MIT Sloan School of Management

MIT Sloan School Working Paper 4733-09
4/15/2009

Measuring Innovation Using Bibliometric Techniques: The Case of Solar Photovoltaic Industry

Georgeta Vidican, Wei Lee Woon, Stuart Madnick

Measuring Innovation Using Bibliometric Techniques: The Case of Solar Photovoltaic Industry

Georgeta Vidican
Wei Lee Woon
Stuart Madnick

Working Paper CISL# 2009-05

April 2009

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Measuring Innovation Using Bibliometric Techniques
The Case of Solar Photovoltaic Industry

Georgeta Vidican[1], *Wei Lee Woon*[1], *Stuart Madnick*[2]

[1]Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
[2]Massachusetts Institute of Technology, Cambridge, MA

Paper submitted to the *Advancing the Study of Innovation and Globalization in Organizations* (ASIGO) Conference in Nurnberg, Germany, May 29-30, 2009

March 15, 2009

Abstract

In this paper, we use feature extraction and data analysis techniques for the elucidation of patterns and trends in technological innovation. In studying innovation, we focus on the role of public research institutions (research universities and national laboratories) in the development of new industries. More specifically, we are interested in measuring innovation through research collaborations between these institutions and the private sector.

The proposed methods are primarily drawn from the field of bibliometrics – i.e. the analysis of information and trends in the publication of text documents, rather than the contents of these documents. In particular, we seek to explore the relationship between joint publication patterns and trends, R&D funding, technology development choices, and the viability and effectiveness of industry-university collaborations.

To focus the discussions and to provide concrete examples of their applicability, this study will have an initial emphasis on the solar photovoltaic (PV) sector in the U.S., though the techniques and general approach devised here will be applicable to a broad range of industries, situations, and locations.

Our analysis suggests that interesting information and conclusions can be derived from this line of analysis. The results obtained using our data extraction techniques allow us to identify early technology focus in different areas within solar PV technologies, and to determine potential technology pathways, which is critical for innovation policy in the renewable energy domain.

1. Introduction

1.1 Problem statement

The increasing challenge of international competitiveness driven by knowledge production and innovation, calls for an assessment of the quality and use of indicators for science, technology, and innovation (OECD 2007, Smith 1998). While innovation is difficult to quantify, some aspects related to key dimensions of inputs and outputs can still be measured (Smith 2007). Measuring innovation is even more important for emerging industries, such as the renewable energy sector, receiving large amounts of governmental spending both for research and development (R&D) as well as for market expansion.

This paper explores the role of public research institutions (universities and national laboratories) in the development of new industries, focusing on measuring innovation outcomes, in the form of knowledge creation, using novel bibliometric techniques. A plethora of studies have stressed that close academic-industry collaborations are critical for the formation of industry clusters (Saxenian 1996, Scott 2003). Moreover, the technical revolutions we are experiencing in fields such as renewable energy, involve complex interactions between government, industry, and the academic system. To capture some of these dynamics, our research focuses primarily on measuring joint publications between researchers in the academia, national labs, and the private sector. In order to narrow the discussions and analyses, our research has an emphasis on the solar photovoltaic (PV) sector in the U.S. Nevertheless, the specific methods are applicable to a broad range of industries and contexts.

1.2 Novelty and motivation

Using bibliometrics to study the progression of research and technological development is not a new idea and there is already a significant body of research addressing this problem (for a good review, the reader is referred to Porter (2005, 2007), Losiewicz et al. (2000), Martino (1993)). Interesting examples include visualizing the inter-relationships between research topics (Porter 2005, Small 2006), identification of important researchers or research groups (Kostoff 2001, Losiewicz et al. 2000), the study of research performance by country (de Miranda et al. 2006), (Kim and Mee-Jean 2007), the study of collaboration patterns (Anuradha et al. 2007,Chiu and Ho 2007, Braun et al. 2000) and the prediction of future trends and developments (Smalheiser 2001, Daim et al. 2005, Daim et al. 2006, Small 2006).

As such, it would appear that the applicability of bibliometric techniques to the study of technology development is quite well established. Godin and Gingras (2000) have used bibliometrics to assess the role of universities in the system of knowledge production. Their study concludes that while federal R&D funding declined overtime, universities became more important through increased collaborations with the private sector. Other studies, such as Zimmerman et al. (2009), use bibliometrics to examine national research collaborations. However, there is hardly any study on the progression of science in the renewable energy field. Tsay (2008) traces the evolution of hydrogen energy literature worldwide. Aside from a few exceptions covering national and international collaboration patterns in the fuel cell technology in Norway (Godo et al. 2003), and co-authorship networks in the area of nanostructured solar cells using bibliometric and social network analysis (Larsen 2008), there is much more insight to be gained from examining different types of collaborations in the field of renewable energy technologies, in particular solar energy technologies.

However, despite the high level of activity in the general area of research, there does not appear

to have been a corresponding level of interest in their use for analyzing industry-university collaborative research. Godin and Gingras (2000) have used bibliometrics to assess the role of universities in the system for knowledge production. Their study concludes that while federal R&D funding declined overtime, universities became more important through increased collaborations with the private sector. Other studies, such as Zimmerman et al. (2009) use bibliometrics to examine national research collaborations. However, there is hardly any study on the progression of science in the renewable energy field. Tsay (2008) traces the evolution of hydrogen energy literature worldwide. Aside from a few exceptions covering national and international collaboration patterns in the fuel cell technology in Norway (Godo et al. 2003), and co-authorship networks in the area of nanostructured solar cells using bibliometric and social network analysis (Larsen 2008), there is much more insight to be gained from examining different types of collaborations in the field of renewable energy technologies, in particular solar energy technologies.

As such, there certainly appears to be a high level of activity in this general area of research. However, interestingly there does not appear to have been a similar level of interest in using these tools for analyzing collaborative research amongst players in industry, academia and in the national laboratories. Another important issue is to study how the characteristics and trends in these collaborations reflect the underlying factors of government funding, societal and environmental developments. The research described in this paper was motivated by, and seeks to address these important issues.

The rest of the paper is structured as follows. The current section presented the background and motivations for the research, while Section 2 describes in detail the data sources and the research methodology used to measure knowledge production from academia and industry on areas related to solar PV technologies in the U.S. In Section 3 we describe the main results along with preliminary observations. Section 4 then discusses these results in the larger context of innovation measurement and challenges in the solar photovoltaic sector. Section 5 concludes the paper with main findings and suggestions for future areas of research.

2. Data and methods

2.1 Data

We focus our research primarily on two states in the U.S., California and Massachusetts, for several reasons. First, the origins of the solar photovoltaic industry worldwide emerged from these two regions (the research labs of two large oil companies, Mobil Tyco in Massachusetts and ARCO Solar in California) (Margolis 2002). Second, over the years the solar photovoltaic industry in the U.S. has been concentrated in these two locales, with California hosting the largest share of companies. Third, the regional economies in Boston and San Francisco are the innovation engines for the U.S. economy, clustered around the research universities establishments (Saxenian 1996). Hence, we argue that publications emerging from universities and research laboratories in these two regions define the scope of research for other research establishments in the U.S.[1]

[1] The National Renewable Energy Laboratory (NREL), playing a key role in coordinating and funding research for the solar industry, is however, located outside these two states, in Golden, Colorado.

2.2 Methods

The basic premise for this study is to investigate the use of bibliometric techniques for studying technological innovation relevant to photovoltaics. These are techniques, which focus on patterns and trends of textual information, rather than on the actual content of the text to be analyzed. In particular, we would like to test the usefulness of *hit count*s, i.e.: the number of academic publications relevant to a particular field, as a measure of the level of research activity or interest in that field. These hit counts were collected in yearly bins, allowing the time evolution of research activity over the corresponding periods of time to be visualized and studied.

To conduct the pilot study, the keywords "photovolt*", "solar cell", "solar PV", "solar energy", "solar generation" and "solar power"" were submitted to ISI's Web of Science database. In addition, it was also necessary to include additional search terms so as to specify the types of institutions in which the research was being carried out. Two general approaches were used. In the first approach, generic searches were generated using terms which indicated authors from industry, national laboratories and from academia. The search terms used were:

> University: Address field to include: "univ" or "inst"
> National Laboratory: Address field to include "lab" or "laboratory" or "labs*"
> Industry: address field to include: "inc" or "corp" or "co"

To study the research activity in different states, an additional term was included as follows:

> AD=("inc" SAME "MA"),

which would admit publications where the address fields include "inc" and "MA" in the same line (i.e. this would identify publications originating from industrial researchers located in Massachusetts).

Using this approach gave us a lot more flexibility as we could now generate searches which target specific subsets of the academic literature, which in turn would reflect the level of research being conducted in the corresponding sectors.

In the second approach, two lists were manually compiled for Massachusetts and California: lists of companies known to be involved in solar photovoltaic research, as well as lists of universities and national research laboratories in the two states. To extract the hit counts from the Web of Science web interface, the results were broken up into batches of 10 companies at a time (this was necessary as the lists of companies were too long to be entered into a single search).

Unfortunately, our initial experimentation with the second method revealed that it was too restrictive and retrieved too few papers to be of use in the present study. In addition, this approach was very labor-intensive as a separate list of companies would have to be compiled for any future study. In addition, maintaining and keeping the lists up-to-date would also not be easy. As such, all the results presented here were extracted using the first approach.

The required computational tools were implemented in the Python programming language, as it facilitated faster development and includes a broad selection of libraries, including those useful for the analysis of text and for data collection from the WWW. Python is also a cross-platform environment and allows applications to be deployed on a variety of operating systems and environments.

3. Results

Annual publication counts from the Web of Science database were collected for all the years between 1975 and 2008, inclusive. In addition, a five tap gaussian filter was used to smooth the resulting time series as they were quite noisy and in some cases the number of publications retrieved were very low. This was a reasonable pre-processing step because the research which results in a publication would have been carried out over a period of time prior to the appearance of the publication; as these publication counts are in fact a proxy for the underlying research activities, smoothing the raw data in this way may be viewed as a means of taking this spread into account. Six different sets of graph are presented here, which reflect research activities carried out in universities, national laboratories, industry, and collaborative efforts involving pairings between each of these three sectors. These graphs are presented in Figures 1 to 6. Initial observations are:

1. While the details of individual graphs varied somewhat, the same high-level trend was observed in the majority of the cases: the number of publications started off high with peaks in the early to mid-80s'. However, as we move into the 90s', there was a marked decline in the number of papers which continued until around 1995, after which publication counts were observed to increase again.
2. The number of papers published in Massachusetts were found to be significantly lower than in California. This was particularly true of collaborative research, where in many cases only one or two papers were identified over a period of several years. As this is clearly insufficient, for collaborative research we will focus on results for California and for the entire U.S. only. Besides Massachusetts, this was also a problem in the case of university-industry collaborations in California, where the numbers of papers produced were close to zero for much of the study period.
3. In general, national laboratories in California and Massachusetts appear to have produced more papers in the initial high activity period (ranging approximately from 1975 to 1985), than in the second half of the study period (see Figure 2). For universities, the results are the opposite where the number of publications produced in the second period of high activity (approximately 1995-2008) exceed the publications produced in the first (see Figure 1). However, note that in both cases we still observe the same broad trend where there is a significant drop in the hit counts for the period of time ranging from around 1985 to 1995.
4. The hit counts corresponding to industrial research are a lot higher in the first period, and for the case of California and Massachusetts, there is no apparent recovery post-1995 (see Figure 3).
5. The results for collaborative research were more difficult to analyze as the number of publications found was significantly lower across all the sectors (this is to be expected as the search terms used were more restrictive in these cases). As such, the results were invariably noisier and were frequently unreliable (cases in point being all of the results for Massachusetts, and the results for university-industry collaborations in California). Having said that, the results for collaborative research activities could often be seen to be combinations of the sectors involved. So, for example, the hit counts for laboratory research in California were a lot higher earlier in the study period while for university research, the opposite is observed; accordingly, for laboratory-university research in California the two periods are quite well balanced (see Figure 4).

6. However, an interesting counter-example to the previous observation is found in the case of industry-laboratory research nationwide. As mentioned previously, hit counts for research conducted in industry and in national laboratories start out relatively high, but eventually end up lower, at least on average (see Figure 6). Surprisingly, the publication trend for industry-laboratory research is exactly the opposite - the number of publications produced post 1995 is actually significantly greater than the number of publications in the preceding period. This implies that, prior to 1995, a lot of the research being conducted in the two sectors were carried out in isolation, whereas a much greater proportion of research was being conducted in collaboration in the post-1995 period. A similar observation can be made about university-laboratory collaborations nationwide: prior to 1995, we see that less than a third of research in national laboratories was in collaboration with universities; however, in the post 1995 period, this figure has increased to around two thirds, as reflected in the number of academic publications (see Figure 5).

4. Discussion

In the previous section, data from the Web of Science database was presented, and pertinent numerical trends were discussed. Importantly, instead of a smooth growth curve, as might have been expected, our analysis revealed that innovation in photovoltaics exhibited a rather discontinuous growth pattern from 1975 to 2008. The decade between late 1975-1985, and after 2000 has registered the largest number of publications in the field of solar photovoltaics (PV) in the U.S. (see Figure 7), while the intervening time saw a marked decline in the number of relevant publications.

To better understand our observations, in this section we examine these trends in reference to the federal research and development (R&D) spending on solar PV, collaboration patterns, and institutions involved in solar PV energy research over the years.

4.1. R&D Spending on Solar PV research

The first energy crisis in the late 1970s called for increased attention to renewable energy technologies as alternatives to conventional fuel sources. Drawing on the experience with solar cells for space applications, there were reasons to believe that with sufficient research funding, solar PV can be used to generate cost-competitive energy for residential and commercial use (Margolis 2002). As a result, increased federal R&D funds have been channeled into research for solar technologies (see Figure 8).

The 1980s, however, saw significant reductions in the amount of federal R&D spending for renewable energy technologies, a trend which has been common in most highly industrialized countries (Margolis 2002). Consequently, overtime, the share of funding for solar energy technologies decreased consistently (see Figure 8).

Our results illustrate that there is a correlation between federal R&D funding and the number of scientific publications, until about 2000. More recently, concerns with climate change and energy security, and government support for commercialization of renewable energy technologies, revived the interest in solar energy technologies at public research institutions. Currently, however, a disproportionate amount of funding originates from the private sector, supporting research at universities.

4.2 Institutions involved in solar PV research

Public research institutions (universities and national research laboratories) have been critical for the advance of knowledge in solar PV energy, in terms of both technology development, as well as system integration. Nevertheless, our research suggests that private companies have also been highly involved in solar energy research primarily before 1990s. Below we discuss in greater detail the institutions involved in the solar energy technologies research and their technological focus in different geographical locations.

4.2.1 National Laboratories

Despite the decline in federal R&D funding, national laboratories were instrumental is supporting research interest in solar energy technologies. As Figure 2 shows, while the number of publications from national labs declined, the trend has not been as dramatic as for the level of federal investment. In Massachusetts and California national labs in the respective regions contributed less after 1990s. We argue that the lower regional presence of national labs in the solar PV research could be due to the shift of research competence on solar PV to National Renewable Energy Laboratory (NREL) located in Colorado, created in 1991.

In California, Jet Propulsion Lab (JPL) at California Institute of Technology (Caltech) recorded the largest share of scientific publications until early 1990s (approximately 50% of all publications on solar energy emerging from national labs). The research originating from JPL has been quite diverse in focus, ranging from space solar cells until early 1980s, to improving efficiency and testing reliability of terrestrial solar technology applications in the 1980s, and more recently on third generation solar technologies.

In addition, about 25% of the publications originated at the Lawrence Berkeley National Laboratory (LBNL) associated with the University of California Berkeley (UC Berkeley). While LBNL's involvement in solar research has been very limited early on, more recently, after year 2000, it has concentrated most of national labs research on solar. The spectrum of solar research at LBNL covers different solar technologies.

Other national labs that contributed to the advancement of knowledge in solar energy technologies over the years have been Lawrence Livermore National Laboratory (from 1973-1999), Sandia National Laboratory (1974-2004). Aside from national labs, other research institutions associated with the industry, played an important role in the research landscape of solar technologies in the 70s and 80s, such as Lockheed Missiles and Space Laboratory, Optical Coating Laboratory, US Air Force Laboratory. Interestingly, until 1980s, the research laboratory of a large Silicon Valley semiconductor multinational company, Varian Associates Inc., published 10% of the scientific papers on solar energy technologies in the region.

In Massachusetts, the lower number of national labs is reflected in fewer number of publications emerging from these research institutions. MIT Lincoln Laboratory has been involved in solar energy technologies research only until late 1980s. More than 60% of the research publications from research laboratories originate at MIT Lincoln Lab. Research at Philips Labs, associated with the US Air Force, has been focused entirely on space applications. The MIT Energy Lab was also important in the early stages of research.

4.2.2 Universities

While the share of research on solar PV technologies originating from universities has been lower that from national laboratories, this trend is likely to change with the new focus on advancing knowledge on renewable energy technology and systems at the global level.

In California, Stanford University and UC Berkeley have been the centers of research along with the national labs in the region, until the mid 1980s. UC Berkeley continues to play a key role in the advancement of science, as reflected by a high number of publications throughout the entire period. Nevertheless, in the past five years, research on organic PV technologies has brought Stanford University back to the research landscape.

In Massachusetts, while MIT has usually been the engine and source of innovation for industries such as semiconductors and biotechnology, our results show that this has not been the case for the solar PV industry. While early on MIT's involvement in the solar energy sector has been through the Lincoln Lab, since 2004 MIT has adopted as its mandate to focus on alternative energy technologies. The creation of MIT Energy Initiative and its engagement with multiple energy related private and public partners suggests that MIT's role in energy (and in particular solar) research is likely to expand significantly in the near term.

University of Massachusetts (UMass) Lowell has played a critical role throughout the entire period in advancing knowledge primarily in solar PV energy systems, but also more recently in cutting edge research focused on organic PV (about 40% of university based publications emerged from UMass Lowell). Other universities in the region have also been active in this research area, such as Harvard University, Boston University and Boston College, Northeastern University, UMass Amherst, UMass Boston, Northeastern University, Tufts University, and Clark University.

4.3 Collaboration patterns

Research collaborations have been identified as important for knowledge creation and knowledge transfer. Sharing knowledge and ideas is even more important for an emerging domain of knowledge like solar energy technologies, which builds on interdisciplinary expertise. In the U.S., the Department of Energy (DOE) has initiated several funding schemes to foster research collaborations between public research institutions and the private sector, such as the PV Manufacturing Technology Project (in 1991), the Thin-Film PV Partnership (in 1994), or the Industry Alliance Project (in 2007). Hence, gaining insights into the outcomes of these investment programs over the years, and in the patterns of collaboration, is a valuable exercise.

4.3.1 Collaboration patterns between universities and national laboratories

While a variety of universities are involved in research collaborations with national labs, the share of publications originating from California, and more recently from Colorado (due to NREL's location) is disproportionably higher. Nevertheless, programs such as those initiated by DOE appear to have been successful in stimulated collaborative research efforts since university-national labs collaborations increased significantly after 1992 (see Figure 4).

In California, the proximity of national research labs to the local universities oftentimes leads to blurry boundaries between the two institutions. Hence, intense collaborations are recorded between, for instance, UC Berkeley and LBNL in Northern California, or between Caltech and JPL in Southern California. This type of collaboration is not present in

Massachusetts in the field of solar technologies, although it is strong in other areas, such as biotechnologies.

4.3.2 Collaboration patterns with the industry

In general, we find only a few collaborations between universities or national laboratories with companies (see Figures 5 and 6). At national level, however, such collaborations increased after 1990s. An explanation for this increasing trend is that companies might have realized that solar PV has potential for becoming a market niche following increased government investment in supporting market deployment worldwide (primarily in Germany and Japan).

Collaborations between national labs and companies predominated in the 70s and 80s when large semiconductor and aerospace companies were interested in exploring potential new niche markets. In California, examples of such companies are Spectrolab, Varian Associates Inc, Standard Oil Co., Hughes Aircraft Co., Rockwell Int. Corp., Lockheed Aircraft Corp., and Applied Materials. The focus of research for these collaborations has been on more mainstream solar technologies such the 1st and 2nd generation of solar technologies. In California and Massachusetts very few industry-national labs collaborations were recorded after late 1990s. The increasing trend in such collaborations at national level (see Figure 6) is primarily due to NREL's significant role mainly after 1990 in solar related research.

Collaborations between the industry and academia have been even less frequent. In California IBM has been the main industry collaborator for universities, followed by Solarmer Energy Inc. a start-up in El Monte. Most of the research has been focused on cutting edge solar technologies such as polymer based PV (3rd generation, organic PV solar technologies). University of California (UC) Los Angeles, UC Santa Cruz, and UC San Diego are the universities most engaged in such collaborations. In Massachusetts the few industry collaborations were with universities or research laboratories from outside the state, reflecting a limited solar research agenda at the established institutions.

Given the emerging nature of the solar industry, we did not identify regional clusters of collaboration between companies and universities or national research laboratories. Rather, we find that companies seek research partners with the desired expertise who are not necessarily in their geographical proximity. In California, where the concentration of national labs and universities is higher, collaborations within the geographical cluster are more prevalent.

5. Conclusions

Below we summarize the main findings, discuss limitations of the current analysis, and offer suggestions for future research in the area of bibliometrics and innovation assessment.

5.1 Methods

The results that we have presented in this paper demonstrate that the proposed methodology, which is based on a bibliometric approach, is capable of extracting valuable information from semi-structured sources of data. While this study is still preliminary, it shows that this information is already useful in helping to improve our understanding of trends and patterns in innovation. In the present study, the emphasis has been on innovation in the field of photovoltaics, and more specifically, in the states of Massachusetts and

California in the U.S. However, the described framework is hugely flexible and can be easily generalized to the study of innovation in different fields, or in different geographical locations.

As with any computational framework which exploits semi-structured data, there were certainly some problems. Firstly, we note that success in tracking the progress of innovation in this way is contingent upon our ability to correctly identify publications which are relevant to our study. So, for example, to study photovoltaic research conducted by industry-linked players in California, an appropriate Web of Science search would have to be generated which matches publications resulting from this specific subset of research. For the current study, the following search term was used:

> ***TI=("photovolt*"OR "solar cell" OR "solar PV" OR "solar energy" OR "solar generation" OR "solar power") AND AD=((inc SAME CA) OR (co SAME CA) OR (corp SAME CA))***

In most cases, this successfully extracts the correct results, however we might anticipate a few potential problems:

1. **False positives/negatives:** There might be publications which contain the terms "solar power" or "solar energy" which are not actually relevant to photovoltaic research. Conversely, there might be publications which are relevant to photovoltaic research, but which do not include any of these terms in the titles. Similarly, there might companies or other industrial research entities which do not explicitly state the terms "co", "corp" or "inc" in their address fields.
2. **Inconsistent database coverage:** This is related to the previous problem; in many cases, a much better retrieval rate might have been achieved had we used title/abstract searches instead of simply using title searches - unfortunately, the Web of Science database was only able to conduct abstract searches for publications dating from 1991 and so for uniformity we had to rely exclusively on title searches.
3. **Inconsistent database capabilities:** To search a larger body of documents, an obvious measure would be to submit searches to a number of different academic search engines (for example, the "Scirus" search engine, or Google's Scholar search engine). Unfortunately, many search engines do not permit the searching of address lines explicitly. Even if it might be possible to include terms like "inc" and "MA" in the full text searches, we would still need to be able to specify that "inc" and "CA" be on the same address line.

To help counter these problems and also increase the overall quality and applicability of the approach, we propose the following avenues for future work:

1. **Intelligent feature extraction** - A variety of techniques from the machine learning and semantic technology communities could be brought to bear. In particular, it would be interesting to see the value of incorporating semantically-enabled features into the search process - i.e. instead of using manually generated keyword searches, computational techniques could be used to group together terms which are either synonymous, or which are observed to co-occur frequently, and to combine these terms appropriately when conducting the searches.
2. **Statistical analysis** of the search results - in this study, the data extraction process has largely been automated; however, the analysis of the results is still largely manual. While the final analysis of the results will likely always be manual, we hope to enrich

this process by providing more information to support users of this system. In particular, text mining techniques could be used to process the abstracts of retrieved documents - this can, for example, help users to identify transitions in the emphasis of research projects, and to visualize the evolution of this emphasis.

3. **Tools development** - thus far the analysis has been carried out using a collection of python scripts. While these have been very useful for our purposes, we plan to make these methods useable by a broader audience by creating a set of user-friendly software applications. These tools will incorporate the functionality of the scripts but in an intuitive and accessible way.

5.2 Innovation assessment

Our study suggests that using bibliometrics offers valuable insights for understanding the outcomes of government research expenditures, the institutional players involved in the emergence of an industry, the technological trajectories over the years, and in general the level of interest in a particular domain of knowledge. Especially for the case of renewable energy, such an analysis is important for laying out the foundation for further explorations.

The results from our analysis point to the close association between federal investment in R&D and knowledge production, as measured by number of publications. Especially in the early stages of industry development, 1970s and 80s, R&D funding programs proved to be critical for advancing science in solar photovoltaic technologies. Hence, policy-makers in the domain of science, technology and innovation, should ensure that especially for emergent industries, consistent investment in R&D is being made.

The high geographical concentration of national research labs in California, and the regional presence of space solar research at companies such as Spectrolab, created opportunities for a natural transition in the scientific community towards terrestrial solar applications research. National labs such as JPL and LBNL, and universities such as Caltech, Stanford, UC Berkeley, became the locus of research for U.S. as a whole. More recently, however, NREL in Colorado, concentrates the highest level of research on solar technologies at the national level.

Collaborative research between industry and the scientific community has been higher early on, having decreased more recently. Two reasons could explain this trend. First, until 1990s R&D funding from the federal government emphasized and required partnerships between different institutions (universities, national labs, and private sector) for caring out research activities. Second, collaborations between companies and academia might be easier and more valuable to engage with in the early stages of the industry. When the industry becomes more mature the level of competition increases, shifting the locus of research in companies' research laboratories.

However, the number of collaborative research papers between universities and industry does not reflect the true level of interaction between these institutions. Because of the different nature of these institutions, a large fraction of research outcomes do not end up in the public domain. Hence, these results need to be supplemented with additional information on specific university-industry contracts for research.

As the solar industry becomes more global and knowledge transfer surpasses national borders, it is interesting and relevant to explore the level of international collaborations thorough joint publications. In an extension of this research, we aim to explore who are the main countries and institutions collaborating with U.S. based scientists. Findings from such an

analysis would shed light on existing and growing research potential abroad.

Lastly, while results from our bibliometric analysis allow us to map the intensity of and interest in research over time in different institutions, we are not able to identify the impact that publications are having on the field of solar technologies as a whole. To get to this aspect, future research will need to take into account an analysis of papers citation patterns.

Bibliography

Anuradha, K., and Shalini (2007). Bibliometric indicators of Indian research collaboration patterns: A correspondence analysis. Scientometrics, 71(2):179–189.

Braun, T., Schubert, A. P., and Kostoff, R. N. (2000). Growth and trends of fullerene research as reflected in its journal literature. Chemical Reviews, 100(1):23–38.

Chiu, W.-T. and Ho, Y.-S. (2007). Bibliometric analysis of tsunami research. Scientometrics, 73(1):3–17.

Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological Forecasting and Social Change, 73(8):981–1012.

Daim, T. U., Rueda, G. R., and Martin, H. T. (2005). Technology forecasting using bibliometric analysis and system dynamics. In Technology Management: A Unifying Discipline for Melting the Boundaries, pages 112–122.

de Miranda, Coelho, G. M., Dos, and Filho, L. F. (2006). Text mining as a valuable tool in foresight exercises: A study on nanotechnology. Technological Forecasting and Social Change, 73(8):1013–1027.

Godin, B. and Gingras, Y. (2000). The Place of Universities in the System of Knowledge Production. Research Policy, 29: 273-278.

Godo, H., Nedrum, L., Rapmund, A. and Nygaard, S. (2003). Innovation in Fuel Cells and Related Hydrogen Technology in Norway-OECD Case Study in the Energy Sector, NIFU report 35/2003.

Kim and Mee-Jean (2007). A bibliometric analysis of the effectiveness of Korea's biotechnology stimulation plans, with a comparison with four other Asian nations. Scientometrics, 72(3):371–388.

Kostoff, R. N. (2001). Text mining using database tomography and bibliometrics: A review. 68:223–253.

Larsen, K. (2008). Knowledge Network Hubs and Measures of Research Impact, Science Structure, and Publication Output in Nanostructures Solar Cell Research. Scientometrics, 74(1): 123-142.

Losiewicz, P., Oard, D., and Kostoff, R. (2000). Textual data mining to support science and technology management. Journal of Intelligent Information Systems, 15(2):99–119.

Margolis, R.K. (2002). Understanding Technological Innovation in the Energy Sector: The Case of Photovoltaics. Doctoral Dissertation, Woodrow Wilson School of Public and International Affairs, Princeton University.

Martino, J. (1993). Technological Forecasting for Decision Making. McGraw-Hill Engineering and Technology Management Series.

Porter, A. (2005). Tech mining. Competitive Intelligence Magazine, 8(1):30–36.

Porter, A. (2007). How "tech mining" can enhance R&D management. Research Technology Management, 50(2):15–20.

Saxenian, A. (1996). Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Cambridge, Harvard University Press.

Scott, A. (2003). Flexible Production Systems and Regional Development: The Rise of New Industrial Spaces in North America and Western Europe. In Barnes, T.J. et al. (Eds.), Reading Economic Geography, Wiley-Blackwell.

Smalheiser, N. R. (2001). Predicting emerging technologies with the aid of text-based data mining: the micro approach, Technovation, 21(10):689–693.

Small, H. (2006). Tracking and predicting growth areas in science. Scientometrics, 68(3):595–610.

Tsay, M. (2008). A Bibliometric Analysis of Hydrogen Energy Literature 1965-2005. Scientometrics, 75(3): 421-438.

Zimmerman, E., Wolfgang, G., and Bar-Ilan, J. (2009). Scholarly Collaboration Between Europe and Israel: A Scientometric Examination of a Changing Landscape. Scientometrics, 78(3): 427-446.

List of Figures

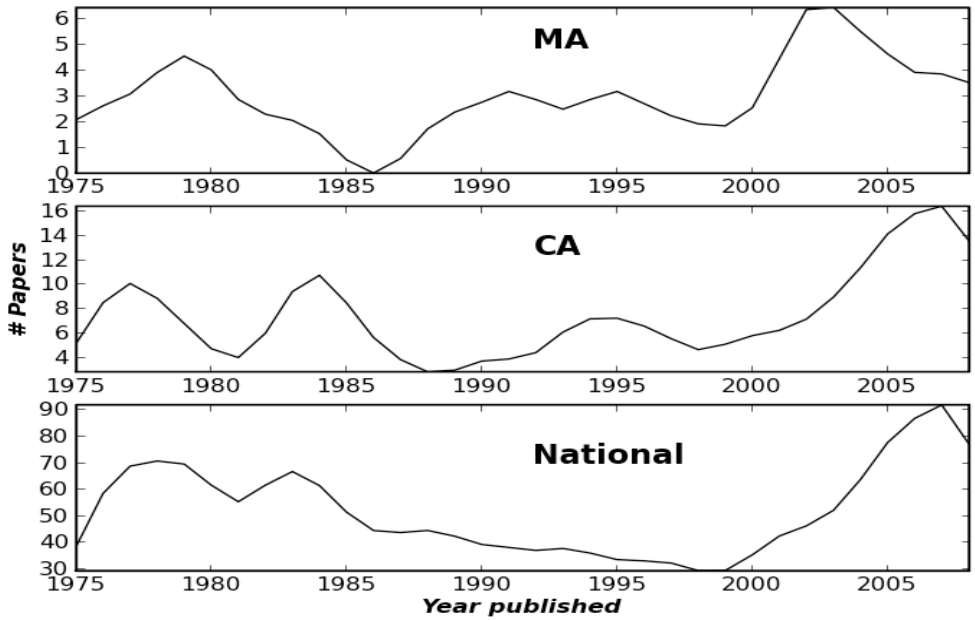


Figure 1: Research at universities

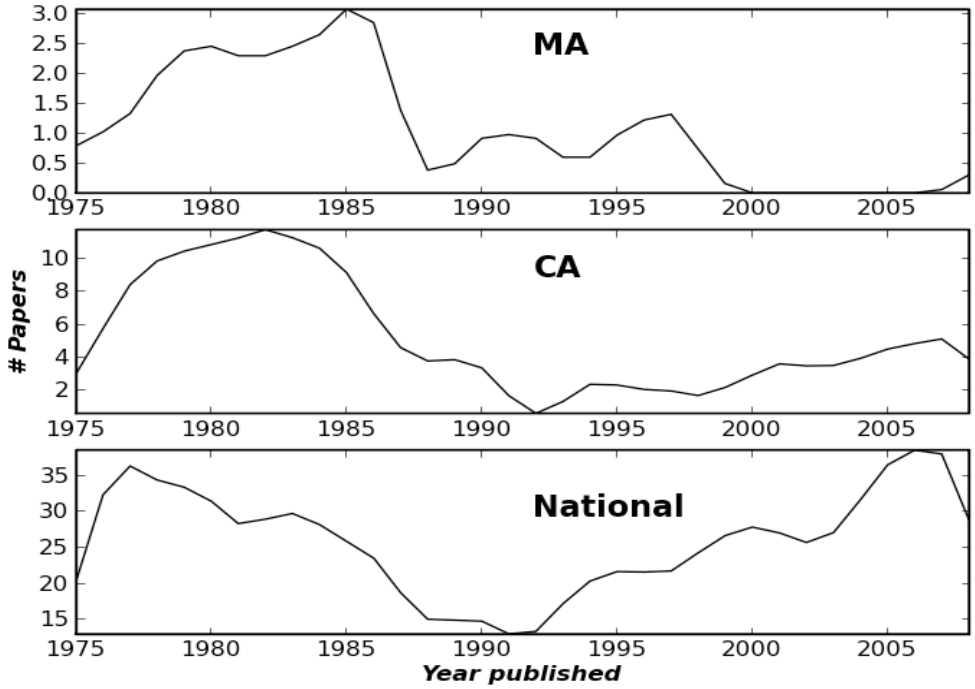


Figure 2:Research at national research laboratories

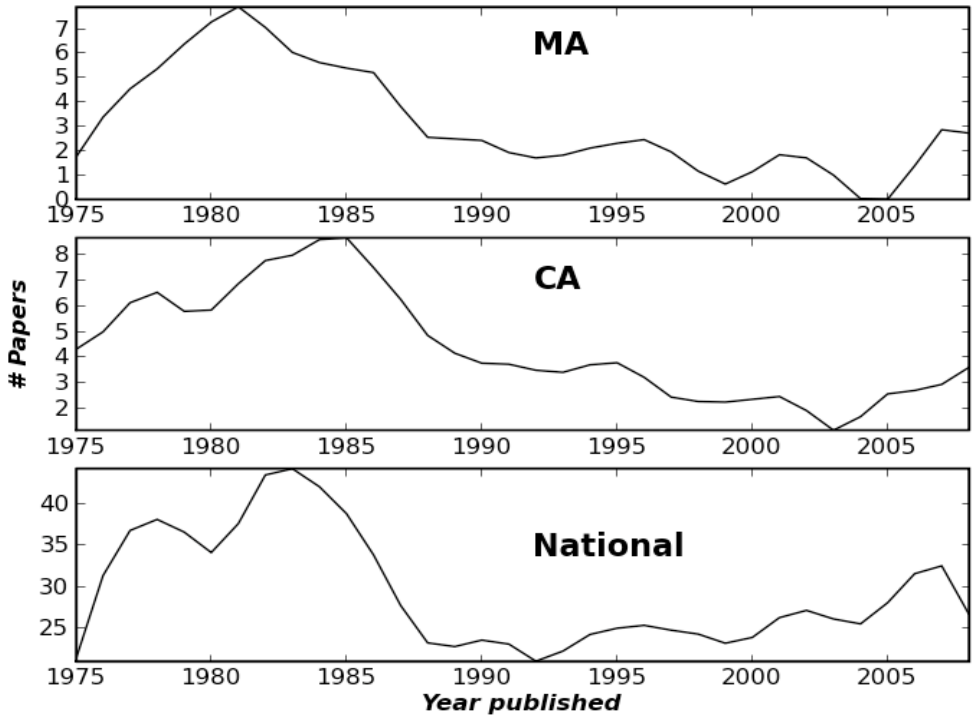


Figure 3: Research by Industry

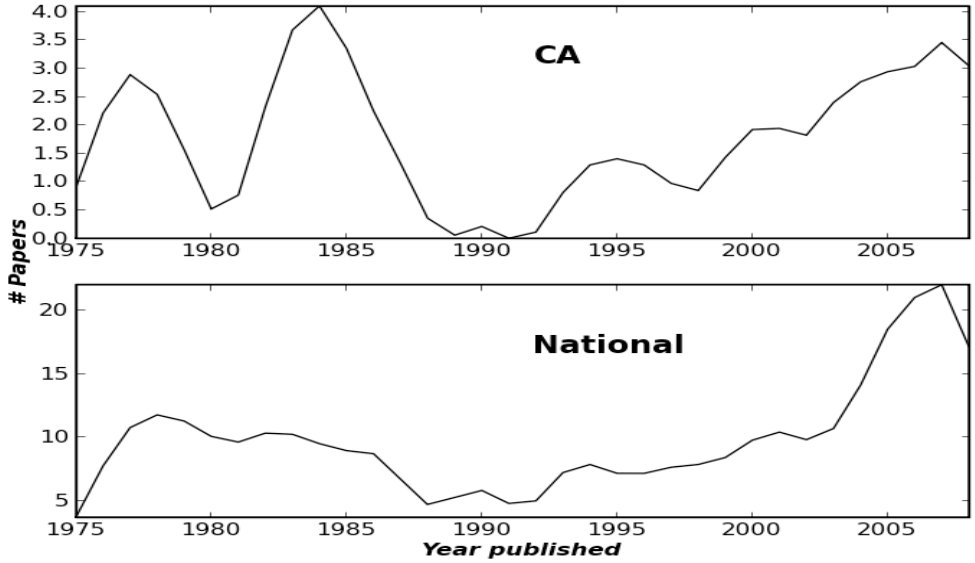


Figure 4: Laboratory-University collaborations

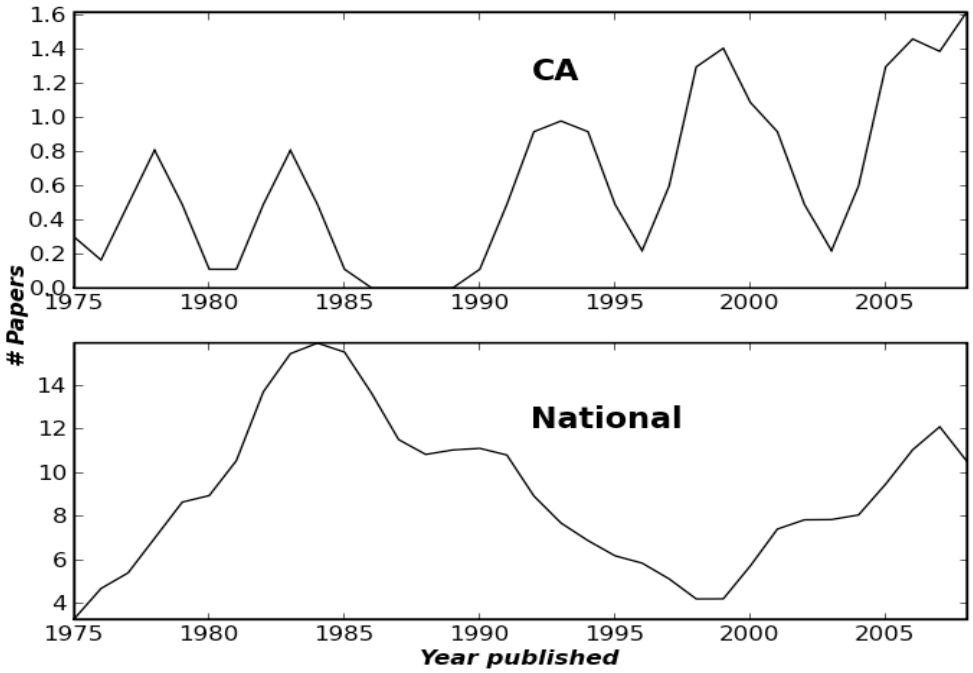


Figure 5: Industry-University collaborations

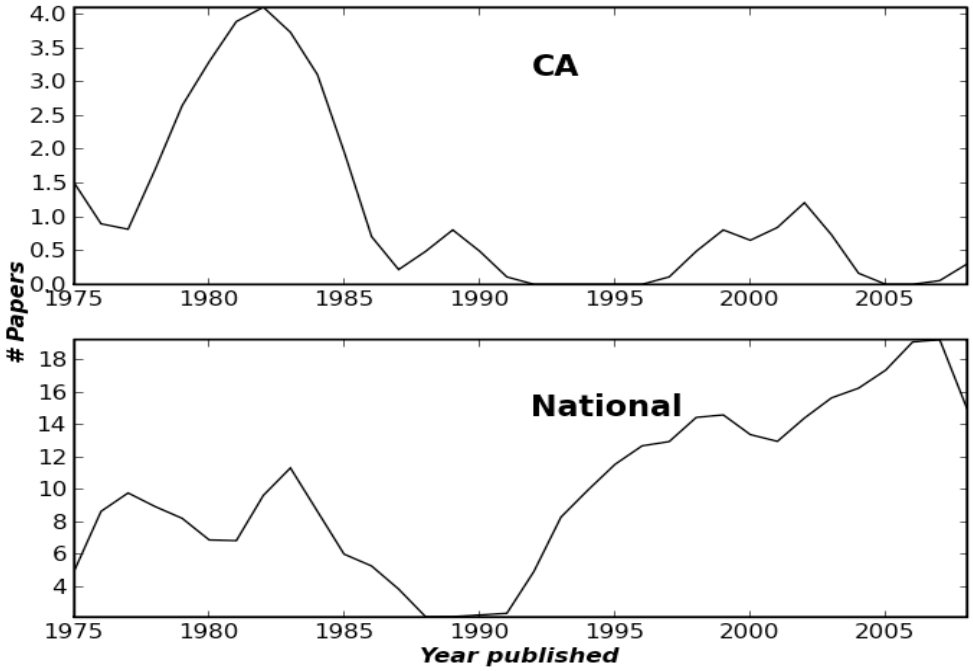


Figure 6: Industry-Laboratory collaborations

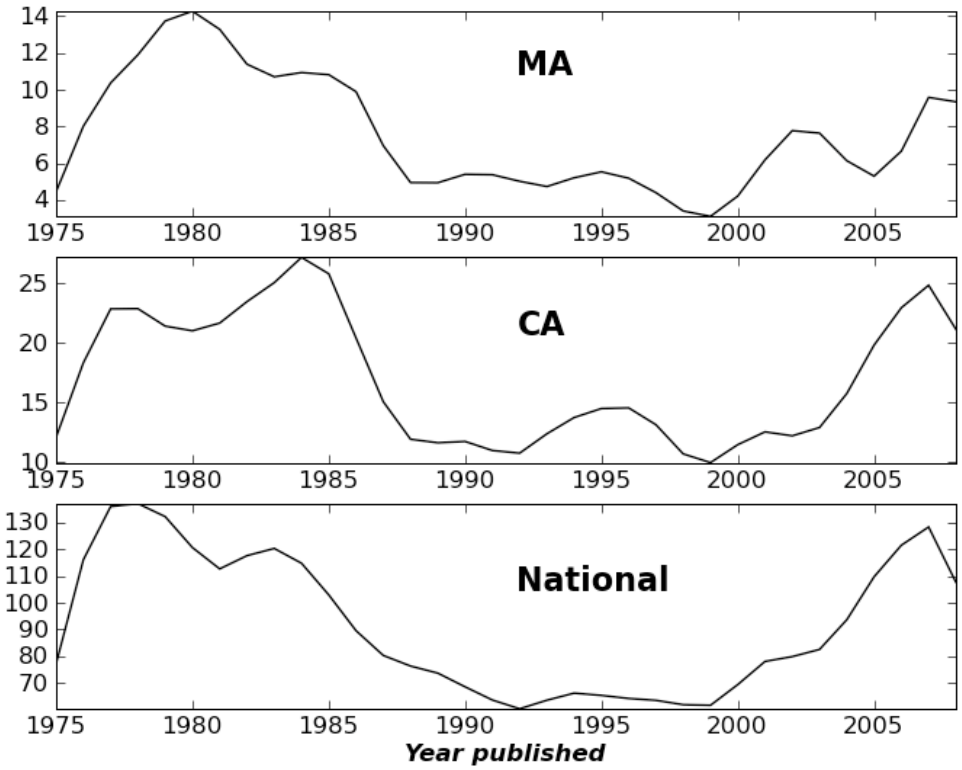
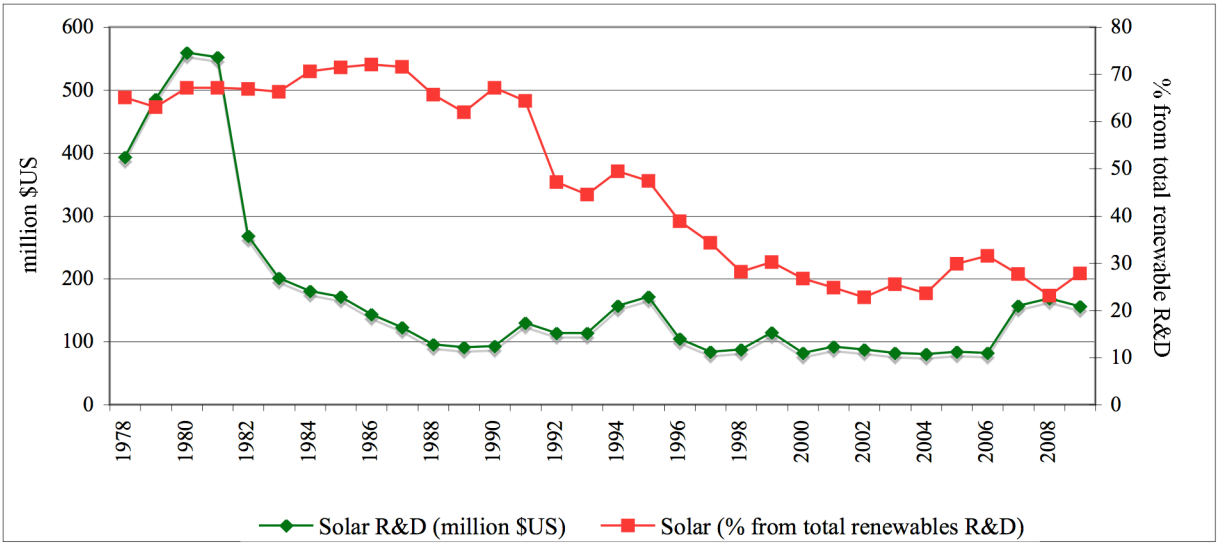


Figure 7: Solar related research in the U.S. (from any type of institution)



Note: Solar includes biofuels, wind, and ocean up to 1998.
Source: Gallagher, K.S., Sagar, A, Segal, D, de Sa, P, and John P. Holdren, "DOE Budget Authority for Energy Research, Development, and Demonstration Database," Energy Technology Innovation Project, John F. Kennedy School of Government, Harvard University, 2006. Database updated by Kelly Gallager, February 2008.

Figure 8: R&D Federal spending on solar related research between 1978 and 2009