What are the main determinants of diffusion speed between academic research and industrial application? Evidence from four major clean-energy technologies

Benedict Probst (bsp26@cam.ac.uk)
University of Cambridge https://orcid.org/0000-0002-1149-8938
Laura Díaz Anadón (lda24@cam.ac.uk)
University of Cambridge
Andreas Kontoleon (ak219@cam.ac.uk)
University of Cambridge https://orcid.org/0000-0003-4769-898X

Research Article

Keywords: Clean technology innovation, knowledge spillovers, technology diffusion

Posted Date: March 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-343866/v1

What are the main determinants of diffusion speed between academic research and industrial application? Evidence from four major clean-energy technologies

Benedict Probst[1*] (bsp26@cam.ac.uk), Laura Diaz Anadon[1*] (lda24@cam.ac.uk), Andreas Kontoleon[1*] (ak219@cam.ac.uk)

1: Cambridge Centre for Environment, Energy and Natural Resource Governance (C-EENRG), Department of Land Economy, University of Cambridge

* Corresponding Authors

Abstract:

Recent evidence suggests a slowdown of economic productivity in major Western and Asian economies. One of the most convincing causes is the slowdown in research productivity in key sectors of the economy, such as low-carbon technologies. The latter trend is particularly worrying as low-carbon technologies play a critical role in keeping global warming well below the 2°C that the Paris Agreement set. We rely on a novel data science method that connects scientific articles with patented technologies. We extract the scientific publications cited in more than 600,000 clean energy technologies (wind, solar, biomass, li-ion) and investigate what determines the diffusion speed between scientific research and patented technologies. We demonstrate that the higher the quality of the scientific article (measured by citations), the lower the distance between scientists and inventors, and the higher the similarity between the content of the scientific article and the patent, the faster the diffusion between research and application. Yet, we also show that while more dissimilar content takes longer to be used in patents, the eventual impact of the patent is greater, possibly because it is more innovative. Our data also reveals that while distance appears to matter for the speed of knowledge diffusion, patents in the four low-carbon technologies on average rely on 81% of foreign sources of science, as scientific knowledge diffuses widely across the world economy. China and the United States play an outsized role as the source of scientific publications used in clean-technology patents globally. Nevertheless, while research is characterised by global spillovers, the application of such knowledge (in a patent) appears to be dominated by national teams, potentially due to greater local spillovers and secrecy issues.

Key words: Clean technology innovation, knowledge spillovers, technology diffusion

1. Introduction

Recent evidence indicates a productivity[1] slowdown in major Western and Asian economies (Goldin *et al.*, 2018). Among the most convincing reasons is an overall decline in research productivity in critical economic sectors, such as semi-conductors. For instance, as Bloom et al. (2018) show that it takes around 18 times as many researchers today compared to the early 1970s to achieve Moore's Law (the famous doubling of the density of computer chips every two years). Yet, this worrying trend is not only limited to semi-conductors but has also been observed in agricultural crop yields, life sciences, and other sectors (Bloom *et al.*, 2017).

Patent data also shows a slowing of innovation in major clean technologies (relative to overall patenting trends), likely also driven by declining fossil fuel prices reducing the savings from low-carbon innovation (i.e., energy-saving technologies) (Popp *et al.*, 2020; Probst *et al.*, 2021). It is clear from climate-economy models and empirical research, however, that a delayed availability of clean technologies, due to lack of innovation or diffusion, could be even costlier than delayed climate policies (Luderer *et al.*, 2012; Probst *et al.*, 2020). This is particularly true for low-carbon technologies, whose availability and price play a critical part in minimising the overall cost of climate change mitigation.

Hence, understanding the lag time between knowledge diffusion from scientific research to technological applications is critical in devising appropriate innovation policy to accelerate technological development in important low-carbon sectors as the market naturally underprovides such incentives (Wilson and Grubler, 2011). However, the majority of existing studies focus on horizontal (rather than vertical spillovers). Horizontal spillovers refer to the application of knowledge at the same technological level, whereas vertical spillovers occur between different levels of development (e.g., scientific research and patented technologies). Existing approaches to study vertical spillovers are largely limited to case-study evidence and interviews (Agrawal and Henderson, 2002), which is difficult to scale across different technological areas and research institutes. In addition, several existing studies focus on the likelihood that academic research is used in patented technologies, but not on the speed of knowledge diffusion (Probst, Kontoleon and Anadón, 2021).

[1] Defined as "the rate at which inputs are turned into outputs" (Goldin *et al.*, 2018; p.2), and so research productivity is the rate at which research inputs (labour and capital) are transformed into outputs (such as academic papers).

In this article, we, therefore, investigate what determines the diffusion speed between scientific publications and their commercial application by analysing the bibliometric and geographic characteristics of scientific publications that patents cite. These paper-patent citation linkages often present a better indicator of knowledge flows than patent-to-patent citations as patent-patent citations are more often added by patent examiners (and not the inventor) (Popp, 2016). By using the new approach described in detail in Probst, Kontoleon and Anadón (2021), which combines econometrics, text-mining, and machine-learning, we are able to link previously unanalysed data on a large scale. This allows us to provide unique insights into the diffusion speed and geographical patterns of key clean technologies needed to mitigate climate change. Moreover, this approach can also be used in other areas outside of clean technology research and provides a promising research avenue in other fields, such as biotechnology and pharmaceuticals.

We next review the relevant parts of the literature (Section 2), discuss the methods (Section 3) and data (Section 4). In Section 5 we discuss our findings and conclude in Section 6.

2. Literature review

Various studies show that knowledge and technological spillovers play a significant role in technological and economic development (Griliches, 1992; Dechezleprêtre, Martin and Mohnen, 2013; Stephan *et al.*, 2017). Yet, the main focus of the existing literature is on horizontal (not vertical) knowledge and technological spillovers. Horizontal spillovers refer to the application of knowledge at the same technological level. For instance, commonly citations between different patents are used – so-called prior art – to understand technological trajectories (Dechezleprêtre *et al.*, 2011).

Despite the absence of a tractable large-scale approach to understanding the diffusion of scientific research to technological applications, the academic literature generally underscores the importance of basic science (financed largely by public R&D) for industrial progress and economic development (Toole, 2012; Li, Azoulay and Sampat, 2017). The central question that emerges is how to make basic science matter most and most quickly for the areas of society's central challenges. It is clear that in certain cases basic research has laid the groundwork for later pathbreaking applications. For instance, Riemann's advances in differential geometry, which were initially ignored, later became essential for Einstein's

general relativity theory and are now critical for the ubiquitous Global Positioning Systems (GPS).

Hence, public R&D and basic science are important contributors to economic progress due to substantial intra- and inter-industry spillovers. Yet, the literature points to several factors that may increase the spillovers from public R&D and government-funded science to industrial application. Section 2.1 sets out the common challenges of measuring the diffusion lag between public R&D, scientific research, and technological application. Section 2.2 then describes the effects of technological characteristics on diffusion speed, whereas Section 2.3 reviews the literature that investigates the impact of geographic distance on diffusion speed.

2.1. Lag between public R&D funding, scientific publications, and technological application

Evaluating the effect of funding on scientific progress and technological innovation is not straightforward. It normally takes many years from initial public R&D funding to industrial application. Yet, the lag differs substantially between industries. In medical science, it may take 17-24 years from the development of a new drug to trials with mice, humans and final approval through FDA or another regulatory agency (Toole, 2012) although the recent development time of various Covid-19 vaccines show that much faster development times are possible. In sectors such as computer science and machine-learning these lags tend to be much shorter, but are also harder to study as computer code is commonly not patentable and rather protected by secrecy. Research on the lag between public clean-energy R&D and subsequent scientific publications indicates that it takes up to 10 years from initial funding to the publication of the research and another few years to the publication of the patent (Popp, 2016). Yet, while Popp's (2016) analysis is intriguing it relies on high-level correlations without investigating the microeconomic foundations.

Another challenge to evaluating the effect of R&D is that many concurrent factors play a role, such as changes in private R&D funding, policy, and regulatory environment, factor prices (labour and energy prices), among many other factors. Hence it is difficult to attribute changes in the metric of interest – such as the increase in scientific papers, drugs, or final technologies – to one single factor, such as changes in R&D funding.

In the absence of discontinuities, randomisation or other (quasi-)experimental approaches cannot be used to isolate cause and effect. Some use funding rules to create

instrumental variables to get around this problem, which requires in-depth knowledge of the funding procedure, and thus restricts applicability to narrower domains (commonly one funding body and one technology). For instance, Azoulay *et al.* (2019) use idiosyncratic[2] rigidities in funding rules of the U.S. National Institutes of Health (NIH) grants over 27 years to evaluate the impact of research investment on patenting. They find that around 10% of investment directly generates a patent (with an average lag of around 10 years), whereas 30% lead to scientific articles that are subsequently cited. Hence, their findings indicate that merely evaluating researchers on their patenting activity may only provide part of scientists' impact.

2.2. Effect of technological and scientific paper characteristics on diffusion

The existing literature has underscored the cumulative process of knowledge production as knowledge created within and outside a given domain contributes to technological progress (Battke *et al.*, 2016). Generally, knowledge diffusion within the same domain is the norm, as the knowledge developed for a specific domain is typically more directly applicable. Hence, for instance, around 60% of patent citations to renewable energy technologies come from patents in the same technological field (Noailly and Shestalova, 2017). Yet, inter-domain knowledge spillovers have also been shown to play an integral role in technological progress (Battke *et al.*, 2016), and may even lead to fundamental breaks in a technological trajectory (Stephan *et al.*, 2017). For instance, empirical evidence from China using survey data suggests that firms relying on a broader knowledge base are more likely to develop breakthrough technologies (Zhou and Li, 2012).

Technological characteristics have been shown to play an integral role in determining the proportion of internal (within the same technological domain) and external (from outside the domain) spillovers. Using evidence from 40,000 battery patents, Battke *et al.* (2016) show that knowledge that is based on more diverse previous knowledge is more likely to flow across different technological classes. In contrast, knowledge that is based on more homogeneous previous knowledge is more likely to diffuse within the same technological class. Similarly, core-knowledge (focused on the central components of a technology) is more likely to remain

[2] Azoulay *et al.* (2019) use an instrument that relies on differences between allocated funding in different sub-fields in medicine (which creates natural variation around funding cut-offs), which they assert is not correlated to the innovation potential of the proposed project.

within the same domain, whereas peripheral knowledge (focused on peripheral components) is more likely to diffuse across different domains.

Scientific paper characteristics have also been shown to play a role in the diffusion of knowledge. For instance, academic studies published by larger research teams are generally associated with more citations from other scientific studies, which is not primarily caused by self-citation (Larivière *et al.*, 2015). In addition, whether studies can be accessed via a paywall (i.e., open access) may have further citation advantages, even when adequately controlling for quality (e.g., because authors may self-select their highest quality work into the open-access category to further increase the impact) (Probst, Kontoleon and Anadón, 2021).

2.3. Effect of geographical distance

The literature generally shows that knowledge diffusion decreases sharply with distance. This is demonstrated by empirical evidence on the effect of geographical distance on different patterns of patent citations (Jaffe, Trajtenberg and Handerson, 1993), R&D and patenting (Branstetter, 2006), productivity, and the sales performance of subsidiaries of multinational firms (Keller and Yeaple, 2013). For instance, Peri and Bottazzi (2003) using R&D and patent data from the EPO, find that knowledge spillovers measured by patent citations are very localised and occur typically within 300 kilometres. These geographical characteristics are critical for economic development, as research by Keller (2004) demonstrates that foreign sources of technology account for more than 90% in the productivity growth of many developing countries.

This localised nature of knowledge has spurred the interest of researchers and policymakers in policies that specifically address and harness the geographic clustering of industries (which relies on localised spillovers). Well-known examples of clusters are Silicon Valley and car manufacturing clusters in Southern Germany, among others, which formed with different levels of public support. Yet, empirical evidence on the effect of these clustering policies is scant. Recent evidence from Germany suggests that cluster-based policies increase the innovative performance between 5.1-11.2 percent of firms within that cluster (compared to a counterfactual group, using differences-in-differences (Falck, Heblich and Kipar, 2010)). In contrast, McDonald, Tsagdis and Huang (2006) uses evidence from 43 European industrial clusters and finds no evidence on the effectiveness of government policies on the growth of these clusters.

While the aforementioned research highlights the importance of technological characteristics and geographical distance on knowledge spillovers, none of these studies investigate to what extent these matter for publication-patent spillovers. While studies on public R&D spillovers have highlighted the importance of these spillovers for economic growth, these remain limited to high-level correlational data. With the approach developed in Probst, Kontoleon and Anadón (2021), we directly link academic research and patented technologies to directly examine which factors drive the application of scientific knowledge in patented technologies. As technological characteristics and geographical distance appear to matter for knowledge diffusion, we investigate these factors specifically in our econometric models, which are described in Section 5. We do not study the effect of different public R&D funding bodies, as it is difficult to obtain accurate data on various funding programmes. Yet, with more and better-quality data in this regard, this is an interesting extension of our research.

3. Methods

Based on the literature in Section 2 (highlighting the transfer channels from scientific research to technological applications and main factors affecting the transfer channels), in this section, we set out how we develop a novel empirical strategy that enables us to compute the characteristics of scientific papers and patents that are expected to influence the uptake of scientific knowledge in patented technologies. We specifically focus on content similarity (3.1) and geographical distance (3.2) of patents and scientific publications. We rely on 600,000 patents from the European Patent Database PATSTAT and extract the scientific articles that are referenced in patents (see Section 4 for an in-depth discussion).

3.1. Content similarity

Previous approaches in measuring the similarity between documents have relied on using the diversity in technological classes of patents that other patents cite (Battke *et al.*, 2016). As patents are commonly filed into different technological categories, these can be used to study whether a patent cites patents from the same technological class or another technological class. While this research has led to interesting findings, a lot of relevant data from the title or the patent abstract are left untouched. Yet, these may hold important information on the patent itself and can be used for topic modelling, which does not classify a patent into one category,

but uses the probability of a patent of falling into different categories (thereby, potentially leading to a more refined and nuanced classification than merely using patent classifications). Patents are normally classified by patent examiners into one or several categories, whereas topic models describe a patent as a mix of various (potentially, unlimited) latent topics. We use the title of patents and scientific papers to understand the topic distribution in each document using unsupervised machine-learning algorithms that are described in more detail below.

In order to study how different or similar patents and scientific articles are we proceed in the following steps:

1. Download all relevant scientific papers via the APIs of the relevant publishers
2. Run the CERMINE algorithm on all PDFs (leave XML files as they are, as these are in the required format)
3. Extract the titles of all papers via text-mining using the XML structure to locate the abstracts.
4. Extract the titles of all patents using the European Patent Office's PATSTAT database
5. Use API of Google Translate to translate all patent and scientific paper titles that are not English (predominantly European languages, such as French, but also Japanese and Mandarin) into English
6. Run the Latent Dirichlet Algorithm on the titles of both papers and patents to determine the topic distribution in the titles
7. Compute the Hellinger Distance (elaborated below) between all scientific papers and patents to determine how topically similar they are
8. Use Hellinger Distance to compute the similarity between two topic distributions

Figure 1 provides an overview of the different steps, which are described in more detail below.

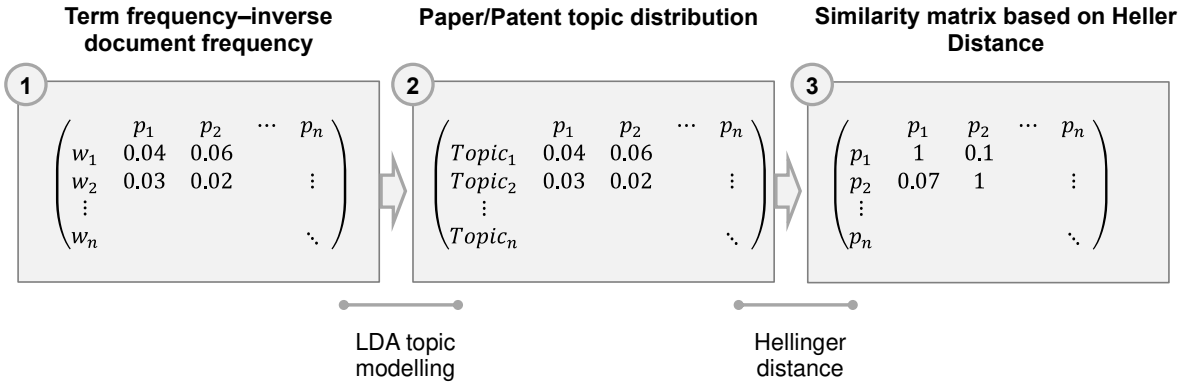

Figure 1: Calculation of similarity matrix between documents based on Heller Distance using a three-step procedure. *Note: p indexes a paper/patent, w is a word, and n indexes the number of the paper, patents, or topics. While two identical documents have a Hellinger distance of 0, this is transformed via the similarity measure S(P,Q) = 1 – H(P,Q). Source: author, based on Kim, Park and Yoon (2016)*

Computing the term frequency-inverse document frequency (tf-idf) consists of two parts: the term frequency $tf\ (t,d)$ and the inverse document frequency $idf\ (t,D)$ (Hogenboom, Capelle and Moerland, 2014). First, the term frequency computes the frequency n of terms $t \in T$ in document $d \in D$. T constitutes all terms, and D all documents in the corpus of interest. The term frequency is calculated with equation (1):

$$tf\ (t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \qquad (1)$$

The inverse-document frequency computes the frequency of a term t in all documents D of a given corpus. This measure corrects for words that occur frequently across all documents (such as words like 'the' that add limited additional information). The logarithmically-scaled inverse document frequency is shown in equation (2)

$$idf\ (t,D) = \log \frac{|D|}{|\{d \in D : t\ \in d\}|} \qquad (2)$$

where D is the number of all documents and $|\{d \in D : t\ \in d\}|$ is the number of documents that contain term t. The td-idf is a metric that determines how important a given word is in a document relying on the two previous calculations. The td-idf is then calculated in equation 3:

$$tfidf\ (t, d, D) = tf\ (t, d) * idf\ (t, D) \qquad (3)$$

In a second step, the topic distribution for each scientific paper or patent is computed. Several algorithms implement topic modelling, but we rely on the latent Dirichlet allocation (LDA) algorithm, which is widely used and performs well when benchmarked to other algorithms in the same class. LDA is an unsupervised machine-learning algorithm that assumes that a document was generated over a mixture of latent topics, with each topic being represented by several words. For instance, two topics (politics and vacation) will have different words that represent each topic. For instance, words in politics could be voter, election, and campaign, whereas words in the vacation topic could be beach, flight, and hotel. Words that do not have a strong topical meaning (such as 'the' or 'a') will have roughly the same likelihood of occurring in any of the two topics.

The algorithm is commonly illustrated using plate notation to show interdependencies among variables in the model (Figure 2). The grey circle in the middle *w* is the only variable that can be observed, which are the words *w* in the documents *D*, all other variables are unobservable (or latent) and need to be computed through the model. The inner plate *N* represents the numerical count of words in a document, and the outer plate *M* the number of documents. The topic of the *k*-th word in document *i* is represented by $z_{i,k}$, and each document i has a topic distributionθ_i and α and β are parameters of the sparse Dirichlet prior, a continuous multivariate probability distribution. The number of topics is fixed and exogenous in the model. The appropriate number of topics can be computed via perplexity scores. We use the same number of topics for this article, as we assume that the latent topics in both datasets are the same (we also corroborate these results via modelling perplexity scores again for this dataset, which indicates the same number of k is appropriate).

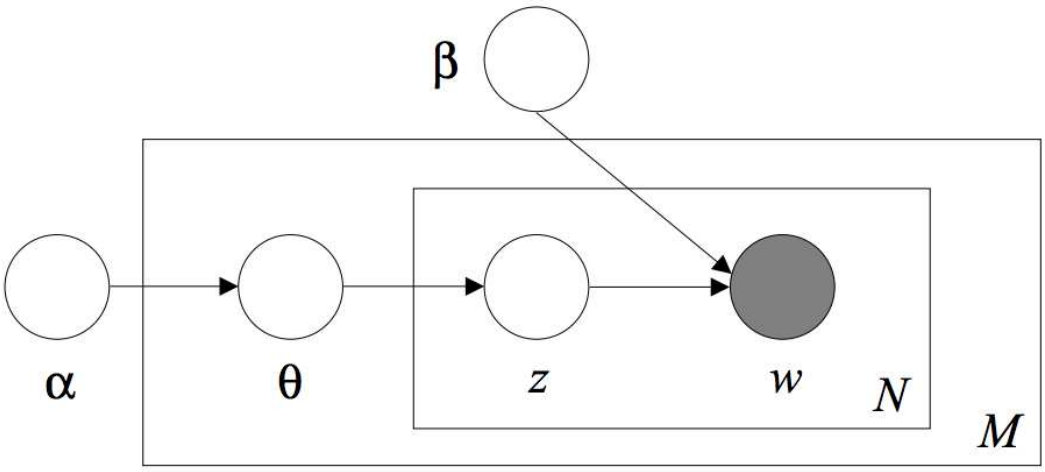

Figure 2: Graphical representation of Latent Dirichlet Allocation (LDA) algorithm. *Source:* Blei *et al.* (2003)

Different distance metrics can be used for computing the distance between documents in a given topic space created through an LDA algorithm. These approaches include the Jensen-Shannon, Hellinger ,and Bhattacharyya distance (Chung *et al.*, 1989). All of these distance metrics compute the divergence between two probability distributions, which has received substantial attention over the last six decades (Kim, Park and Yoon, 2016). As topic modelling assigns each document a probability distribution of different topics, these distance measures compute the similarity between two document probability vectors. As the distance measures provide relatively similar results (Chung *et al.*, 1989), we opt for the Heller distance as it is most widely used in LDA modelling and software implementation.

The Hellinger distance $0 \leq H(PQ) \leq 1$ between two discrete probability distributions P = (p_1,…, p_k) and Q = (q_1,…, q_k) , is succinctly defined in equation (4):

$$H^2(P,Q) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2 \qquad (4)$$

where *H* is the Hellinger distance between two continuous probability distributions *P and Q.* For a full exposition and derivation of the formula, see Hellinger (1909). The resulting output is a similarity matrix between two documents, which is shown in Figure 1, where a distance of

0 indicates an equal distribution of topics between two documents. In line with Kim, Park and Yoon (2016) we then define a similarity measure between the probability distributions P and Q in equation (5).

$$S(P,Q) = 1 - H\,(P,Q) \qquad (5)$$

3.2. Geographical distance

As discussed in the section on geographical distance, understanding the geographical configuration of technological systems has long interested researchers (Sturgeon, Van Biesebroeck and Gereffi, 2008).

In order to study how close researchers and scientists are co-located in space, we use the following measure to determine to what extent geographical distance impacts the use of scientific knowledge in patents. We proceed as follows:

1. Directly extract the geographical location of inventors from the patent data. We rely on PATSTAT and additional data provided by ECOOM, which is a consortium of Flemish University and Eurostat, which substantially improves coverage
2. Use the scientific articles (XML-transformed through CERMINE algorithm) to directly extract the researcher' location via text-mining
3. As this is unstructured data (sometimes misspelled, sometimes the same country is written differently), we use the R-package 'countrycode' to iterate over all countries to convert them into iso2c code.
4. However, this information is not commonly available for all inventors as only 44% of patents have at least one inventor with detailed geographic information, whereas around 88% of patents have at least one inventor with a country code. Hence, we analyse whether the inventor and scientist come from the same country, which we include in our regression as a variable. Similarly, the retrieval rate of the CERMINE for affiliation data from the scientific publications is around 80%, which means we lose between 20-30% of our observations.

3.3. Empirical strategy

We proceed with our empirical analysis as follows: We first run a regression on the impact of content similarity between papers and patents and geographic proximity between scientists and inventors on diffusion speed between academic research and patent publication. Yet, as patent value is highly skewed (Harhoff *et al.*, 1999), we also investigate to impactful the resulting patent is on other patents (which we proxy through the citations the patent received in the five years following its publication). The data is described more in detail in Section 4.

3.4. Linear Regression for time difference

We run a linear OLS regression to investigate what factors influence the diffusion speed between scientific research and the subsequent use in a patent, which is described in equation (6):

$$d_{i,p} = \alpha_0 + \beta\phi_i + \zeta_i + \omega_i + \vartheta_{i,k} + \lambda_{i,k} + \varepsilon_{i,k} \qquad (6)$$

where d is the time in years between the publication of the scientific article *i* and the patent *p*. $\phi_{i,k}$ are the characteristics of a given paper (number of authors, average citation impact, technological field). $\zeta_{i,k}$ indicates whether the paper is full Open Access (OA) and $\omega_{i,k}$ whether it can be accessed in an OA repository. $\vartheta_{i,k}$ is the similarity measure between the content of the scientific paper and the patent. $\lambda_{i,k}$ is a dummy that indicates whether at least one scientist and at least one inventor were located in the same country. α is the constant and ε is the error term.

3.5. Negative Binomial Regression for subsequent patent impact

Apart from the diffusion speed, we are also interested to what extent patents that rely on scientific research with particular characteristics (e.g., topic similarity between scientific

publication and patent) are more impactful on other patents. Measuring patent impact is important as patent values are generally high skewed, and hence it allows us to not only investigate the characteristics of scientific papers that influence the diffusion time but also the subsequent impact of the patented invention on other patents (Harhoff *et al.*, 1999).

We proxy this by the citations that a given patent received in the five years after its publication, which is commonly used in the literature to gauge the impact of a patent on other patents (Dechezleprêtre, Ménière and Mohnen, 2017). As many patents do not receive any citations from other patents (and might therefore have limited use for subsequent technological advances), patent citations are used to control for differences in patent quality. As the dependent variable is count data, we rely on a negative binomial regression to estimate the effects:

$$\psi_{i,k} = \alpha_0 + \beta\Phi_{i,k} + \zeta_{i,k} + \omega_{i,k} + \vartheta_{i,k} + \lambda_{i,k}\,\varepsilon_{i,k} \qquad (7)$$

where $\psi_{i,k}$ is a count of patents that rely on the scientific publication weighted by the citation the publication received in the first five years, whereas $\Phi_{i,k}$ represents the characteristics of a given paper (number of authors). $\zeta_{i,k}$ indicates whether the paper is full OA and $\omega_{i,k}$ whether it can be accessed in an OA repository. α is the constant and ε is the error term. As above, $\vartheta_{i,k}$ is the similarity measure between the content of the scientific paper and the patent. $\lambda_{i,k}$ is a dummy that indicates whether at least one scientist and at least one inventor were located in the same country.

After investigating the determinants of the time lag and subsequent patent impact, we analyse to what extent the patents in our dataset rely on scientific knowledge that comes from outside of the home country of the inventor (or the inventors, for that matter).

4. Data

We rely on 600,000 patents from the European Patent Database PATSTAT in the period 2005-2013 and extract the scientific articles that are referenced in patents (2,917). Figure 3 shows the yearly publications of articles in our dataset for the four different clean-energy technologies that our analysis relies on: 1) li-ion batteries, 2) biofuels, 3) solar PV and 4) wind energy. It is noteworthy that wind energy patents only to a very limited extent cite scientific knowledge.

This could be due to the more limited size of the wind energy market and incremental manufacturing improvements not coming from public R&D.

As the publication count of scientific publication in a specific technological field alone may not be representative of the underlying quality of the counted articles[3], we also show the average citations that the articles received to control for quality. The second Y-Axis in Figure 3 also shows the Average Citation Impact (ACI), which uses 102,301 scientific articles based on keywords in Popp (2016) and (Probst, Kontoleon and Anadón, 2021) to estimate how the scientific articles in biofuels, li-ion batteries, solar PV, and wind energy cited by patents compare to all other articles published in the same technological field in the same year. It shows that across years, scientific literature cited by EPO patents is between two to six times more cited than the average article in a technological field. Whereas for batteries and biofuels the trend has remained largely stable over time (and for wind it is not that telling due to the low sample number), for solar PV the ACI has doubled on average from around two to four to six between 2005 and 2013.

[3] A research group *A* might publish many scientific articles of low quality (proxied by subsequent citations by other articles) and another research *B* group very few articles of high quality; in both scenarios a mere count would overestimate the output of the research group A and underestimate that of the research group B.

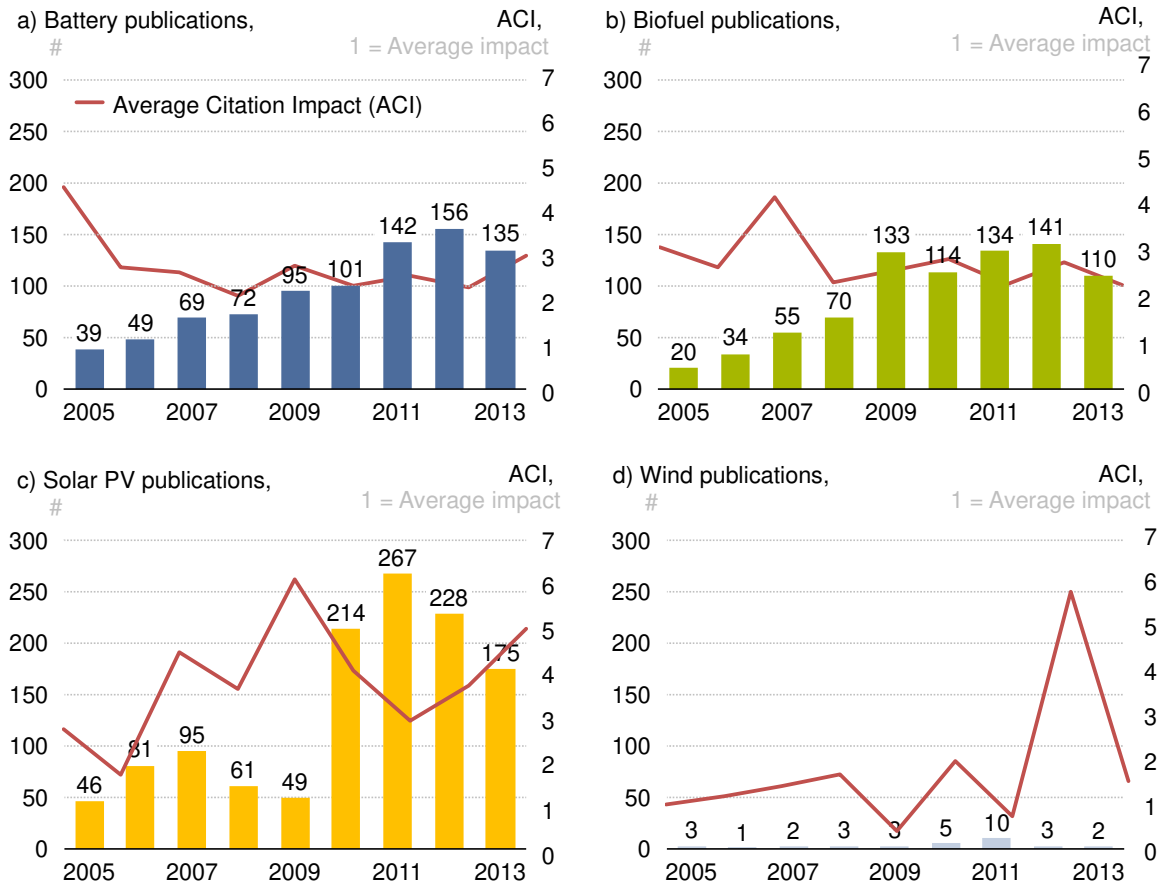

Figure 3: Yearly publications in a) li-ion batteries, b) biofuels, c) solar PV, and d) wind energy that were cited by EPO patents between 2005-2013*. Note: Average Citation Impact is the citations that the article received compared to the average citations article received in the same field in the same year (hence, an article with ACI = 1 received the same citations that a scientific article received in the same field in the same year). Source: author*

Table 1 further shows descriptive statistics of our main dataset of scientific articles. As can be seen, most publications are in solar PV (1,216), with li-ion batteries and biofuels with roughly equal counts (858 and 811, respectively), and very few wind articles (32). The articles used in li-ion battery patents demonstrate the lowest level of OA (9%), whereas 25% of biofuel articles are OA. The number of authors across technological fields is between 3.33 (wind) and 6.05 (for solar PV). The average citation impact (ACI) across articles differs substantially, with solar PV articles receiving more twice as many citations than wind articles.

Descriptive Statistics (Publications Used in Patents)

Full Dataset (2005-2013)

Publications	#n	Number authors	Average Citations	No open access	Full Open Access (Gold)	Open Access Repository (Green)
Technology						
Batteries (Li-Ion)	858	5.13	2.70	91%	5%	4%
Biofuels	811	4.74	2.68	75%	19%	6%
Solar PV	1,216	6.05	3.82	90%	3%	7%
Wind	32	3.33	1.71	82%	4%	14%

Note: #n is the number of open access publications, no open access are those with a paywall (but no open access through repositories), full open access are those that can be found on the website with an open access version, number of authors are those on the paper and average citation impact is the average citation that the paper has received over its lifetime divided by the average citation papers received in a technological field i that were published in the same year t (so an average citation impact of one would be the average citation of papers in a given year and scientific field).

Table 1: Descriptive statistics of 2,917 scientific articles in our main dataset. *Source: author*

Of the 2,917 articles in our dataset, 2,681 (91.9%) could be downloaded through the APIs of various publishers using the CrossRef R-Plug-in. The rest (236 articles) were downloaded manually. We do this for the completeness of our dataset, but this step is not necessary if the articles that could not be retrieved are not systematically different (in terms of location, technology, or some other relevant metric) from the articles that could be downloaded.

Most inventors come from the same country (only 11% of patents have inventors from multiple countries), whereas for scientists this number is around twice as high, with 22% of scientific teams in our dataset working with international colleagues. This might be an indication that the geographical diversity in bringing a technology to market might narrow from highly geographically diverse scientific teams to more national inventor teams that apply these insights, possibly due to secrecy issues and local factors, such as acquiring venture capital.

5. Results

The following section provides our findings on the impact of semantic and geographic distance on the use of scientific knowledge in patented technologies (5.1) and the reliance of investors (who registered patents at the EPO) on knowledge developed outside of their own country (5.2).

5.1. Impact of semantic and geographical distance (and other attributes) on paper-patent diffusion lag

Our results in Table 2 show that an important determinant of the diffusion speed is the scientific impact of the article. The baseline regression also shows that more dissimilar content (between publication and patent) takes longer to diffuse. Specification (5) and (6) then add geographic information to the regression to show that research diffuses faster when the scientist and the inventor are located in the same country – a result that to our knowledge has not been shown before on a large scale. As discussed before, geographic information on patentees and inventors is often lacking from patent documents and is sometimes not correctly classified by the CERMINE algorithm we employ. Therefore, the results from specifications (5) and (6) rely on a dataset of around 1,316 observations (clean-tech-patents citing scientific publications), whereas our entire dataset has 2,917 paper-patent citations. Hence, the last two specifications should be seen as exploratory.

An additional point in the ACI (set to 1 if the paper received the average citations in the field in a given year)[4] reduces the diffusion time by ~0.012 years or 4.38 days. A one-unit increase in the standard deviation in the ACI (which is 8.17), therefore reduces the average diffusion lag between scientific research and application in a patent by 35 days. Hence, scientific articles that are more impactful in the literature tend to diffuse faster from scientific research to application, possibly, due to higher quality or visibility, or both.

It does not appear that OA. impacts the diffusion lag. These results hold also for disaggregating the results for each technological class (biofuels, solar PV, wind, and batteries). Hence, while OA appears to increase the visibility of articles, it does not have an impact on diffusion speed in the case of the four clean-energy technologies investigated in this paper. These results are surprising as one might expect that OA has an impact on both the likelihood of diffusion as well as the diffusion speed (or none of them).

We also include time dummies in our analysis to capture macro trends (such as the rise of information technology) that affect all technologies in our sample (Qiu and Anadon, 2012). These include the large-scale diffusion of information technology, such as the Internet, a host of targeted policies to increase the commercialisation of scientific knowledge (Chan, 2014),

[4] Need to explain this in depth, perhaps with formula

among other factors. Over time these dummies show a linear negative trend, which indicates an increase in diffusion speed over time.

The similarity measure is significant in specifications 3 and 4 and indicates that dissimilar content takes more time to diffuse. Yet, it loses its significance in specification 5 and 6, likely due to the limited sample size. It shows that more similar content diffuses faster between scientific articles and technological applications.

The impact of geography is strongly significant in our smaller sample. It indicates that when scientists and inventors are located in the same country, the diffusion lag is reduced by an average of around 0.37 years (135 days). The diffusion speed in our dataset also tends to be quicker for biofuels and solar PV than for li-ion batteries (which is the base category), possibly because battery patents rely on a more diverse set of scientific research (measured by the content similarity score developed in Section 3.1, which increases the diffusion speed (compared to solar PV and biofuels)). Yet, this last question should be the focus of future research.

Main Regression (All technologies, 2005-2013)

Dependent variable: Time lag between scientific publication and use in patent						
	OLS					
	(1)	(2)	(3)	(4)	(5)	(6)
Average Citation Impact (ACI)	-0.014***	-0.012***	-0.012***	-0.012***	-0.006	-0.005
	(0.005)	(0.004)	(0.004)	(0.004)	(0.006)	(0.006)
Similarity Measure			-0.662*	-0.658*	-0.252	0.094
			(0.358)	(0.359)	(0.537)	(0.557)
Same Country Inventor and Scientist					-0.371***	-0.369***
					(0.114)	(0.115)
Number Authors				0.006		
				(0.010)		
Open Access				0.0259		
				(0.091)		
Open Access Disaggregated						
Base: Closed						
OA-Repository						0.201
						(0.175)
Full OA via publisher						-0.018
						(0.282)
Technologies Disaggregated						
Base: Batteries						
Biofuels						-0.370***
						(0.129)
Solar PV						-0.296***
						(0.115)
Wind						-0.375
						(0.499)
Year Controls	No	Yes	Yes	Yes	Yes	Yes
N	2,917	2,917	2,917	2,917	1,306	1,306
Adj. R2	0.002664	0.334	0.3345	0.3341	0.3082	0.3116

Average Citation Impact (ACI) measures the impact a scientific publication published in year *t* had in a technological field *i* by dividing the total citations received compared to all other scientific articles published in the same year in the same technological field. The ACI is set to 1, which represents the average citation received by scientific articles in a technological field for a given year. Number of authors refers to the number of authors of the scientific article. Similarity Measure indicates how similar the citing patent is to the scientific paper.
* p< 0.10, ** p<0.05,*** p < 0.01.

Table 2: Main regression on the influence of different characteristics on the diffusion lag between scientific publications and patents. *Source: author*

Nonetheless, we are not only interested in what extent the diffusion lag is impacted by different factors, but also how the patent performs that relies on scientific articles with specific attributes, which we analyse in Table 3. The ACI of the scientific paper that the patent relies on is a strong and significant predictor of the citations that the patent receives in the five years following its publication. In the previous regression, the similarity measure indicated that more dissimilar paper-patent connections tend to diffuse more slowly. The regression in Table 3, however, shows that while more dissimilar content takes longer to be used in patents, the eventual impact of the patent is greater.

However, the number of authors on the scientific articles that the patents cite does not have a significant impact on the eventual impact of the patent, which stands in contrast to findings that co-authored scientific publications receive more citations from other publications due to greater visibility (Abramo and D'Angelo, 2015). The dummy for the same country of inventor and scientist is significant but only at the 10% level. This may indicate that while knowledge might flow quicker between more closely co-located scientists and inventors, these are not necessarily more impactful (measured by patent citations). Yet, the results from specifications 5 and 6 should be taken more carefully, as we lose more than half of our sample due to missingness in the data. Across the technological classes, solar PV inventions tend to be cited the most (possibly due to a greater stock of solar PV patents increasing the likelihood of receiving citations by another patent (Dechezleprêtre *et al.*, 2011)) but there is no significant difference in our dataset between biofuels, batteries and wind technologies in terms of citation.

Patent impact (All technologies, 2005-2013)

Dependent variable: Citations received in five years following publication of patent by other patents						
Negative Binomial Regression (Marginal Effects)						
	(1)	(2)	(3)	(4)	(5)	(6)
Average Citation Impact (ACI)	0.131*** (0.030)	0.136*** (0.027)	0.136*** (0.027)	0.137*** (0.028)	0.314*** (0.142)	
Similarity Measure			-3.997** (1.920)	-3.6964** (1.900)	-6.234 (4.949)	
Number authors				0.066 (0.053)	0.264 (0.160)	
Same Country Inventor and Scientist					-1.520* (0.891)	
Technologies Disaggregated						
Base: Batteries						
Biofuels						0.202 (0.190)
Solar PV						0.635*** (0.168)
Wind						0.359 (0.7363)
Year Controls	No	Yes	Yes	Yes	Yes	Yes
N	2,917	2,917	2,917	1,306	1,306	1,306
AIC	11,759	11,683	11,631	11,631	5245.1	5244.6

Average Citation Impact (ACI) measures the impact a scientific publication published in year *t* had in a technological field *i* by dividing the total citations received compared to all other scientific articles published in the same year in the same technological field. The ACI is set to 1, which represents the average citation received by scientific articles in a technological field for a given year. Number of authors refers to the number of authors of the scientific article. Similarity Measure indicates how similar the citing patent is to the scientific paper. Same Country Inventor and Scientist indicate whether the authors of the scientific paper and the inventor of the patent are from the same country.

* p< 0.10, ** p<0.05,*** p < 0.01.

Table 3: Impact of the subsequent patent on other patents (measured as citations received from other patents in the five years after publication). *Source: author*

5.2. Geographic dynamics of diffusion

It is also important to explore how much of the knowledge used in patents comes from inside and outside of the countries the inventors reside in. Figure 4 shows on the left-hand side the location of the scientists (which we took directly from the affiliation data of academic articles)

and on the left side the country of residence of the inventors. Our dataset contains 63 'scientist-countries' linked to 54 'inventor-countries'. These are the 1,306 paper-patent links from specifications 5 and 6 from our previous regressions.

Several things are noticeable: First, on a general level, countries generally only capture a small fraction of the output of their scientists as knowledge diffuses widely across different economies. However, although most countries only capture a fraction of their own research, their inventors rely on scientific advances done elsewhere. Second, as these are patents registered in Europe, it appears that many European countries have a bigger share of the eventual patents than in the research these patents rely on. Yet, as these are patents registered at the European Patent Office (EPO), these are by nature overrepresented (so-called 'home bias effect' (De Rassenfosse *et al.*, 2013)). Third, several big players produce a substantial number of scientific publications that patents rely on, particularly in Asia (e.g., China, Japan, South Korea), the United States, and major European economies. Smaller countries from the Global South are notably absent.

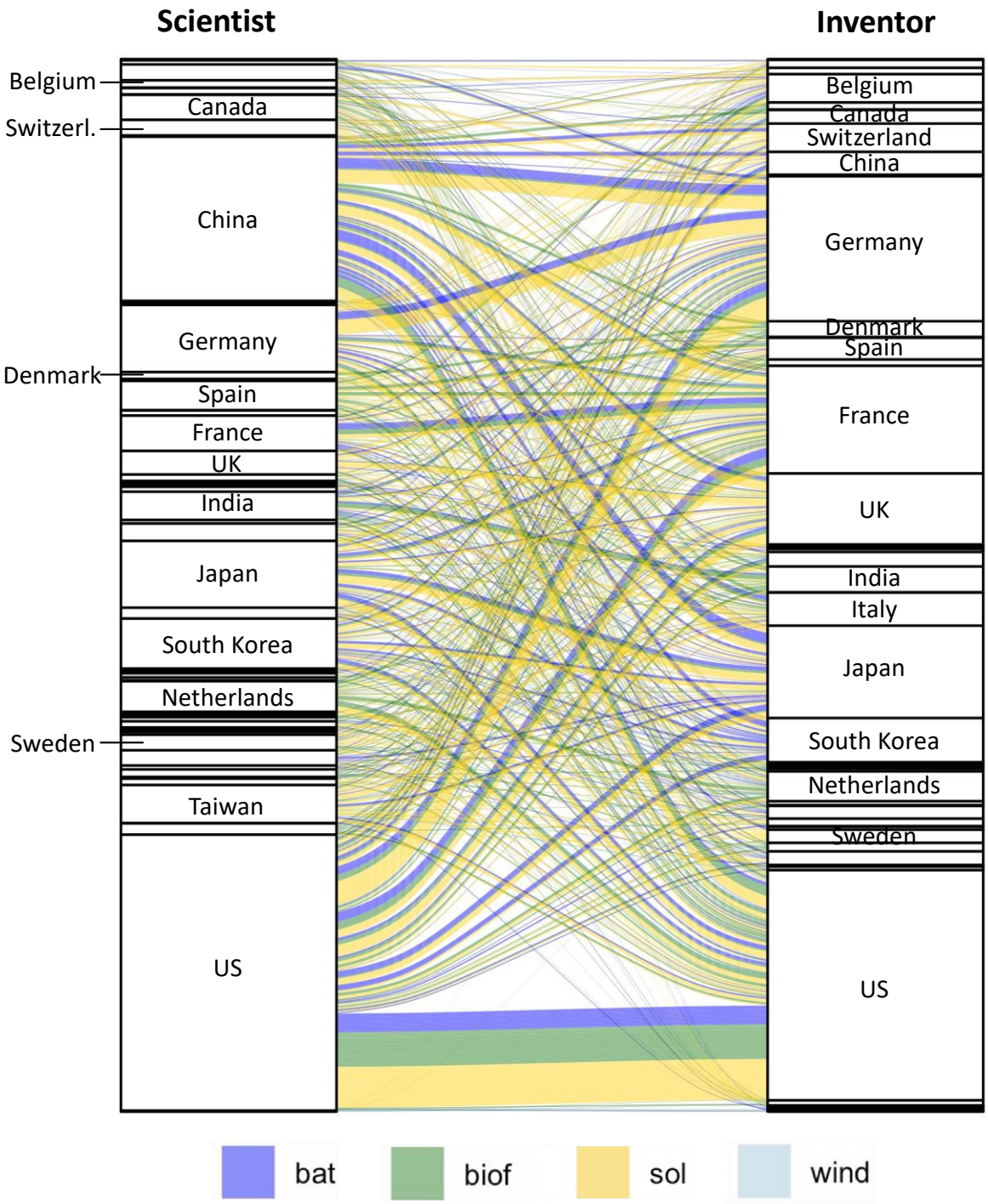

Figure 4: Reliance of EPO inventors on the work of scientists by country. *Note: citations are fractional counts, meaning that if a paper has three authors from three different countries each country receives one-third. Source: author*

However, countries in our dataset differ substantially on the reliance on outside knowledge as can be seen in Figure 5. As our dataset relies on 1,306 patent-paper linkages, these numbers should be interpreted carefully, but underscore the international nature of

research. The countries can be divided into three main classes: medium reliance (50-75%), medium-to-high reliance (76-90%) and high reliance (and more than 90%). Countries with medium reliance includes big countries, such as China and the United States, which themselves have many universities (allowing them to explore more niche topics). Medium to high reliance are mostly European countries (such as Germany, Spain, but also Japan), which rely to more than 75% on scientific sources that originate outside of the country. Countries that are very highly reliant on outside research include countries such as the UK, Italy and South Korea. These are countries that are very open economies. The average for our entire dataset is that for each country around 81% of science used in patents is from outside of the inventors' country.

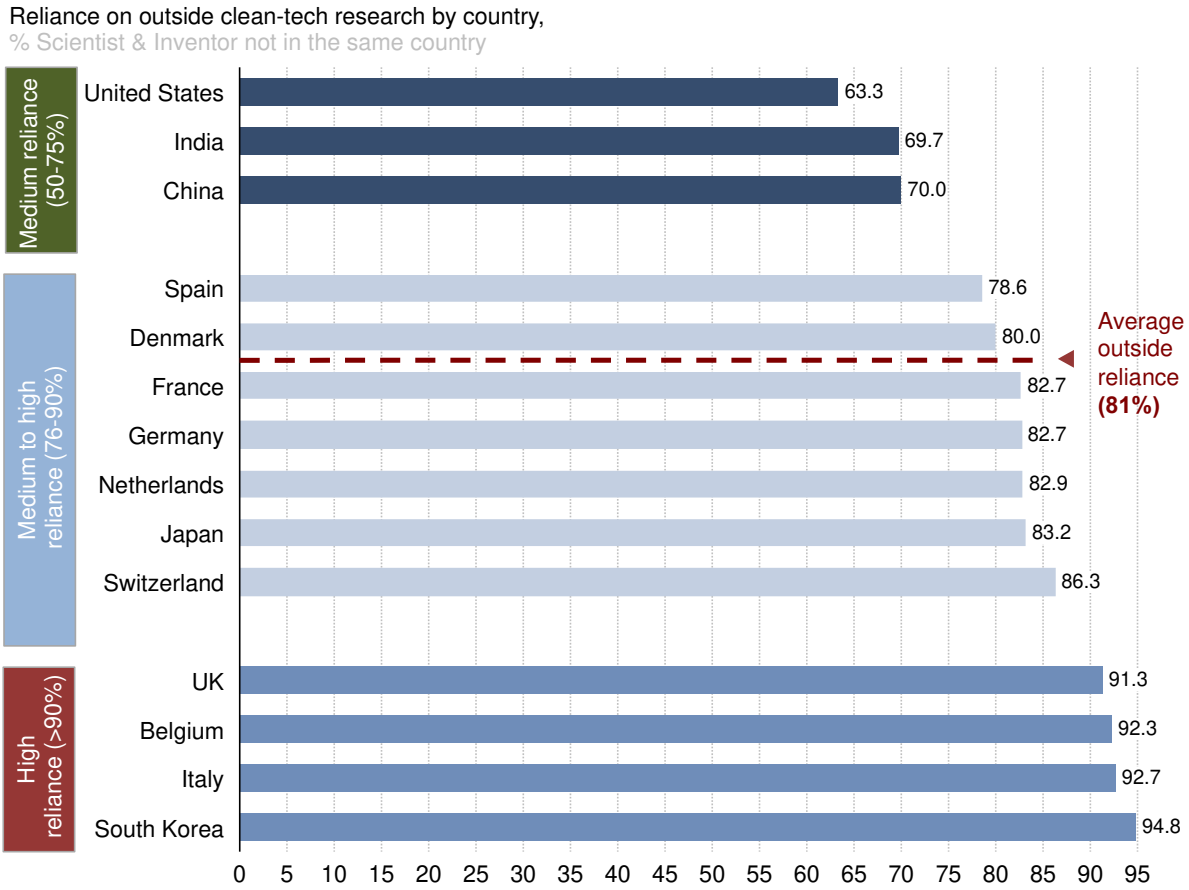

Figure 5: Reliance on outside clean-tech research by country. *Source: author*

Figure 6 shows the top-5 sources of knowledge for patents registered at the EPO in the four clean energy technologies. The rise of China in all four technologies is visible for batteries, biofuels, and solar PV. Our data for wind is more limited as we only have few observations and therefore only provides more limited information. The dominance of the US publications for European patents remains relatively high but is only slightly higher than China in all four clean-energy technologies. This is in line with recent findings of the European Innovation

Scoreboard (2018; p.1), which noted that Europe's lead over China is "decreasing rapidly with China having improved almost three times [its innovation performance[5]] as fast as the EU". Similarly, recent research estimated that scientific articles published in China received 37% of global academic citations (Xie and Freeman, 2019).

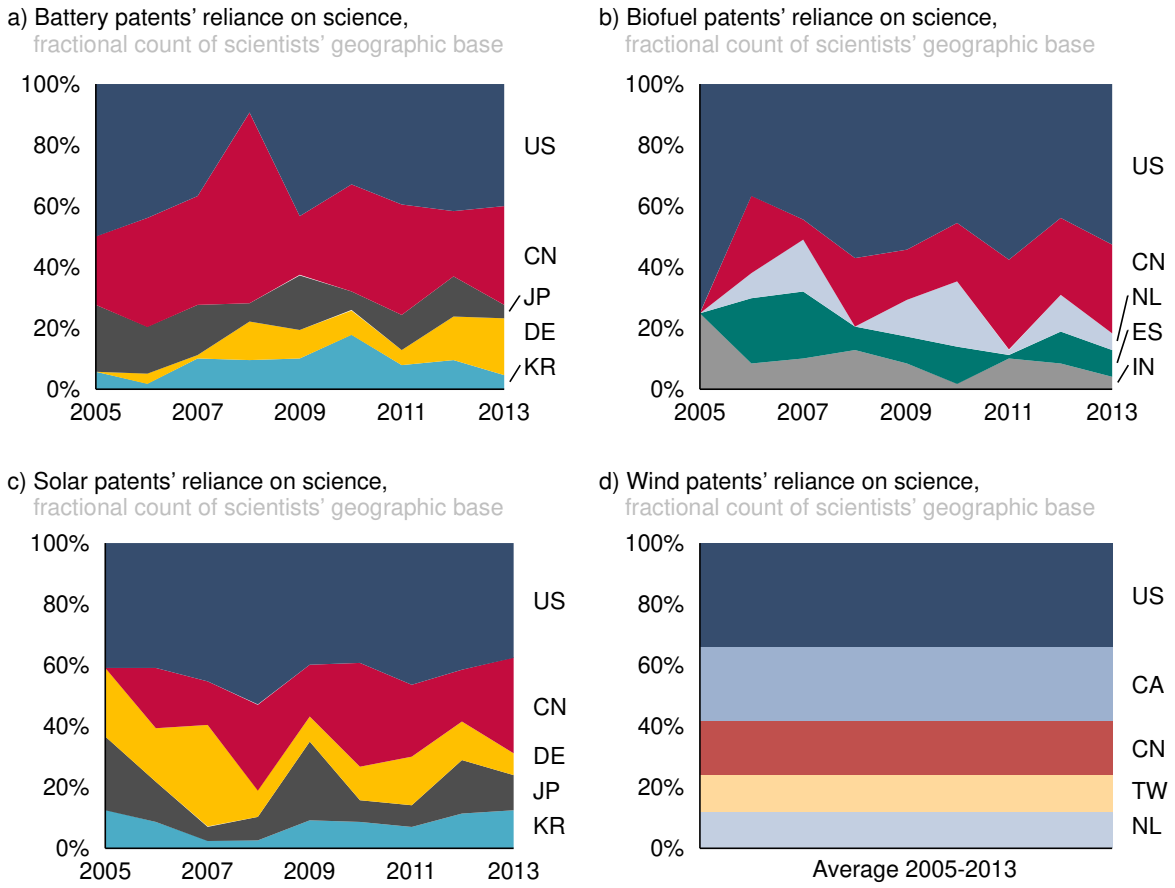

Figure 6: Geographic location of outside knowledge. Clean-tech patents registered at EPO. Source: author; excludes patents where scientists and inventor are from the same country, which are around 20% of all patents. Only average reliance of wind patents on science due to low numbers. *Source: author*

6. Conclusion

As research productivity across major economic sectors is falling (Goldin *et al.*, 2018), understanding the determinants of knowledge diffusion between scientific research and industrial application is critical (Bloom *et al.*, 2017). This is particularly true for fields such as

[5] The European Innovation Scoreboard (2018) relies on various metrics related to the performance of national innovation systems, which includes framework conditions, investments, and innovation activity.

clean energy technologies that are already playing – and are poised to play – are a critical role in bringing the world economies' trajectory in line with the two-degree target set out in the Paris Agreement (Dechezleprêtre, 2016).

We develop a novel data science method that connects scientific articles with patented technologies in a manner not undertaken before in the literature. Our analysis using the linkages between almost 3,000 paper-patent in four clean-energy technologies, shows that the diffusion speed between scientific publications and patents in these technologies has fallen from around six to two years between 2005-2013. As we observe this trend across all four technologies, it appears that larger forces are at play. We hypothesise that several factors influence this trend, such as increased competition and technological maturation (Hoppmann *et al.*, 2013; Huenteler *et al.*, 2016), the ascent of fast and widespread communication technologies (Greenstein, 2010), and declining research productivity (hence, needing more research to maintain the same level of output) (Bloom *et al.*, 2017).

Our detailed analysis of semantic and geographic proximity of the scientific sources used in patents in four main clean energy technologies registered at the EPO shows that articles that are more widely cited in the scientific literature also diffuse quicker from universities to firms, possibly due to higher visibility and quality for inventors. We also show that while scientific research that is more similar to the content of the patent diffuses quicker, more dissimilar content takes longer to be used in patents but is more impactful (possibly due to greater novelty of the patent, which combines more different sources of technology). In addition, our analysis suggests that physical proximity between scientists and inventors plays a role in knowledge diffusion, which has been documented in the literature before.

Our analysis also indicates that scientific teams are substantially more international than inventors. Scientific teams are around twice as likely to rely on international collaboration than inventor teams, which suggests that geographic distance for bringing a technology to market may be more important than for scientific research (possibly due to secrecy issues in patents and a greater reliance on local venture capital). We show that scientific research is used only to a small extent by the 'home country' (where the researchers reside), but diffuses widely across different countries. In our dataset, on average 81% of the scientific articles used are outside of the inventors' country (or countries, for multinational teams). Yet, countries also benefit from their openness towards the research communities in other countries, as they absorb spillovers from research conducted abroad. Yet, our analysis also demonstrates that countries with large domestic markets and many universities (possibly allowing for a greater division of labour) rely to a larger extent on domestic sources, which can be seen for countries such as

China, India, and the United States. In contrast, particularly European economies (such as the UK, Germany, and Spain) and smaller Asian countries (compared to India and China), such as South Korea and Japan, show a higher degree of reliance on outside knowledge. Hence, while scientific research appears to rely on international research networks, the application of such research has a greater domestic component.

Several limitations should be highlighted. There are various ways in which knowledge can 'flow out' of universities. While scientific publications are ranked by scientists as an important way to bring scientific insights towards technological application (Agrawal and Henderson, 2002), there are several other transfer channels not analysed in this paper, which include co-supervision of students between universities and firms, as well as consulting services provided by professors to firms. In addition, as only a few clean-energy patents cite scientific studies, our sample size is limited to around 3,000 patent-paper linkages. In addition, the similarity-content measure introduced in 3.1 relies on using the title of papers and patents (instead of abstracts due to missingness). Future research should use abstracts to introduce even more granularity in the similarity-content measure once better data becomes available.

Further research in this domain should analyse closely how more detailed measures of geographic distance influence diffusion speed (once better and more detailed inventor data becomes available). It would also be interesting to investigate to what extent the gender of scientists (female/male/diverse) leads to differences in the uptake of research. Here advances in machine learning can provide an interesting revenue for research, as most gender coding software rely on historical data, which is mostly limited to European names (which is not appropriate for the international nature of scientific research and patenting). In addition, expanding this analysis to technologies that go beyond the low-carbon technologies analysed in this paper (and patent offices outside of the EPO) is another fruitful future research avenue.

Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation programme project INNOPATHS [Grant agreement No. 730403], the Department of Land Economy and the School of Humanities and Social Sciences, University of Cambridge and the Heinrich Böll Foundation.

Conflict of Interest

We declare no conflict of interest.

References

Abramo, G. and D'Angelo, C. A. (2015) 'The relationship between the number of authors of a publication, its citations and the impact factor of the publishing journal: Evidence from Italy', *Journal of Informetrics*, 9(4), pp. 746–761. doi: 10.1016/j.joi.2015.07.003.

Agrawal, A. K. and Henderson, R. (2002) 'Putting patents in context: Exploring knowledge transfer from MIT', *Advances in Strategic Management*, 26, pp. 13–37. doi: 10.1108/S0742-3322(2009)0000026033.

Azoulay, P. *et al.* (2019) 'Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules', *The Review of Economic Studies*, 86(1), pp. 117–152. doi: 10.1093/restud/rdy034.

Battke, B. *et al.* (2016) 'Internal or external spillovers - Which kind of knowledge is more likely to flow within or across technologies', *Research Policy*. Elsevier B.V., 45(1), pp. 27–41. doi: 10.1016/j.respol.2015.06.014.

Blei, D. M. *et al.* (2003) 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, 3, pp. 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993.

Bloom, N. *et al.* (2017) 'Are Ideas Getting Harder to Find?' doi: 10.3386/w23782.

Branstetter, L. (2006) 'Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States', *Journal of International Economics*, 68(2), pp. 325–344. doi: 10.1016/j.jinteco.2005.06.006.

Chan, G. A. (2014) 'The Commercialization of Publicly Funded Science : How Licensing Federal Laboratory Inventions Affects Knowledge Spillovers', *Working Paper*.

Chung, J. K. *et al.* (1989) 'Measures of distance between probability distributions', *Journal of Mathematical Analysis and Applications*, 138(1), pp. 280–292. doi: 10.1016/0022-247X(89)90335-1.

Dechezleprêtre, A. *et al.* (2011) 'Invention and transfer of climate change-mitigation

technologies: A global analysis', *Review of Environmental Economics and Policy*, 5(1), pp. 109–130. doi: 10.1093/reep/req023.

Dechezleprêtre, A. (2016) *Why aren't we investing enough in low-carbon technologies?*, *WE Forum*. Available at: https://www.weforum.org/agenda/2016/10/how-to-reverse-the-dangerous-decline-in-low-carbon-innovation.

Dechezleprêtre, A., Martin, R. and Mohnen, M. (2013) 'Knowledge spillovers from clean and dirty technologies : a patent citation analysis Centre for Climate Change Economics and Policy', *Grantham Research Institute and the Environment Working Paper No 151*, (151). doi: 10.1016/S0890-8508(03)00045-8.

Dechezleprêtre, A., Ménière, Y. and Mohnen, M. (2017) 'International patent families: from application strategies to statistical indicators', *Scientometrics*, 111(2), pp. 793–828. doi: 10.1007/s11192-017-2311-4.

European Innovation Scoreboard (2018) *European Innovation Scoreboard 2018*. Available at: https://ec.europa.eu/growth/content/european-innovation-scoreboard-2018-europe-must-deepen-its-innovation-edge_en (Accessed: 3 September 2019).

Falck, O., Heblich, S. and Kipar, S. (2010) 'Industrial innovation: Direct evidence from a cluster-oriented policy', *Regional Science and Urban Economics*. Elsevier B.V., 40(6), pp. 574–582. doi: 10.1016/j.regsciurbeco.2010.03.007.

Goldin, I. *et al.* (2018) 'Why is productivity slowing down?', *University of Oxford*, pp. 1–35. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=3867664&site=ehost-live.

Greenstein, S. (2010) *Nurturing the Accumulation of Innovations :*, *Management*. doi: 10.3386/w15905.

Griliches, Z. (1992) 'The Search for R&D Spillovers', *The Scandinavian Journal of Economics*, 94(1992), p. S29. doi: 10.2307/3440244.

Harhoff, D. *et al.* (1999) 'Citation Frequency and the Value of Patented Inventions', *Review of Economics and Statistics*, 81(3), pp. 511–515. doi: 10.1162/003465399558265.

Hellinger, E. (1909) 'Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen.', *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1909(136), pp. 210–271. doi: 10.1515/crll.1909.136.210.

Hogenboom, F., Capelle, M. and Moerland, M. (2014) 'News Recommendation Using Semantics with the Bing-SF-IDF Approach', in, pp. 160–169. doi: 10.1007/978-3-319-14139-8_18.

Hoppmann, J. *et al.* (2013) 'The two faces of market support - How deployment policies affect technological exploration and exploitation in the solar photovoltaic industry', *Research Policy*. Elsevier B.V., 42(4), pp. 989–1003. doi: 10.1016/j.respol.2013.01.002.

Huenteler, J. *et al.* (2016) 'Technology life-cycles in the energy sector - Technological characteristics and the role of deployment for innovation', *Technological Forecasting and Social Change*, 104, pp. 102–121. doi: 10.1016/j.techfore.2015.09.022.

Jaffe, A., Trajtenberg, M. and Handerson, R. (1993) 'Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations', *The Quarterly Journal of Economics*, 108(3), pp. 577–598. doi: 10.2307/2118401.

Keller, W. (2004) 'International Technology Diffusion', *Journal of Economic Literature*, 42(3), pp. 752–782. doi: 10.1257/0022051042177685.

Keller, W. and Yeaple, S. R. (2013) 'The gravity of knowledge', *American Economic Review*, 103(4), pp. 1414–1444. doi: 10.1257/aer.103.4.1414.

Kim, M., Park, Y. and Yoon, J. (2016) 'Generating patent development maps for technology monitoring using semantic patent-topic analysis', *Computers and Industrial Engineering*. Elsevier Ltd, 98, pp. 289–299. doi: 10.1016/j.cie.2016.06.006.

Larivière, V. *et al.* (2015) 'Team size matters: Collaboration and scientific impact since 1900', *Journal of the Association for Information Science and Technology*, 66(7), pp. 1323–1332. doi: 10.1002/asi.23266.

Li, D., Azoulay, P. and Sampat, B. N. (2017) 'The applied value of public investments in biomedical research', *Science*, 356(6333), pp. 78–81. doi: 10.1126/science.aal0010.

Luderer, G. *et al.* (2012) 'The economics of decarbonizing the energy system-results and insights from the RECIPE model intercomparison', *Climatic Change*, 114(1), pp. 9–37. doi: 10.1007/s10584-011-0105-x.

McDonald, F., Tsagdis, D. and Huang, Q. (2006) 'The development of industrial clusters and public policy', *Entrepreneurship and Regional Development*, 18(6), pp. 525–542. doi: 10.1080/08985620600884636.

Noailly, J. and Shestalova, V. (2017) 'Knowledge spillovers from renewable energy technologies: Lessons from patent citations', *Environmental Innovation and Societal Transitions*. Elsevier B.V., 22, pp. 1–14. doi: 10.1016/j.eist.2016.07.004.

Peri, G. and Bottazzi, L. (2003) 'Innovation and spillovers in regions: Evidence from European patent data', *European Economic Review*, 47(4), pp. 687–710.

Popp, D. (2016) 'Economic analysis of scientific publications and implications for energy research and development', *Nature Energy*, 1(4), p. 16020. doi: 10.1038/nenergy.2016.20.

Popp, D. *et al.* (2020) 'Innovation and entrepreneurship in the energy sector', *NBER Working Paper Series*.

Probst, B. *et al.* (2020) 'The short-term costs of local content requirements in the Indian solar auctions', *Nature Energy*, 5(11), pp. 842–850. doi: 10.1038/s41560-020-0677-7.

Probst, B. *et al.* (2021) 'Global Trends in the Innovation and Diffusion of Climate Change Mitigation Technologies', *Research Square Preprint*. doi: 10.21203/rs.3.rs-266803/v1.

Probst, B., Kontoleon, A. and Anadón, L. D. (2021) *Connecting scientific advances and patented technologies: The role of open access scientific publishing in clean-technology innovation*. doi: 10.21203/rs.3.rs-320565/v1.

Qiu, Y. and Anadon, L. D. (2012) 'The price of wind power in China during its expansion : Technology adoption , learning-by-doing , economies of scale , and manufacturing localization', *Energy Economics*. Elsevier B.V., 34(3), pp. 772–785. doi: 10.1016/j.eneco.2011.06.008.

De Rassenfosse, G. *et al.* (2013) 'The worldwide count of priority patents: A new indicator of inventive activity', *Research Policy*. Elsevier B.V., 42(3), pp. 720–737. doi: 10.1016/j.respol.2012.11.002.

Stephan, A. *et al.* (2017) 'The sectoral configuration of technological innovation systems: Patterns of knowledge development and diffusion in the lithium-ion battery technology in Japan', *Research Policy*. Elsevier B.V., 46(4), pp. 709–723. doi: 10.1016/j.respol.2017.01.009.

Sturgeon, T., Van Biesebroeck, J. and Gereffi, G. (2008) 'Value chains, networks and clusters: Reframing the global automotive industry', *Journal of Economic Geography*, 8(3), pp. 297–321. doi: 10.1093/jeg/lbn007.

Toole, A. A. (2012) 'The impact of public basic research on industrial innovation: Evidence from the pharmaceutical industry', *Research Policy*. Elsevier B.V., 41(1), pp. 1–12. doi: 10.1016/j.respol.2011.06.004.

Wilson, C. and Grubler, A. (2011) 'Lessons from the history of technological change for clean energy scenarios and policies', *Natural Resources Forum*, 35(3), pp. 165–184. doi: 10.1111/j.1477-8947.2011.01386.x.

Xie, Q. and Freeman, R. B. (2019) 'Bigger Than You Thought: China's Contribution to Scientific Publications and Its Impact on the Global Economy', *China and World Economy*, 27(1), pp. 1–27. doi: 10.1111/cwe.12265.

Zhou, K. Z. and Li, C. B. (2012) 'How knowledge affects radical innovation: Knowledge base, market knowledge acquisition, and internal knowledge sharing', *Strategic Management*

Journal, 33(9), pp. 1090–1102. doi: 10.1002/smj.1959.

Figures

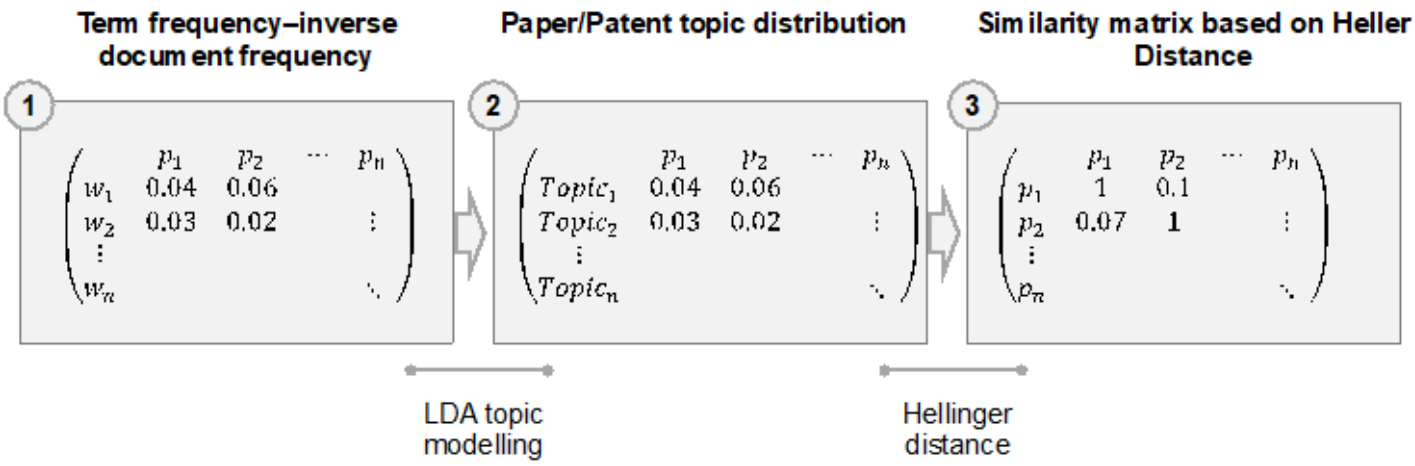

Figure 1

Calculation of similarity matrix between documents based on Heller Distance using a three-step procedure. Note: p indexes a paper/patent, w is a word, and n indexes the number of the paper, patents, or topics. While two identical documents have a Hellinger distance of 0, this is transformed via the similarity measure S(P,Q) = 1 – H(P,Q). Source: author, based on Kim, Park and Yoon (2016)

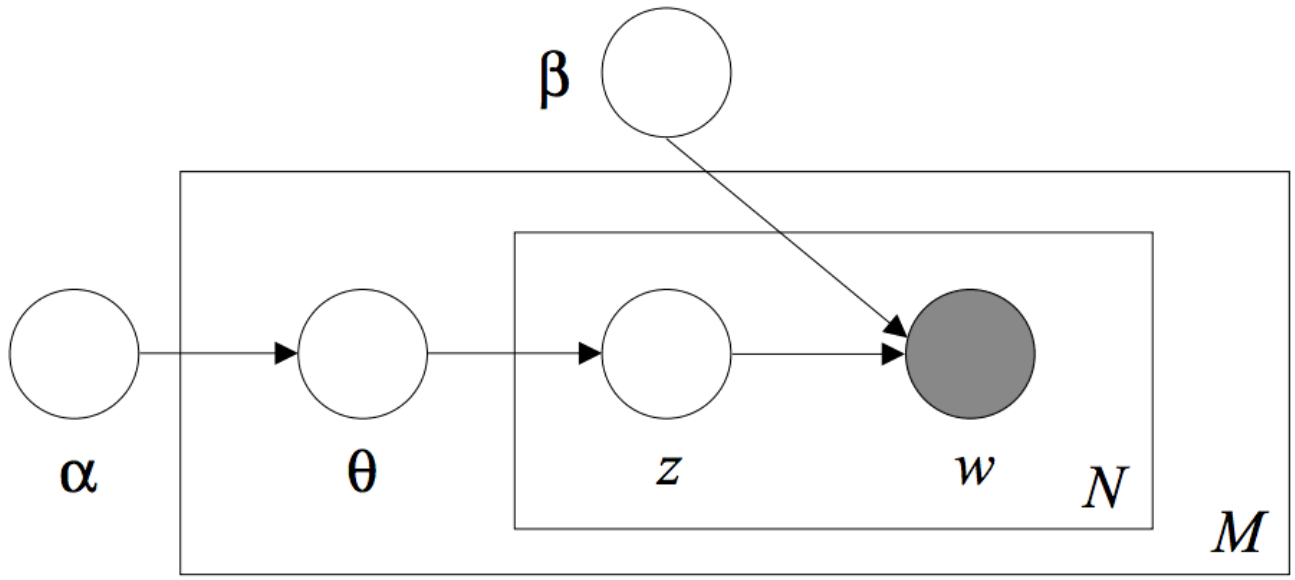

Figure 2

Graphical representation of Latent Dirichlet Allocation (LDA) algorithm. Source: Blei et al. (2003)

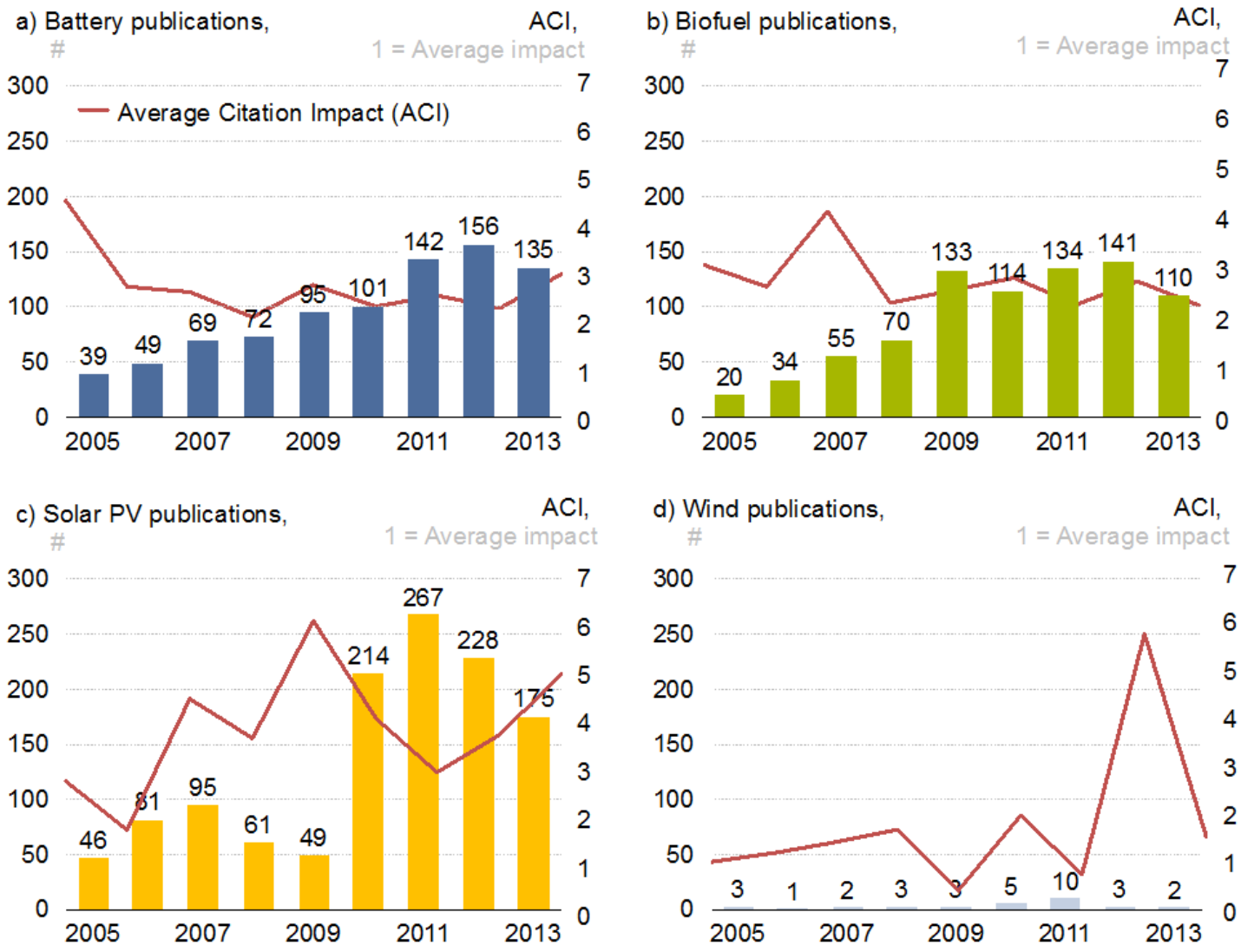

Figure 3

Yearly publications in a) li-ion batteries, b) biofuels, c) solar PV, and d) wind energy that were cited by EPO patents between 2005-2013. Note: Average Citation Impact is the citations that the article received compared to the average citations article received in the same field in the same year (hence, an article with ACI = 1 received the same citations that a scientific article received in the same field in the same year). Source: author

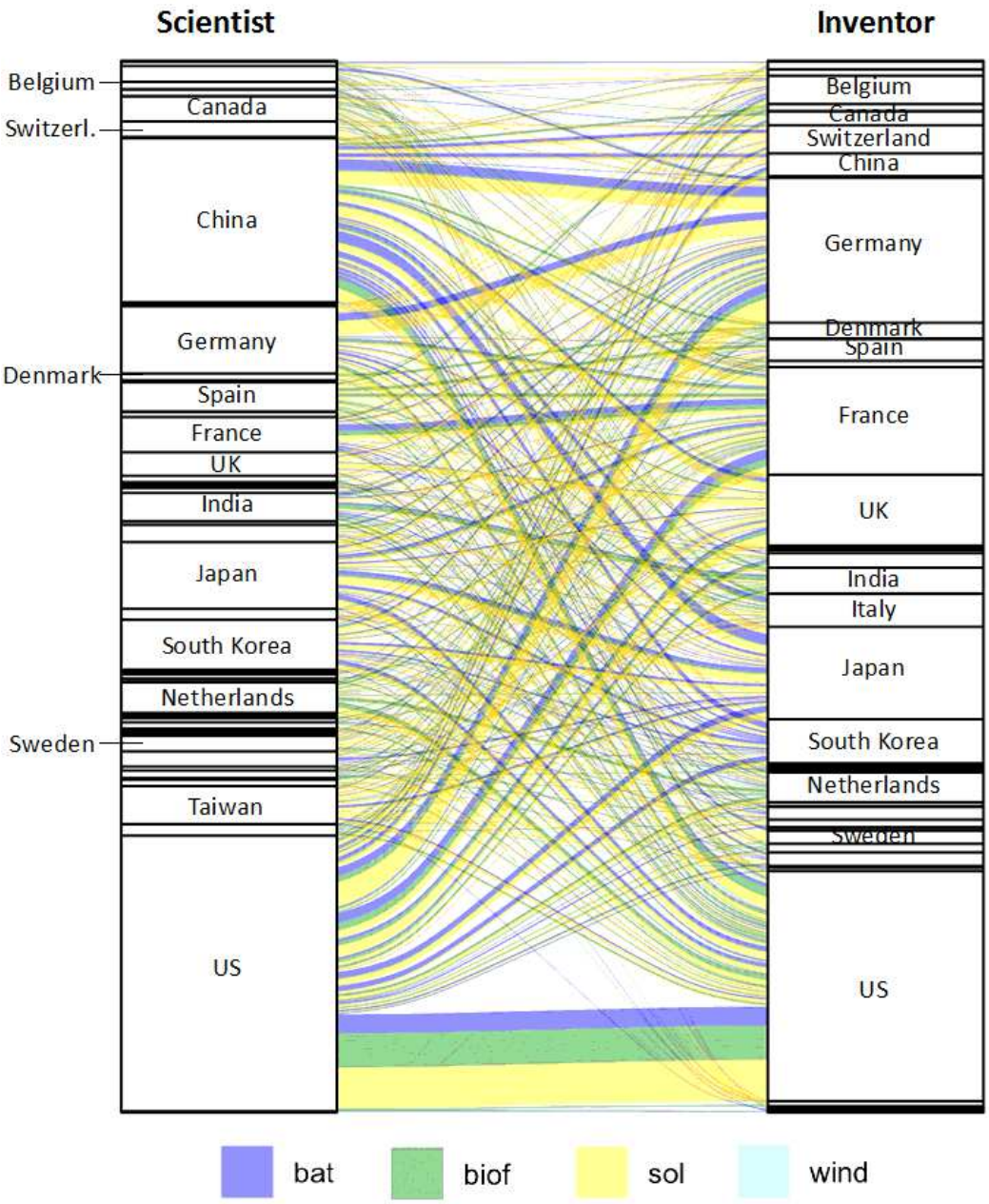

Figure 4

Reliance of EPO inventors on the work of scientists by country. Note: citations are fractional counts, meaning that if a paper has three authors from three different countries each country receives one-third. Source: author

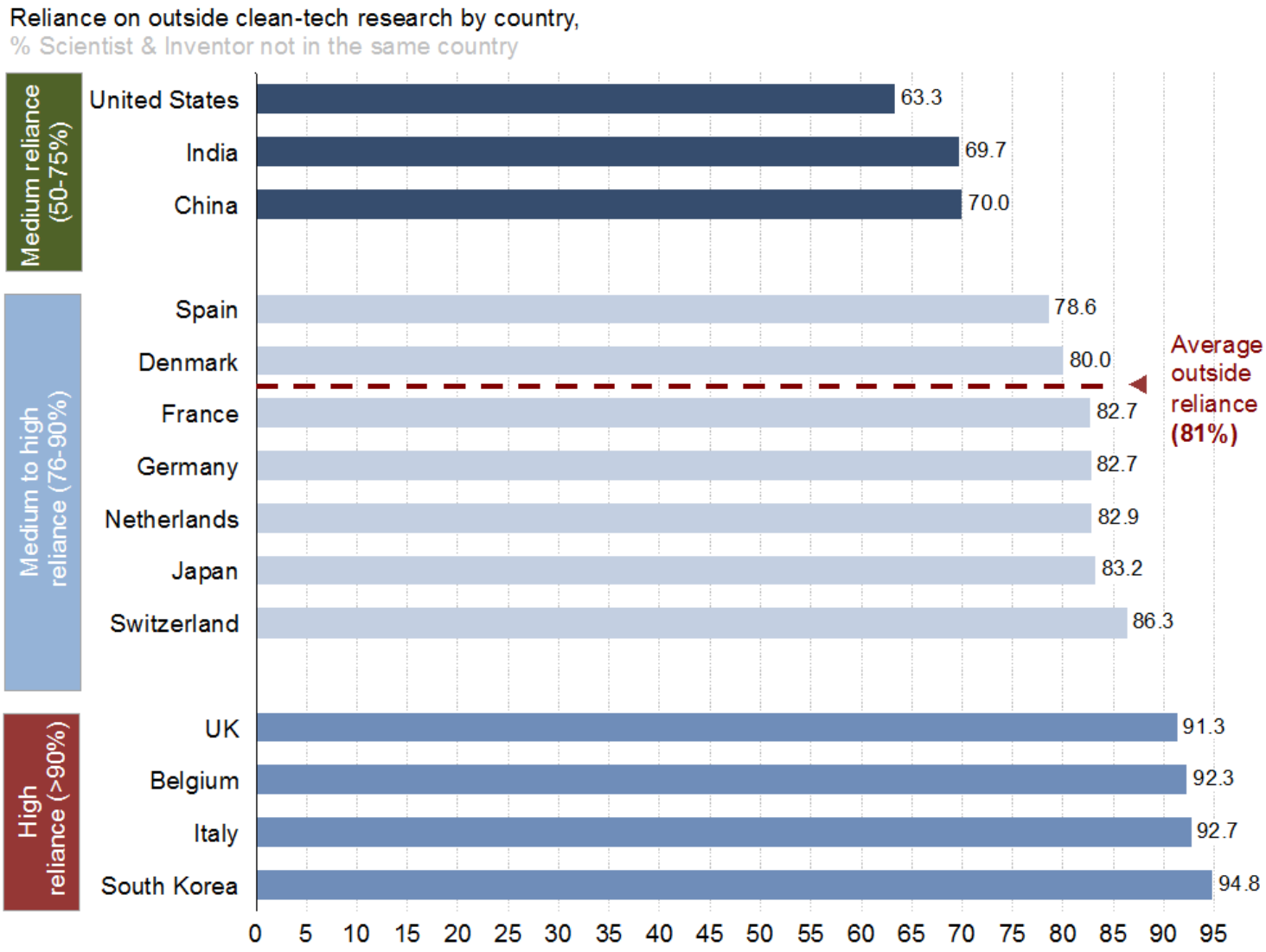

Figure 5

Reliance on outside clean-tech research by country. Source: author

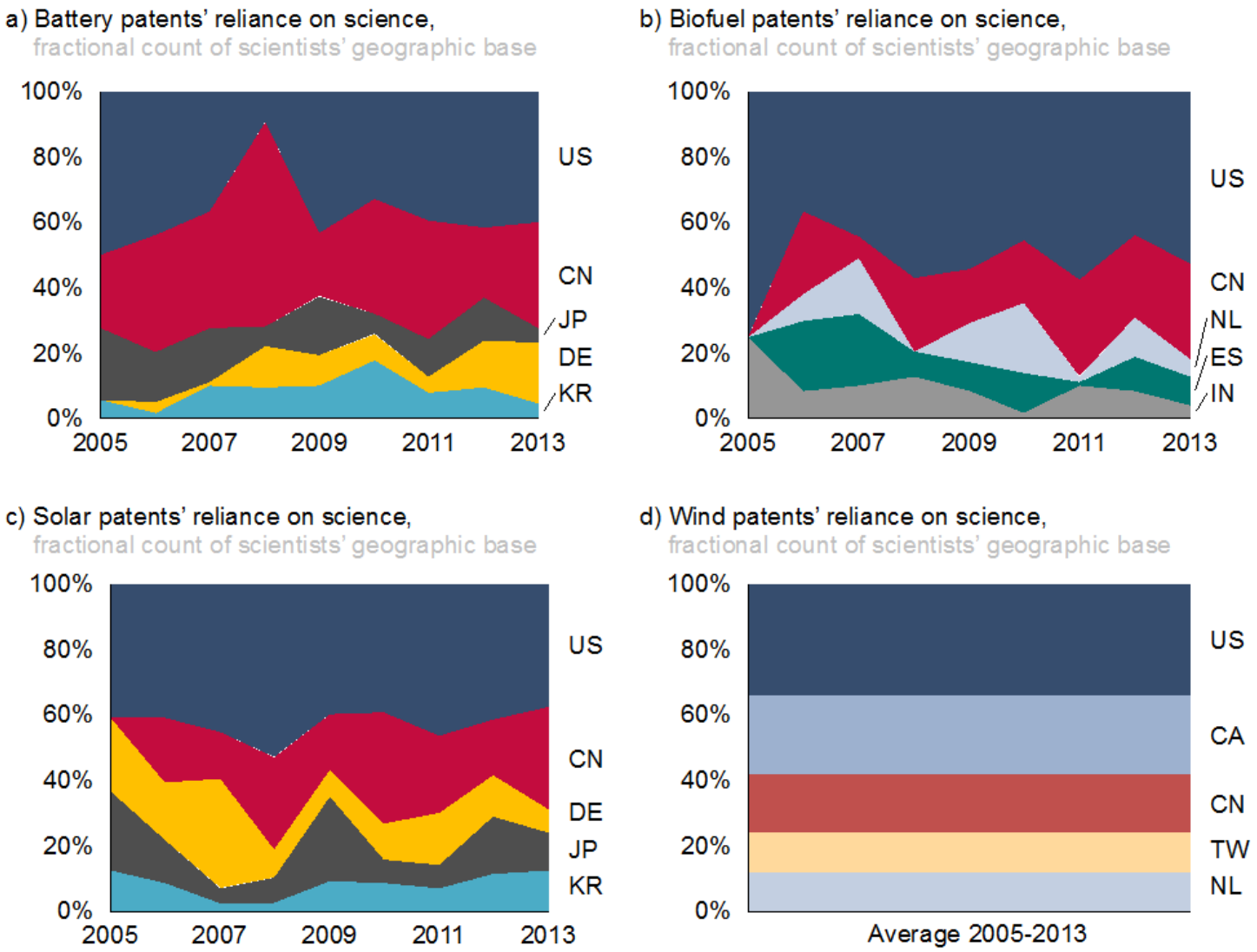

Figure 6

Geographic location of outside knowledge. Clean-tech patents registered at EPO. Source: author; excludes patents where scientists and inventor are from the same country, which are around 20% of all patents. Only average reliance of wind patents on science due to low numbers. Source: author