10.128

[13] Yu, N. Innovation of renewable energy generation technologies at a regional level in China: a study based on patent data analysis. *Int Econ Econ Policy* **14**, 431–448 (2017). https://doi.org/10.1007/s10368-017-0382-6

[14] Johnstone, N., Haščič, I. & Popp, D. Renewable Energy Policies and Technological Innovation: Evidence Based on Patent Counts. *Environ Resource Econ* **45**, 133–155 (2010). https://doi.org/10.1007/s10640-009-9309-1

[15] Doblinger, C., Surana, K., Li, D., Hultman, N., & Anadón, L. D. (2022). How do global manufacturing shifts affect long-term clean energy innovation? A study of wind energy suppliers. *Research Policy*, *51*(7), 104558. https://doi.org/10.1016/j.respol.2022.104558

[16] Dubarić, E., Giannoccaro, D., Bengtsson, R., & Ackermann, T. (2011). Patent data as indicators of wind power technology development. *World Patent Information*, *33*(2), 144-149. https://doi.org/10.1016/j.wpi.2010.12.005

[17] Hain, D. S., Jurowetzki, R., Konda, P., & Oehler, L. (2020). From catching up to industrial leadership: Towards an integrated market-technology perspective. An application of semantic patent-to-patent similarity in the wind and EV Sector. *Industrial and Corporate Change*, *29*(5), 1233–1255. https://doi.org/10.1093/icc/dtaa021

[18] Altuntas, F., Gok, M.S. A data-driven analysis of renewable energy management: a case study of wind energy technology. *Cluster Comput* (2023). https://doi.org/10.1007/s10586-023-03966-3

[19] Schleich, J., Walz, R., & Ragwitz, M. (2017). Effects of policies on patenting in wind-power technologies. *Energy Policy*, *108*, 684-695. https://doi.org/10.1016/j.enpol.2017.06.043

[20] A. Malhotra, T. S. Schmidt, and J. Huenteler, "The role of inter-sectoral learning in knowledge development and diffusion: Case studies on three clean energy technologies," *Technological Forecasting and Social Change*, vol. 146, pp. 464–487, Sep. 2019, doi: 10.1016/j.techfore.2019.04.018.

[21] Y. C. Tsai, Y. F. Huang, and J. T. Yang, "Strategies for the development of offshore wind technology for far-east countries – A point of view from patent analysis," *Renewable and Sustainable Energy Reviews*, vol. 60. Elsevier Ltd, pp. 182–194, Jul. 01, 2016. Doi: 10.1016/j.rser.2016.01.102.

[22] Lee, K., & Lee, S. (2013). Patterns of technological innovation and evolution in the energy sector: A patent-based approach. *Energy Policy*, *59*, 415-432. https://doi.org/10.1016/j.enpol.2013.03.054

[23] Odam, N., & de Vries, F. P. (2020). Innovation modelling and multi-factor learning in wind energy technology. *Energy Economics*, *85*, 104594. https://doi.org/10.1016/j.eneco.2019.104594

[24] EPO/PATSTAT Support, "Coverage of PATSTAT 2020 Spring Edition", Tableau Public.

[25] de Rassenfosse, G., & Seliger, F. (2021). Imputation of missing information in worldwide patent data. *Data in Brief*, *34*, 106615. https://doi.org/10.1016/j.dib.2020.106615

[26] Kang, B., & Tarasconi, G. (2016). PATSTAT revisited: Suggestions for better usage. *World Patent Information*, *46*, 56-63. https://doi.org/10.1016/j.wpi.2016.06.001

[27] Pasimeni, F. (2019). SQL query to increase data accuracy and completeness in PATSTAT. *World Patent Information*, *57*, 1-7. https://doi.org/10.1016/j.wpi.2019.02.001

[28] Veefkind, V., Hurtado-Albir, J., Angelucci, S., Karachalios, K., & Thumm, N. (2012). A new EPO classification scheme for climate change mitigation technologies. *World Patent Information*, *34*(2), 106-111. https://doi.org/10.1016/j.wpi.2011.12.004

Appendix A: Keywords

Table A1. Keyword Comparison Analysis of the Conventional Classification Codes (WEDD1) vs. New Additions.

Keyword	SQL Query	WEDD1	Percentage of WEDD1	New Additions	Percentage of New Additions	Presence in PATSTAT Excluding WEDD1
Wind Power	%wind_power%	19741	23%	3737	48%	16596
Wind turbine	%wind_turbine%	18979	22%	1299	17%	3691
Wind Energy	%wind_energy%	7513	9%	893	11%	5579
Wind Generator	%wind_generat%	5057	6%	853	11%	3944
Wind Farm	%wind_farm%	949	1%	175	2%	367

Windmill	%windmill%	3944	5%	234	3%	3036
Energy of Wind	%energy_of_wind%	164	0%	19	0%	101
Energy from wind	%energy_from_wind%	122	0%	3	0%	19
Wind rotor	%wind_rotor%	322	0%	7	0%	83
Wind axis	%wind_ax_s%	14	0%	15	0%	107
Wind blade	%wind_blade%	292	0%	33	0%	529
Wind generating set	%wind_generating_set%	2610	3%	389	5%	1431
Wind array	%wind_array%	1	0%	0	0%	2

Wind plant	%wind_plant%	90	0%	38	0%	76
Wind park	%wind_park%	306	0%	2	0%	5
Wind platform	%wind_platform%	1	0%	3	0%	9
Wind foundation	%wind_foundation%	0	0%	0	0%	0
Wind substructure	%wind_substructure%	1	0%	0	0%	0
Wind base	%wind_base%	22	0%	7	0%	78
Wind hub	%wind_hub%	0	0%	0	0%	8
Wind nacelle	%wind_nacelle%	1	0%	0	0%	0

Wind gearbox	%wind_gearbox%	0	0%	0	0%	0
Wind control	%wind_control%	25	0%	30	0%	588
Wind installation	%wind_installation%	17	0%	2	0%	18

APPENDIX B: Classification Codes

Table B1. Expanded Code List.

Classification Code	**Subgroup Description**	**"YES" Occurrence Percentage**	**IPC, CPC, or BOTH**
H02J	CIRCUIT ARRANGEMENTS OR SYSTEMS FOR SUPPLYING OR DISTRIBUTING ELECTRIC POWER; SYSTEMS FOR STORING ELECTRIC ENERGY	14%	BOTH
H02P	CONTROL OR REGULATION OF ELECTRIC MOTORS, ELECTRIC GENERATORS OR DYNAMO-ELECTRIC CONVERTERS; CONTROLLING TRANSFORMERS, REACTORS OR CHOKE COILS	8%	BOTH

G01M	TESTING STATIC OR DYNAMIC BALANCE OF MACHINES OR STRUCTURES; TESTING OF STRUCTURES OR APPARATUS, NOT OTHERWISE PROVIDED FOR	7%	BOTH
H02M	APPARATUS FOR CONVERSION BETWEEN AC AND AC, BETWEEN AC AND DC, OR BETWEEN DC AND DC, AND FOR USE WITH MAINS OR SIMILAR POWER SUPPLY SYSTEMS; CONVERSION OF DC OR AC INPUT POWER INTO SURGE OUTPUT POWER; CONTROL OR REGULATION THEREOF	6%	BOTH
G06Q	DATA PROCESSING SYSTEMS OR METHODS, SPECIALLY ADAPTED FOR ADMINISTRATIVE, COMMERCIAL, FINANCIAL, MANAGERIAL, SUPERVISORY OR FORECASTING PURPOSES; SYSTEMS OR METHODS SPECIALLY ADAPTED FOR ADMINISTRATIVE, COMMERCIAL, FINANCIAL, MANAGERIAL, SUPERVISORY OR FORECASTING PURPOSES, NOT OTHERWISE PROVIDED FOR	6%	BOTH
G06F	ELECTRIC DIGITAL DATA PROCESSING	5%	BOTH
F16H	GEARING	5%	BOTH
E02D	FOUNDATIONS; EXCAVATIONS; EMBANKMENTS; UNDERGROUND OR UNDERWATER STRUCTURES	5%	BOTH
C22C	ALLOYS	4%	BOTH

B63B	SHIPS OR OTHER WATERBORNE VESSELS; EQUIPMENT FOR SHIPPING	4%	BOTH
B29C	SHAPING OR JOINING OF PLASTICS; SHAPING OF MATERIAL IN A PLASTIC STATE, NOT OTHERWISE PROVIDED FOR; AFTER-TREATMENT OF THE SHAPED PRODUCTS, e.g. REPAIRING	4%	BOTH
E04H	BUILDINGS OR LIKE STRUCTURES FOR PARTICULAR PURPOSES; SWIMMING OR SPLASH BATHS OR POOLS; MASTS; FENCING; TENTS OR CANOPIES, IN GENERAL	4%	BOTH
H02B	BOARDS, SUBSTATIONS, OR SWITCHING ARRANGEMENTS FOR THE SUPPLY OR DISTRIBUTION OF ELECTRIC POWER	4%	BOTH
H05K	PRINTED CIRCUITS; CASINGS OR CONSTRUCTIONAL DETAILS OF ELECTRIC APPARATUS; MANUFACTURE OF ASSEMBLAGES OF ELECTRICAL COMPONENTS	4%	BOTH
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	4%	CPC

Table B2. Excluded Code List.

Classification Code	Subgroup Description	"NO" Occurrence Percentage	IPC, CPC, or BOTH
Y02B	CLIMATE CHANGE MITIGATION TECHNOLOGIES RELATED TO BUILDINGS, e.g. HOUSING, HOUSE APPLIANCES OR RELATED END-USER APPLICATIONS[KF1]	11%	CPC
F24F	AIR-CONDITIONING; AIR-HUMIDIFICATION; VENTILATION; USE OF AIR CURRENTS FOR SCREENING	10%	BOTH
F21S	NON-PORTABLE LIGHTING DEVICES; SYSTEMS THEREOF; VEHICLE LIGHTING DEVICES SPECIALLY ADAPTED FOR VEHICLE EXTERIORS	8%	IPC
Y02E	REDUCTION OF GREENHOUSE GAS [GHG] EMISSIONS, RELATED TO ENERGY GENERATION, TRANSMISSION OR DISTRIBUTION	7%	CPC
F21V	FUNCTIONAL FEATURES OR DETAILS OF LIGHTING DEVICES OR SYSTEMS THEREOF; STRUCTURAL COMBINATIONS OF LIGHTING DEVICES WITH OTHER ARTICLES, NOT OTHERWISE PROVIDED FOR	7%	BOTH

F21W	RELATING TO USES OR APPLICATIONS OF LIGHTING DEVICES OR SYSTEMS	6%	BOTH
F04D	NON-POSITIVE-DISPLACEMENT PUMPS	5%	BOTH
C02F	TREATMENT OF WATER, WASTE WATER, SEWAGE, OR SLUDGE	5%	BOTH
F21Y	RELATING TO THE FORM OR THE KIND OF THE LIGHT SOURCES OR OF THE COLOUR OF THE LIGHT EMITTED	5%	BOTH
Y02T	CLIMATE CHANGE MITIGATION TECHNOLOGIES RELATED TO TRANSPORTATION	4%	CPC

Table B3. Confusion Matrix for Inclusion of Code Filter 1

		Actual class		
		Wind	Not wind	
Predicted class	Wind	65	26	91
	Not wind	44	122	166
		109	148	257

Table B4. Confusion Matrix for Inclusion of Code Filter 2

		Actual class		
		Wind	Not wind	
Predicted class	Wind	69	26	95
	Not wind	40	122	162
		109	148	257

Table B5. Confusion Matrix for Code Filter 3

		Actual class		
		Wind	Not wind	
Predicted class	Wind	61	18	79
	Not wind	48	130	178
		109	148	257

Table B6. Comparing Precision and Recall Rate of 3 Different Alternative Code List Sets

	Code Filter Criteria	**Precision Rate**	**Recall Rate**
Code Filter 1	Inclusion of Malhotra Code List	71%	60%
Code Filter 2	Inclusion of 15 Relevant Codes from Occurrence Analysis	73%	63%
Code Filter 3	Inclusion of 15 relevant codes from occurrence analysis and exclusion of 10 irrelevant codes (Alternative Code Set)	77%	56%

	Manual Review of Alternative Code Set	83%	39%

APPENDIX C: Sampling

Equation C1. Sample Size Determination

$$n = Z_{0.95}^2 \frac{P(1-P)}{E^2} \qquad \text{(Eq. C1)}$$

n = Sample size

P = Population proportion

E = Margin of error

Z = Z score

For the application to wind energy in Section 3, the population proportion (P) is assumed to be 50% since determining whether a patent is related or unrelated to the technology in question is a binary outcome. This assumption also results in the largest sample size for the population. The margin of error is then tested at 5% and 10%. Furthermore, the Z-score is determined based on an assumed confidence level of 0.95. After completing both calculations, the resulting sample sizes are 96 for a margin of error at 10% and 384 for a margin of error at 5%.

Table C1. Manual Review Sample Results

Sample Number	Patent Criteria	Number of Patents Sampled	Yes	No	Unsure
1	Any of the Conventional Classification Codes (WEDD1)	100	98%	2%	0%
2	Keyword Set - WEDD1	257	42%	67%	1%
3	(Keyword Set – WEDD1) – Code Set	100	32%	67%	0%
4	New Additions	100	83%	17%	0%

The number of patents used in all four samples on the wind energy application are outlined in Table C1. The sample size used for Sample 2 (Keyword Set – WEDD1) is 257, which is between 5-10% margin of error. The sample size used for this criterion originally started at 300 patents, but after the evaluations, only granted patents were included in the considerations for the study. The number of patents used in Sample 1 (Any of the Conventional Classification Codes (WEDD1), Sample 3 ((Keyword Set – WEDD1) – Code Set), and Sample 4 (New Additions) are all 100 patents, which is aligned with an estimated margin of error at 10%.

APPENDIX D: Comparing Domain Definitions

Table D1. Confusion Matrix for WEDD1

		Actual class		
		Wind	Not wind	
Predicted class	Wind	84,682	1,728	86,410
	Not wind	13,731	the rest of PATSTAT	
		98,413		

Table D2. Confusion Matrix for Confirmed Code Set on Keyword Additions Set

		Actual class		
		Wind	Not wind	
Predicted class	Wind	5,414	1,109	6,523
	Not wind	8,317	16,885	25,202
		13,731	17,994	31,725

Table D3. Confusion Matrix for WEDD3

		Actual class		
		Wind	Not wind	
Predicted class	Wind	90,096	2,837	92,933
	Not wind	8,317	the rest of PATSTAT	
		98,413		

APPENDIX E: Wind Energy Patent Reviewers

The wind energy experts who reviewed the patents in this study are UMass Wind Energy Fellows, who are PhD Candidates at the University of Massachusetts Amherst. The patent

reviewers were selected on a volunteer basis. Two of the four patent reviewers suggested keywords for this study after they completed the patent review process.

Each volunteer was given the same protocol to follow for the review process. The protocol included written instructions with an excel sheet of 100 randomly selected patents with twenty-five patents that overlapped all volunteer samples. Volunteers were not notified about the overlapping patent set.

The protocol given asked each volunteer to review a patent's title and abstract and assign whether the patent is related or unrelated to wind energy based on whether it encompasses grid-connected stationary electricity generation powered by wind, either offshore or onshore. If it was evident from the title and abstract that the application is wind energy related, then the volunteer marked the patent as being related to wind energy and moved on to the next patent. However, if the volunteer marked the patent as being not related to wind energy, then the volunteer was asked to provide a short justification for their evaluation. If it was unclear as to whether the patent was related or not related to wind energy, instructions were given to obtain the full text of the patent for more context, where the reviewer was asked to reevaluate the patent again based on their best judgement.

APPENDIX F: Application on Geothermal Energy

We applied a limited version of our approach (without expert verification) to geothermal energy. The classification codes and keywords are identified from Albino, et al. (2014) and by searching for geothermal patent instances. Table F1 showed results using conventional codes versus our approach. Table F2 listed the conventional codes, keywords, and alternative codes.

Figure F1. Comparing percentage of patent count using conventional classification codes and New Additions by Patent Authority for Geothermal Energy

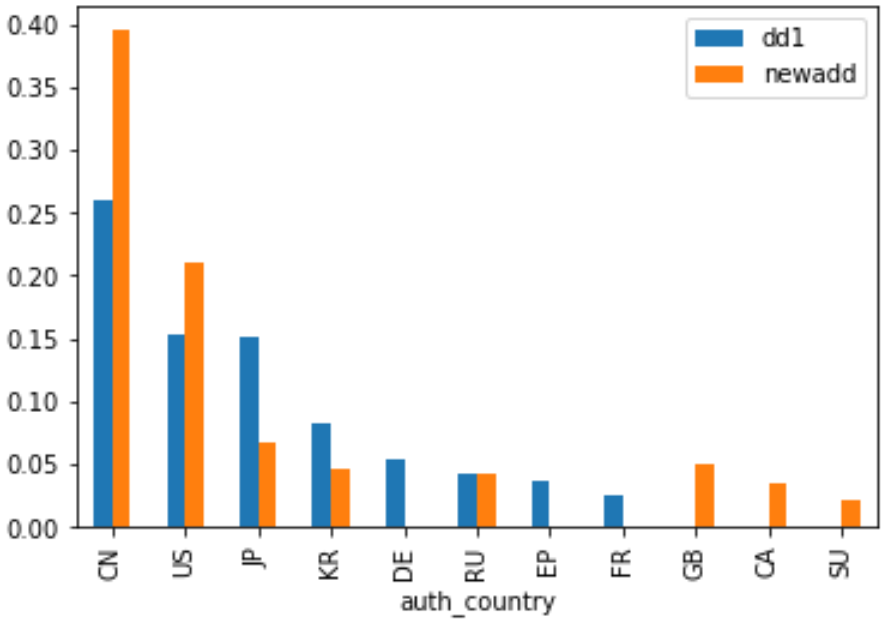


Figure F2. Comparing percentage of patent count using conventional classification codes and New Additions by Earliest Publication Year for Geothermal Energy

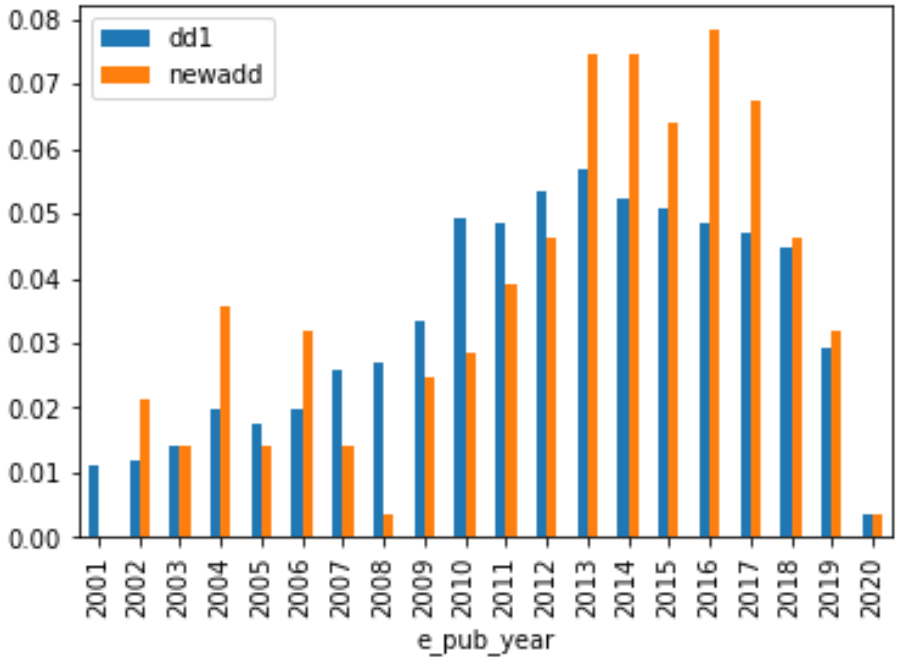


Table F1. Descriptive statistics for DD1, DD2 and DD3 of Geothermal Energy.

	DD1	DD2	DD3
Patent Criteria	Any of the conventional classification codes	Any of the conventional classification codes or targeted keywords	Any of the conventional classification codes, or a combination of targeted keywords and additional classifications
Number of Granted Patents	13,023	140,311	13,304
Average Year Granted	2002	2002	2002

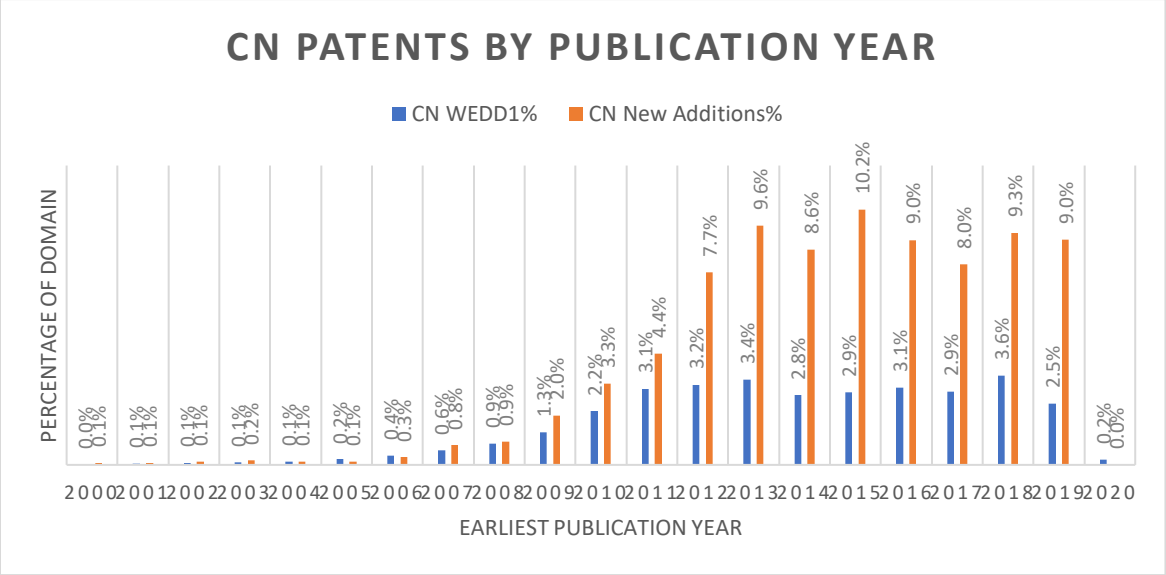
Number of Granted Patents since 2010	6,299	75,097	6,455
Percent of Granted Patents since 2010	48%	54%	49%
Number of Granted Patents from China	3,382	67,951	3,493
Percent of Granted Patents from China	26%	48%	26%

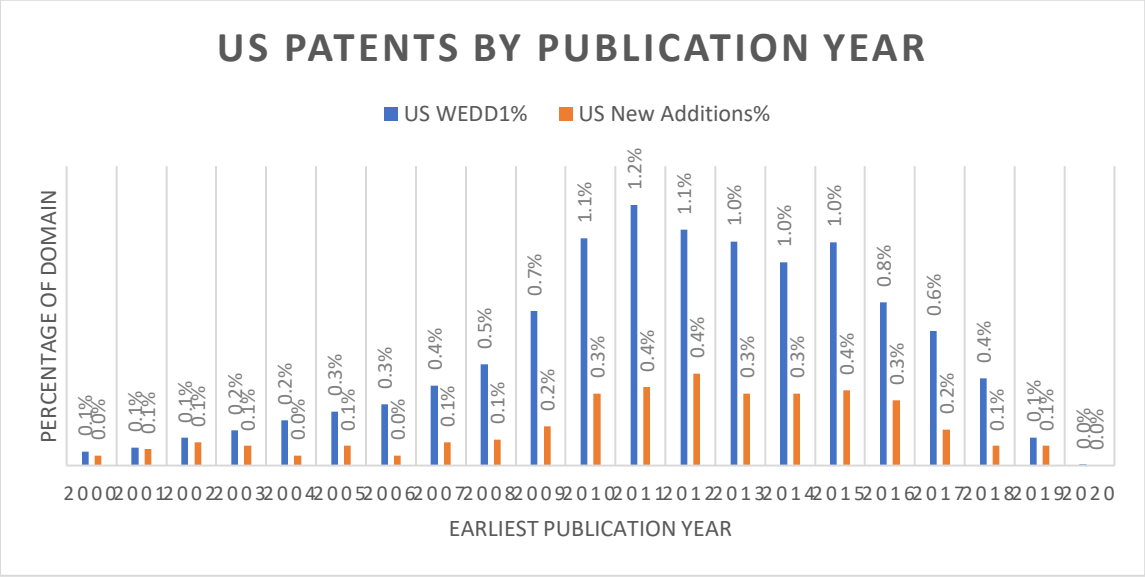
Table F2. Conventional Codes, Keywords and Alternative Codes for Geothermal Energy Definition

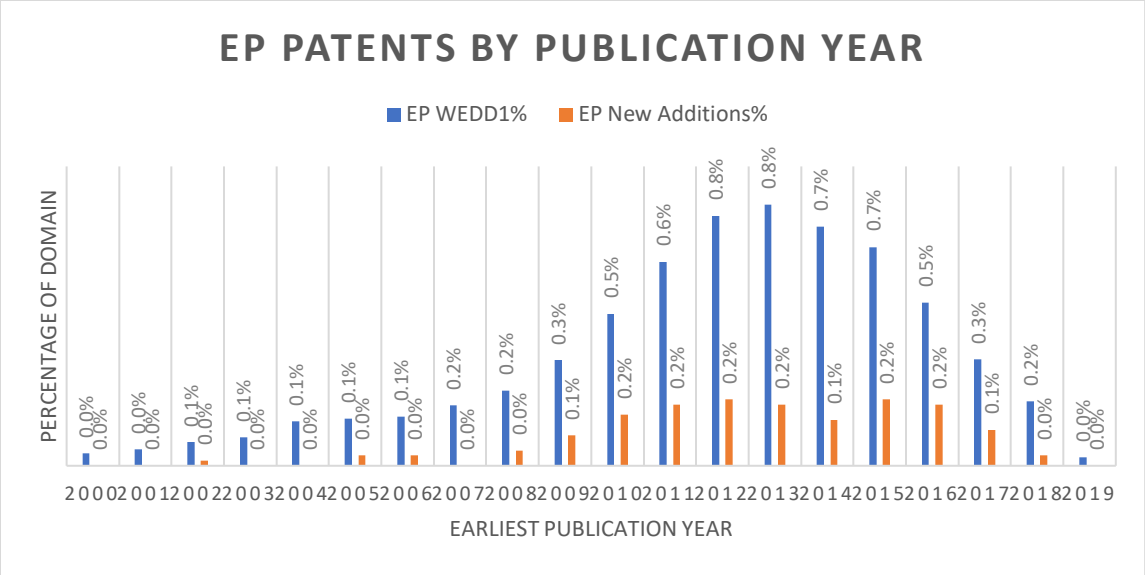
Conventional codes	**F24T** geothermal collectors; geothermal systems **F03G 4** Devices for producing mechanical power from geothermal energy **F03G 7/04** Mechanical-power-producing mechanisms using pressure differences or thermal differences occurring in nature **F24J 3** Other production or use of heat, not derived from combustion, using natural heat/geothermal heat **F24F2005/0057** Air-conditioning systems or apparatus receiving heat-exchange fluid from a closed circuit in the ground (cpc only) **F17C2227/032** Heat exchange with the fluid using geothermal water (cpc only) **Y02B10/40** Geothermal heat-pumps (cpc only) **Y02E10/10** Geothermal energy (cpc only)
Keyword list	**Geothermal; Hydrothermal; Geo-heat; "Natural heat"; "Ground heat/thermal"; "Earth heat/thermal"**
Alternative code set	**F01K** Steam engine plants; steam accumulators; engine plants not otherwise provided for; engines using special working fluids or cycles **H02N 10** Electric motors using thermal effects

APPENDIX G: Patent by Publication Year for Each Patent Authority

Figure G1. Comparing percentage of patent count using conventional classification codes and New Additions by earliest publication years from 2000-2020 for each Patent Authority (China, US, and EPO).






EP PATENTS BY PUBLICATION YEAR
EP WEDD1%
EP New Additions%
PERCENTAGE OF DOMAIN
EARLIEST PUBLICATION YEAR
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
0.0% 0.0% 0.0% 0.0% 0.1% 0.0% 0.1% 0.0% 0.1% 0.0% 0.1% 0.0% 0.1% 0.0% 0.2% 0.0% 0.2% 0.0% 0.3% 0.1% 0.5% 0.2% 0.6% 0.2% 0.8% 0.2% 0.8% 0.2% 0.7% 0.1% 0.7% 0.2% 0.5% 0.2% 0.3% 0.1% 0.2% 0.0% 0.0% 0.0%