Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Department of Economics

Knowledge production and spillovers

A patent citation study of the development of PV technology

Li Stoorhöök

Master's thesis · 30 hec · Environmental Economics and Management , Master's programme · Degree thesis No 1071 · ISSN 1401-4084
Uppsala 2017

Knowledge production and spillovers - A patent citation study of the development of PV technology

Li Stoorhöök

Supervisor: Rob Hart, Swedish University of Agricultural Science, Department of Economics

Examiner: Ing-Marie Gren, Swedish University of Agricultural Science, Department of Economics

Credits: 30 hec
Level: A2E
Course title: Independent Project/ Degree in Economics
Course code: EX0537
Programme/Education: Environmental Economics and Management, Master's Programme
Faculty: Faculty of Natural Resources and Agricultural Sciences

Place of publication: Uppsala
Year of publication: 2017
Name of Series: Degree project/SLU, Department of Economics
No: 1071
ISSN 1401-4084
Online publication: http://stud.epsilon.slu.se

Keywords: DTC, patent data, PV, spillovers, technological change

Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Department of Economics

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Rob Hart, for the invaluable guidance, support and encouragement he has given me throughout the process.

My sincere thanks also go to Geer Boedt at the European Patent Office (EPO), whose generous technical support with the data collection in PATSTAT, was absolutely crucial for the thesis.

Abstract/ Summary

The purpose of this study is finding better understanding of the process of knowledge growth, and how technological knowledge builds on existing knowledge.

To be able to empirically examine the subject we performed a case study of the photovoltaic (PV) energy sector. The theoretical framework used is Hart's (2016) model of knowledge production, which takes into account flows of knowledge between sectors. The model is tested through OLS- and logistic regression, using European patent data for time time period 1977-2009.

We find that spillovers have had a significant impact on the knowledge growth of the PV sector, and that the most important contributions have come from closely related technology fields. We also find that geographical origin is an important parameter for determining spillover patterns.

Table of Contents

1 INTRODUCTION **1**

2 LITERATURE REVIEW **3**

2.1 SPILLOVERS 3
2.2 ECONOMIC THEORY OF TECHNOLOGICAL CHANGE 4
2.3 EMPERICAL RESEARCH OF SPILLOVERS 6

3 THEORETICAL FRAMEWORK **9**

3.1 KNOWLEDGE PRODUCTION 9

4 METHOD AND DATA **14**

4.1 PATENT CITATION DATA 14
4.2 DATA COLLECTION 15
4.3 CATEGORIZATION PROCEDURE 15
4.4 VARIABLES 18
4.4.1 Geographical Origin 22
4.4.2 Technological Origin 24

5. MODEL ESTIMATIONS **27**

5.1 MODELS 27
5.1.1 Results 27
5.1.2 Diagnostics 29
5.1.3 Conclusions 33
5.2 LOGISTIC MIXED EFFECT MODELS 34
5.2.1 Models 34
5.2.2 Results 35
5.2.3 Robustness check 37
5.2.4 Conclusions 39

6 DISCUSSION **41**

7 CONCLUSIONS **43**

REFERENCES **45**

APPENDICES **47**

APPENDIX A: PV-ENERGY 47
APPENDIX B: CPC CLASSIFICATION SCHEME 49
APPENDIX C: PATENT STOCK SIZE 50
APPENDIX D: CITINGS PER CPC SYMBOL 51
APPENDIX E: ADDITIONAL REGRESSION OUTPUTS 52
APPENDIX F: Q-Q PLOTS 54

1 Introduction

Peak oil and acknowledgement of the threat of global warming manifests the need for radical shifts on either the demand side, or the supply side for energy. In the public debate focus has foremost been set on the supply side – finding alternative energy sources that can cope with the growing demand. Solar energy is one of those alternatives – it has low environmental impact, and the supply of radiation from the sun is constant. However, today renewable energy only responds to five percent of the total global energy production, and in order to be able to in the long run replace fossil fuels as the dominating energy source, vast technological development in the field is necessary. Technical development is however by definition uncertain, as is the political will to achieve it.

Arthur (2007) argues that a difficulty all theories of innovation face is that modern research show that the process of invention varies significantly from historical case to historical case, thus it is questionable whether universalities actually exists. Although, he continues, the *incentives* driving invention are less of a mystery: social needs, economic opportunities, perceived risk, factor price changes are some of the suggested explanatory variables behind technological change. Clarke et al. (2008) claim that R&D, learning-by-doing and spillovers as the three main sources of technological change. In this thesis we will focus on the last of those three sources.

In economic theory of innovation spillovers have often been overlooked. Acemoglu's (2002, 2009) influential model of directed technological change is based on the assumption that technological knowledge in different sectors grows separately, and that there are no knowledge spillover between the sectors. This assumption is unintuitive and needs to be both theoretically and empirically examined.

The aim of this study is finding better understanding of the process of knowledge growth. To be able to empirically examine the subject we have chosen to perform a case study of the photovoltaic (PV) energy sector. The theoretical framework of the study is Hart's (2016) model of knowledge production, which takes into account flows of knowledge between sectors. We hope to find some of the determinants of knowledge growth and in particular how new technological knowledge builds on existing knowledge.

The reason for choosing the PV sector is that it is a relatively young, and fast growing, knowledge field that only existed since late 1950's and found market adaptations in the 1970's. Thus it will be possible to map the technology's evolution from its beginning until present date. Our interest is to see how the knowledge stock of PV is linked to the broader concept of “general knowledge”. We further want to see whether these linkages changes over time, or if they are constant.

Hart’s (2016) model of knowledge production will be tested empirically using European patent citation data for the period 1978-2012. The results indicate that we have inter-industry knowledge spillovers, and that they have had a significant impact on the knowledge production in the PV sector. Further, we find that the marginal productivity of within sector knowledge is decreasing.

This thesis will solely focus on the *knowledge production* of photovoltaic energy, and particularly on the role of spillovers for the technological field. Thus demand mechanisms and growth paths will not be discussed more that briefly, even though spillovers are highly relevant for understanding those processes. The results of this thesis are based on the particular case of PV-energy. To gain greater understanding of the knowledge production process in a broader set of industries and knowledge fields, further research on the subject is needed, including other technology fields. For a brief review of the technological development of PV-energy, see appendix A.

The disposition is as follows: in section 2 we review the existing literature in the field of directed technological change and spillovers, in section 3 we present the theoretical framework, in section 4 we present the method and data, in section 5 we perform the model estimations, in section 6 we discuss the results and in section 7 we present our conclusions.

2 Literature review

2.1 Spillovers

Clarke et al. (2008, p. 413) defines spillovers as: ”technological change in one firm, industry, country or domain of technology that arises from innovative activities in another firm, industry, country or domain of technology”.

Spillovers can be either “direct” or “indirect”. Direct spillovers alter technology with no required effort on the part of the receiving industry. Indirect spillovers creates a “pool of opportunities” that can be exploited by receiving industries, but which will have no impact unless these opportunities are taken.

It is also possible to distinguish between “rent spillovers” and “knowledge spillovers”, where the first refers to the transfer of economic benefits, and the latter to the transfer of knowledge. Rent spillovers lower the cost or the quality of inputs in receiving industries, and thus improve their profitability. Knowledge spillovers can either be direct or indirect, depending on how much effort receiver must put in to use the knowledge spillover (Clarke et al. 2008). Direct knowledge spillovers have no influence on own-industry activities. They can simply be added to the knowledge production already carried out by the actors in the industry. In contrast, indirect knowledge spillovers have a more complementary relationship with own-industry activities. Spillovers cannot be utilized without own-industry effort, and furthermore, the technological opportunities created by the spillovers are necessary for the ongoing knowledge production of the industry (Clarke et al. 2008).

The pool of technological opportunities that is created by indirect knowledge spillovers grows when it is not exploited, and creates a room for rapid boost for the own-industry knowledge growth once it is utilized. The knowledge growth rate will decline over time as the pool is being exhausted. That implies that if the rate of exploitation of indirect spillovers is high relative to their creation, the growth rate of spillovers may be defining for the rate of the technological change in that industry (Clarke et al. 2008).

It is important to note the distinction between spillovers and own-industry activities is a matter of level of aggregation chosen for the particular study (Clarke et al. 2008). A more narrowly defined industry implies a higher degree of spillovers, simply by the fact that fewer activities are considered as own-industry efforts. Therefore it is meaningless to compare spillover rates between different industries, not taking this into account. What can be more interesting is comparing spillover rates within the same industry at different points in time and between countries.

2.2 Economic Theory of Technological Change

The process of technological change has long been discussed and modeled by economists. In early models of economic growth (see e.g. Solow, 1956) technological development is the driving force behind economic growth, but its mechanism is left unexplained and the phenomenon is viewed as something exogenously given. Later theorists have tried to explain this force as a result of investment in research, performed by profit maximizing firms as a method for reducing production costs (Aghion & Howitt 2008, pp 12-14). However, not only the acknowledgement of an active process of knowledge production is enough to explain the *differences* in growth between different sectors in the economy. Some sectors experience a rapid technological development, while others hardy develop at all. Already in the 1930's John R. Hicks introduced the hypothesis of induced bias in innovation, as an attempt to explain the direction of technological change. According to the hypothesis, a rise in the price of labor relative to capital will induce labor-saving innovations (Hicks, 1932). The idea behind the hypothesis is that R&D is a profit maximizing investment activity, and that innovation in a certain field is responding positively to increases in relative prices (Popp et al., 2009). In the 1960's this hypothesis was served new attention when it became subject of a lively debate.

The debate was centered on two alternative models – a growth-theoretic approach and a microeconomic version. The most formally developed model was the growth theoretic approach introduced by Kennedy (1964, 1966 and 1967) and Samuelson (1965 and 1966) (Ruttan, 2001). Kennedy (1964) presents the Innovation Possibilities Frontier (IPF), which shows that there is a trade-off frontier between capital- and labor augmenting rates. Unlike Hicks, Kennedy claims that relative factor prices are not essential for the theory. Instead he suggests that the entrepreneur will search for improvements that reduce the *total unit cost in greatest proportion*. Thus if labor-costs are higher than capital-costs, investments in labor-saving innovations are encouraged. The model can be seen as an attempt to explain the empirical phenomenon of constant factor shares. The reason technical development has been labor-saving is that the labor force is more or less fixed, while the capital stock is possible to increase through investments. That means that in order to increase the efficient number of worked hours you need to increase labor productivity. Since there is a trade-off between the capital- and labor augmenting rates the increasing labor costs will induce labor saving innovations.

Nordhaus (1973) criticizes the theory of induced innovation for building upon unrealistic, and unmentioned, assumptions. His critique especially concerns the assumption that the shape of the IPF is independent of the levels of factor augmenting knowledge - that is that previous investments in labor- or capital-saving innovations will not affect future possibilities for technological improvements. Binswanger (1974) gives a microeconomic approach to induced innovation that offers support to Nordhaus's critique. Going through the implicit assumptions of Kennedy's IPC, he claims it only to be a disguised example of exogenous technical change. Instead he argues that the variables that determine induced bias in innovation are the relative productivity and price of alternative research lines, the scale of output and changes in present value of factor costs, since these changes the optimal research mix.

The microeconomic approach was developed by Ahmad, (1966) and build directly on Hicks induced bias hypothesis. They use the concept of the historic innovation possibility curve (IPC), according to which it at each point in time it exists a set of potential production processes available to be developed. Which these processes are will be determined by the basic state of knowledge at the given time. Each process in the set can be characterized by an isoquant, and the IPC is the envelope of all unit isoquants at a given time (see also Ruttan, 2001).

More recent work in the field of directed technological change (DTC) is done by Acemoglu, (1998, 2002); and Acemoglu et al., (2009), whose framework has been widely used in subsequent work in the field. Acemoglu (2002) presents a model for knowledge-based R&D, which allows for "state dependence". State dependence means that current investments in R&D in a particular sector enhance the productivity of future knowledge production of that sector. That has the direct consequence that lagging sectors will have trouble catching up since the leading sectors are much more productive.

From this set-up, Acemoglu (2002) presents two major results. The first is the "weak induced-bias hypothesis", which states that as long as the elasticity substitution between factors is not equal to 1, an increase in the relative abundance of a factor will always create some amount of technical change directed towards it. The second result is the "strong induces-bias hypothesis, which says that given that the elasticity of substitution is sufficiently large (between 1 and 2), directed technological change towards the more abundant factor can result in an upwards sloping long-run relative demand curve for that factor.

Acemoglu et al. (2010) further develops the framework and applies it to an economy with environmental constraint, producing one final good using either a clean- or a dirty input, or a combination of the two. The dirty good will produce environmentally damaging externalities, and exhaust the stock of environmental capacity. When the stock is fully exhausted we have an environmental disaster. An intermediate goods sector of monopolistically completive research firms choses in each period individually to direct the firm's research to either the dirty or the clean sector. The research activity enhances the productivity of the chosen sector. Without policy intervention, directed at distorting the historically given advantage of the dirty sector, the economy will reach a socially suboptimal equilibrium. However, given that the inputs are strong substitutes, only a temporary intervention is necessary in order to tilt the balance and start a positive feedback loop that will offset the previous advantage of the dirty sector. This scenario is based on the implicit assumption of extreme path dependence, i.e. that the knowledge stocks of the two sectors evolve separately.

Hart (2016) suggests in a working paper an alternative model of directed technological change. In the model, which will be referred to as *knowledge spillovers*, the knowledge stocks of different sectors are linked together, so that the knowledge production in one sector can feed of the existing knowledge in other sectors. These knowledge spillovers countervail imbalances between knowledge stocks, as knowledge spillovers from the leading sector to the lagging sector enhance the relative growth of the latter. As the lagging sector catches up, it can make less and less use of the knowledge spillovers from the leading sector, and becomes more and more dependent on within sector knowledge.

2.3 Emperical research of spillovers

Geographical and technological proximity is often seen as the main factors fostering innovation. Geographical clusters of firms leads to knowledge spillovers that they all can benifit from in terms of innovative capabilites, but it also comes with the risk of spreading key knowledge to competitors which could be harmful for the business (Wersching, 2005, p. 2).

The amount of knowledge spillover a firm is able to make use of is called the absorptive capacity. It sets a lower and upper limit for the heterogeneity of the knowledge a firm is able so absorb (Cohen and Levinthal, 1989, 1990). A knowledge inflow that is too far from the firms' knowledge base is too difficult for the firm to incorporate in its own business. Similarly, knowledge that is too close the firm's knowledge base becomes trivial and will not boost the innovative capacity of the firm. The relationship is by Werchish (2005) described as an inverse U-shape, depending on technological distance.

Orlando (2002) showed that the largest R&D spillovers flow between firms in the same industry. Although location can play a part, they found the geographical distance to between narrowly defined technology groups doesn't impact on the spillovers. However, for spillovers from firms outside the industry, geographical distance is an important factor. Jaffe (1986, p. 2) similarly argued that firms with higher R&D investments than the average of the industry benefits from R&D spillovers from "neighbouring firms in technological space", while the opposite is true for firms with lower than average R&D spending. Griliches (1992) gives examples of a large number of studies showing that R&D spillovers contributes significantly to the productivity increases of both the agriculture and the industry.

Nemet, (2012a) examined the impact of inter-industry knowledge flows on the inventions' value. They presented the c*umulative synthesis hypothesis*, according to which new inventions arise from new combinations of current knowledge. Using United States patent data for the time period 1976-2006, where backwards citations were used as a proxy for knowledge flows, and forward citations were used as a proxy of the value of the invention, they found that citations from external technological fields were less positively – or even negatively – correlated to the patents value. Nemet discussed several reasons for this result to appear. One possible explanation was that knowledge flows from external technological fields are associated with higher risk. Even if the invention resulting from inter-industry knowledge flows is commercially successful itself, it might be difficult to assimilate the knowledge into new inventions. Other explanations presented evolved around measurement errors.

In a similar study Nemet (2012b) used a slightly different approach by only looking at knowledge spillovers in energy technologies. Here, in contrast, inter-technology knowledge spillovers played an important role in the evolution of the technologies. Important energy patents have drawn heavily from external knowledge, especially in the fields of chemical, electronics and electrical technologies.

Jaffe et al. (2000) surveyed both cited and citing inventors around specific innovations, and found that patent citations did represent knowledge flows, but with a substantial amount of noise. Jaffe (International Knowledge Flows: Evidence from Patent Citation 1999) showed that patents are 100 times more likely to cite a patent with the same patent class, than patents from other classes. Further they showed that patents from inventors that resided in the same country were 30-80 percent more likely to cite each other than inventors from other countries.

Liu et al., (2011) analyzed the photovoltaic energy growth trajectory, by examining US patents for the time period 1974-2007. They found that the technology growth was highly correlated with the crude oil price, by a time lag of about one year. Furthermore, they found that the time gap between technology growth and market growth is ten years, meaning that it took about ten years from the time a new technology is developed, to the time it reached the market. In their analysis they separated PV energy into five categories: the Emerging PV, CdTe, CIS/CIGS, Group III-V, and Silicon and argued that all of the five categories showed a S-shaped growth trajectory, with two distinct phases of technological growth. The first phase generally halts around 1985-1990, coinciding with the fall of the crude oil prices. It is also shown that two of the technology groups, CdTe an CIS/CIGS have reached the mature phase of their life-cycle, while the other technologies has yet to reach that point. They are expected to reach that point between 2014-2018.

3 Theoretical framework

In this section we will present Hart's (2016) model for knowledge production and knowledge flows. According to the model, knowledge production is a function of the number of existing patents in the different sectors of the economy. Knowledge flows are represented by patent citations, through which knowledge embodied in patents spawns new innovations (and thus new patents).

The model is based on four concepts; *knowledge stocks*, *patent stocks*, *patent flow and knowledge flow*. The *patent stocks* are the existing number of usable patents in each sector *i*, and time period *t*. Usable is referring to the fact that the information contained in patents can only be used if it is available to inventors, and that the information is not outdated.

The second important concept *is knowledge stocks*, which are built up by patent stocks. Since the model allows for spillovers between the sectors, the patent stock of one sector can affect the size of the knowledge stocks in other sectors. The knowledge stocks can be described as the collected amount of existing knowledge that can be used to produce new patents in each sector.

The third concept - the *flow of new patents* - is an increasing function of the patent stocks and investment. Finally, *knowledge flows* from old patents to new through patent citations. The number of patent citations is in proportion to the value of the patent stocks.

3.1 Knowledge production

We assume two technology sectors – PV, which we denote as *p*, and non-PV, which we denote as *n*. The productivity growth of each sector depends on new innovations that are obtained as a result of investments in research, *R*.

We further assume that there exists two types of patents, PV-patents, $x_{p,t}$, and non PV-patents, $x_{n,t}$, and that all knowledge is embodied in these.

All new knowledge builds upon previous findings, so that the stock of knowledge increases step by step. Each innovation is represented by a patent, which once it is made public contributes to the existing knowledge stock of that sector. We have

$$\boldsymbol{x_{p,t} = K_{pt}^{\Psi} R_{pt}^{\emptyset}}, \tag{1}$$

where $x_{p,t}$ is the number of new PV patents produced in period *t*, i.e. the *PV patent flow*. The flow is an increasing function of the PV knowledge stock $K_{p,t}$ and sector specific investments

$R_{p,t}$. Time is discrete and Ψ and $\emptyset$ are positive parameters. Correspondingly, for the non-PV sector we have

$$x_{n,t} = K_{nt}^{\Psi} R_{nt}^{\emptyset}. \tag{2}$$

Thus new patents build upon the sector specific knowledge stock. However, the *PV knowledge stock* is generally an increasing function of *both the PV- and non-PV patent stock*, since knowledge valuable for new PV inventions can exist in both sectors. The non-PV knowledge input in the PV production is what we call knowledge spillovers. The two knowledge stocks, K_{pt} and K_{nt} are modeled as two CES-functions,

$$K_{pt} = \left(\left(1 - \frac{\sigma}{2}\right) X_{pt}^{\varepsilon} + \frac{\sigma}{2} x_{nt}^{\varepsilon}\right)^{\frac{1}{\varepsilon}} \tag{3}$$

and

$$K_{nt} = \left\{\left(1 - \frac{\sigma}{2}\right) X_{nt}^{\varepsilon} + \frac{\sigma}{2} x_{pt}^{\varepsilon}\right\}^{\frac{1}{\varepsilon}} R_{pt}^{\emptyset} \tag{4}$$

where X_{it}^{ε} is the patent stock of sector *i,* and σ and ε are parameters between 0 and 1. σ is a scale parameter that determines the relative weight of PV knowledge compared to non-PV knowledge for making new PV innovations. If σ is smaller than 1, then PV knowledge is relatively more important for new PV innovations than non-PV knowledge, and vice versa. ε is the degree of substitutability between PV and non-PV knowledge. It is these two parameters that together determine the size of the spillover between the two sectors. For instance, when σ is 0 we have no spillovers. When σ is 1, the value of the knowledge is sector neutral, and citation rates will only depend on the relative size of the two sectors.
By definition the information contained in patents are, once it is published, free and open for everyone to exploit. Hence the knowledge stocks $K_{i,t}$ and patent stocks $x_{i,t}^{\varepsilon}$ are public goods which cannot be traded, and have no market price. Nevertheless, for researchers, the knowledge is highly valuable, and therefore we can talk about shadow prices, i.e. the price the researcher would be willing to pay for the knowledge. The shadow prices will in this context be used to predict citation rates.

Hart (2016) exemplifies this by considering a firm that wants to invest in research on PV technology. The research success is measured in the number of patents a given amount of research effort yields. The expected shadow value of one patent is denoted v_t, and depends on the how the knowledge embodied in the patent can be used. v_t can be used to find expressions for the shadow prices of the knowledge stocks K_{pt} and K_{nt}.

We substitute the expression for the PV-knowledge stock into the patent production function (1) to obtain

$$x_{p,t} = \left\{\left[\left(1-\frac{\sigma}{2}\right)X_{p,t}^{\varepsilon} + \frac{\sigma}{2}x_{n,t}^{\varepsilon}\right]^{\frac{1}{\varepsilon}}\right\}^{\Psi} R_{p,t}^{\emptyset} \qquad (5)$$

The firm wants to maximize its shadow revenue minus costs:

$$\max \; v_t x_{p,t} - w_{r,t}R_{p,t} \qquad (6)$$

where $w_{r,t}$ is the price of the research input. Given the maximization problem, the shadow prices of the patent stocks X_{it}^{ε} must equal the marginal shadow revenues resulting from an increase in knowledge. We denote the shadow prices of the knowledge stocks ξ_p and ξ_n respectively and get the following expressions:

$$\xi_{pt} = v_t \frac{\partial x_{p,t}}{\partial X_{p,t}} \qquad (7)$$

and

$$\xi_{nt} = v_t \frac{\partial x_{nt}}{\partial X_{nt}}. \qquad (8)$$

We solve these to yield:

$$\xi_{pt} = v_t \Psi \frac{x_{pt}}{K_{pt}} K_{pt}^{1-\varepsilon}\left(1-\frac{\sigma}{2}\right)X_{pt}^{\varepsilon}X_{pt}^{-1}, \qquad (9)$$

and

$$\xi_{nt} = v_t \Psi \frac{x_{pt}}{K_{pt}} K_{pt}^{1-\varepsilon}\left(\frac{\sigma}{2}\right)X_{nt}^{\varepsilon}X_{nt}^{-1}. \qquad (10)$$

So the relative shadow values of the patent stocks $X_{p,t}^{\varepsilon}$ and $X_{n,t}^{\varepsilon}$ are:

$$\frac{\xi_{p,t}X_{p,t}}{\xi_{n,t}X_{n,t}} = \frac{2-\sigma}{\sigma}\left(\frac{X_{p,t}}{X_{n,t}}\right)^{\varepsilon}. \qquad (11)$$

We will use these shadow values to find the *citation rates*. We define c_{pt} as the average number of citations to (old) PV patents per new PV patent, and c_{nt} as the average number of citations to (old) non-PV patents per new PV patent. We assume that the probability of a researcher citing a certain patent is proportional to the shadow value of that patent. Hence, the expected PV citation rate, $\frac{c_{pt}}{c_{nt}}$ is proportional to the shadow value of the PV patent stock:

$$\boldsymbol{E}\left[\frac{c_{pt}}{c_{nt}}\right] = \frac{\xi_{pt}X_{pt}}{\xi_{nt}X_{nt}}. \tag{12}$$

Insert equation (4) in (5) to yield:

$$\boldsymbol{E}\left[\frac{c_{pt}}{c_{nt}}\right] = \frac{2-\sigma}{\sigma}\left(\frac{X_{pt}}{X_{nt}}\right)^{\varepsilon} \tag{13}$$

From (13) we see that the citation ratio depends on the relative size of the two patent stocks $\frac{X_{pt}}{X_{nt}}$, the spillover effect, σ, and the degree of substitutability between PV and non-PV, ε. The closer σ is to zero, the smaller is the spillover effect, and the more dependent is the PV knowledge production on within sector knowledge for making new advances. If σ is large, existing PV knowledge is less essential for the development of the sector, since it can make use of the vast pool of knowledge existing in the non-PV sector. Similarly, a high degree of substitutability dampens the negative effect of a small PV patent stock, since the citing ratio can be adjusted to the knowledge supply. If ε is smaller than 1, the marginal productivity of X_{pt} is decreasing. The intuition behind that is that when the PV-patent stock is small, each patent is very valuable for spurring new innovations. While the patent stock is growing, each PV-patent becomes less essential for new related findings.

Special case 1: σ=0

When σ is zero, we have no knowledge spillovers. The PV knowledge sock is simply a function of the PV patent stock: $K_{pt} = X_{pt}$. From that it follows that the degree of substitutability, ε, also must be zero since there is only one input. The citation ratio is therefore:

$$\frac{c_{p,t}}{c_{n,t}} = \infty. \tag{14}$$

This is the special case that Acemoglu refers to as “extreme path dependence”, which implies that the historical advantage of dominating sectors offsets the opportunity for smaller sectors to grow and gain market shares (Acemoglu 2009).

With values of σ close to zero it is very difficult, or even impossible, for new technologies and small sectors to evolve over time, as they are depending on the limited (or zero) existing knowledge within the sector.

Special case 2: 0>σ>1, ε=0

When ε approaches zero equation (3) becomes a Cobb-Douglas function (cf. Acemoglu 2002, 6201X):

$$K_{pt} = X_{pt}^{\left(1-\frac{\sigma}{2}\right)} X_{nt}^{\left(\frac{\sigma}{2}\right)} .$$

(15)

When $\varepsilon=0$, the relative citation ratio is independent of the relative size of the patent stocks, and thus is constant over time (assuming σ is constant):

$$E\left[\frac{C_{pt}}{C_{nt}}\right] = \left(\frac{2}{\sigma} - 1\right).$$

(16)

When the PV patent stock is small relative to the non-PV patent stock, each PV patent has a higher marginal product, so that the total value of the patent stock in terms of contribution to new knowledge production is constant.

Special case 3: 0>σ>1, ε=1

When ε=1 we have a linear function, where the marginal product of the patent stocks is constant:

$$K_{pt} = \left(1 - \frac{\sigma}{2}\right) X_{pt} + \frac{\sigma}{2} X_{nt}.$$

(17)

In this setting it translates to that the knowledge kept in a single patent has the same marginal value no matter the size of the patent stock. Therefore, the relative number of PV-citations increases linearly with the relative size of the PV patent stock:

$$E\left[\frac{C_{pt}}{C_{nt}}\right] = \left(\frac{2}{\sigma} - 1\right) \frac{X_{Pt}}{X_{nt}}$$

(18)

4 Method and data

4.1 Patent citation data

A patent is a temporary monopoly awarded to inventors for the commercial use of a newly invented device. In exchange for the monopoly rights, the inventor is obliged to publicly disclose the technical solution of the invention. In order for the patent to be issued, the invention must be considered non-trivial, and to have industrial application (PRV 2013).

Innovations are functions of previous technological findings and scientific breakthroughs. In patents, the linkages between new and proceeding innovations are revealed in the patent citations. The patent applicant has legal duty to disclose any knowledge of the prior art, by referring to other patents and scientific articles. This serves an important legal function by delimiting the scope of the property rights of the patent holder (Trajtenberg et al., 1992, p. 7). Patent data have been vastly used in studies concerning technological development and is a good source of information for this purpose, since it includes detailed information about each invention's contribution to the knowledge stock, and knowledge spillovers between technological groups, counties, companies etc.

A patent document usually include information on: the name of the applicant, the name of the inventor, the application date (priority date), legal status, patent family[1], citations and classifications according to the CPC-, USTPO, and/or the IPC-classification system. The IPC is a hierarchical system, which uses language-independent symbols to classify patent documents[2] according to the area of technology to which they relate, and is used by patent offices worldwide.[3] The Cooperative Patent Classification (CPC) came into force in January 2013, and is a bilateral system of patent classifications, jointly developed by the European Patent Office (EPO) and the ***United States Patent and Trademark Office*** (USPTO).

Whether patent data is to consider representative for innovative activity have been questioned. Many studies have shown strong correlation between R&D inputs and patent counts and patent citations. Some researchers, however, question the strength of patent counts a proxy for innovations, since these may reflect two types of errors. First, many innovations never get patented. Instead, inventors use different measures, such as secrecy, to protect their

[1] The patent family is all the patents in different countries belonging to the same invention
[2] Patent documents refers to: published patent applications, inventors' certificates, utility models and utility certificates (International Patent Classification (Version 2012), "Guide", available online: http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf)
[3] Espacenet (2013), "International Patent Classification (IPC) "http://ep.espacenet.com/help?locale=en_EP&method=handleHelpTopic&topic=ipc

innovations. Surveys of European firms show that these omissions can be very large, and vary significantly between different branches. The extents of these omissions are however unknown, since there exists no systematic data about unpatented innovations (For a review, see Nelson 2009). Second, patent data represents inventions, and not innovations. Nelson (2009) defines inventions as technological development, and innovations as those inventions that are useful and diffused. Moser (2004) showed that only 5-20 percent of all patents are economically useful innovations. Despite the shortcomings of the patent data, it is the best data available for understanding the process of technological knowledge production.

4.2 Data collection

All data used is in this study is downloaded from PATSTAT Online, which EPO's worldwide statistical patent database. It contains bibliographic data on 90 million patents- and patents applications from around the world, including information about patent family, legal status, references to other types of technical literature and citations (EPO, 2013). The data is selected and sorted using SQL queries.

EP-patents are patents that are published by the European Patent Office (EPO). That doesn´t mean that they were invented in Europe, or that this is where they were first published. A patent can (and often will) be granted by many different patent offices, in order to earn monopoly rights for a larger geographic area. It is therefore important not to confuse *invention country* and *granting authority.*

4.3 Categorization procedure

We assume that there exist three types of patents: photovoltaic (*p*), non-photovoltaic (*n*) and all else (a), and that these categories are mutually exclusive. All existing patents are defined as belonging to one and only one of these three categories. The exact definition of these categories will vary slightly between the different variables, which will be explained more closely in section 4.4.

The technological categorization is based on the CPC-classification system. All patents are given one or more CPC-symbols by the patent office, unveiling the technological nature of the patent. We define the *p* patents as*: patents that have received at least one CPC symbol that we categorize as p*. The *n* patents are defined as patents *that have received at least one CPC symbol defined as n, and no CPC symbols defined as p*. The third type*, a, are patents that haven't received either any p or n CPC symbols (*Table 1*).*

Table 1. ***Patent categories***

Category	Description	Categorization rule
p	PV-patents	Patents that have received at least one CPC symbol that we categorize as p.
n	Non-PV patents	Patents that have received at least one CPC symbol defined as n, and no CPC symbols defined as p
a	All else	Patents that haven't received either any p or n CPC symbols.

For *p* two definitions will be used. *Definition 1* only includes patents given the CPC symbol Y02E10/50 (photovoltaic energy) or one of its subcategories. This definition is only used for the selection of the *citing patents*. The reason for using such conservative definition in this case is to simplify the search and also to guarantee that the citing patents which are the base of the analysis really are PV-patents. Theoretically all patents that are defined as PV should be given this category, but in reality some PV-patents are instead given other related CPC-symbols. For this reason we are using the more generous *definition 2* in all other cases of the categorization, for which the categorization scheme of Johnstone et al. (2009) will be use as far as possible. They use the IPC-scheme to categorize patents into different technology fields related to renewable energy. Most times the IPC symbol corresponds to an identical CPC-symbol. However, the CPC-scheme is higher in detail than the IPC, and includes more than three times as many entries, which are mainly made up by subcategories to the existing IPC-symbols. That implies that some classes related to PV energy has been added to the CPC-scheme, and thus needs to be considered for our classification. To find these we have searched for keywords relating to PV energy, using the Espacenet[4] classification search. First, the search was performed using the following keywords: "solar", "PV", "photovoltaic", "photoelectric", "solar energy", and "solar cells". After that we have manually gone through the definitions of the CPC-classes found and categorized them according to our schedule. Secondly we have looked at each section of the CPC scheme to see if some classification symbol is missing, but no additional CPC-symbols where found in this way.

Moreover, some additional modifications to the Johnston et al (2009) scheme had to be done since they use the category "solar energy", which includes both photovoltaics and other solar energy. Hence some of their categories have been excluded from our definition of PV. The

[4] Espacnet is an online patent database offered by the EPO

distinction is made based description of the categories - most times it is straightforward to see which of the two categories, PV (*p*) and non-PV (*n*), the IPC-class belongs to. The full categorization scheme is found in appendix B.

Table 2. ***Category definitions***

	PV (p)	NON-PV (n)
Definition 1	CPC=Y02E10/5*	CPC not in list 1
Definition 2	List 1(appendix B)	List 2 (appendix B)

*Note: *or any of its subcategories*

After we have found a scheme for categorizing *p*, we need to define *n*. Also for *n* two definitions will be used, one simple definition (*def. 1*) and one more refined version (*def. 2*). *Definition 1* includes all patents that are not *p*, whereas *definition 2* includes only CPC-categories that (i) are not *p*, and (ii) have some probability to be cited by PV-patents (see table 2). The reason for making this distinction is that only a small part of all patent categories is ever cited by PV-patents, and thereby contribute to the technological development of that industry. The first definition is used for categorizing the *citing patents*, $C_{i,t}$. In this case it is more reasonable to include all CPC symbols that are not *p* in the definition of *n*, since we know that all of these patents (by definition) have had an impact on the PV knowledge production.

The cited patents only represent a small fraction of all existing CPC symbols, which shows that the pool of usable knowledge for technological development of PV-energy is just a small fraction of all existing knowledge. Since the patent stocks, $P_{it,}$ are to represent the existing "usable knowledge" for PV knowledge production, the technological nature of knowledge stocks should mirror the cited patents. Therefore, to identify the CPC symbols of interest we have gone through the dataset of cited patents to see which categories are actually cited. CPC sections symbols (the first letter of the CPC-symbol) that are cited by less than 5 percent of the patents are then excluded from the definition of *n*-patents. Using this rule, about one third of all published patents are considered to belong to the *n*-category (find the complete definition in appendix B). Letting all non- "*p*" patents be "*n*" would increase the data set with about 100 million observations, which would only add noise to the set. This is still a rather generous definition and in practice, the outcome from using this definition instead of including all existing patents, only have a marginal effect on the results.

The third category, *a*, is only necessary when we are using the refined definition of *n*. It is given to patents that are 1) not *p*, and 2) not *n*. These patents are not considered in the model and are assumed to have no value for the knowledge production of PV.

4.4 Variables

The data set includes four main variables; citing patents (P_{pt}), cited PV patents (C_{pt}), cited non-PV patents (C_{nt}), and all published PV and non-PV patents (Table 3).

Table 3. ***Data description***

Variable	Granted	Priority year	Categorizaton	Autority
Citing patens (P)	8 842	1977-2009	PV (def. 1)	EPO
Cited PV patents (C_{pt})	13 238	<2009	PV (def. 2)	All
Cited non-PV patents (C_{nt})	8 716	<2009	Non-PV (def. 1)	All
All published PV patents	62 990	1900-2009	PV (def.2)	All
All published non-PV patents	16 790 277	1900-2009	Non-PV (def. 2)	All

Table 4 shows which type of information we have on each variable. For example, for *P* we have micro data on priority year, granting office, citations made and recieved and CPC-class. For the cited patents, C_{pt} and C_{nt}, we have data on publishing year, granting office and CPC-class. For the variables all published PV and non-PV patents we only have data on publishing year and CPC-class.

Table 4. ***Data description***

	Citing patents (P)	Cited PV patents (C_{pt})	Cited non-PV patents (C_{nt})	All published PV patents	All published non-PV patents
Priority year	X	-	-	-	-
Publishing year	-	X	X	X	X
Granting office	X	X	X	-	-
Citations made	X	-	-	-	-
Citations recieved	X	-	-	-	-
CPC-class	X	X	X	X	X

Citing patents

x_{pt} is an integer variable [0,∞] that sums the number of granted PV patents in year *t*, using definition 1 (Table 2). The study includes all patents granted by the EPO between 1977 and 2009, which responds to a number of 8 842 (Table 3). The time period chosen is based on data availability; there exists no EU-patent data before 1977 since the EPO was set up in that year. Excluding the patents that are not citing any patents we have 8 713 left. The data is ordered by priority year, which is the year when the application first was sent in to the patent office.

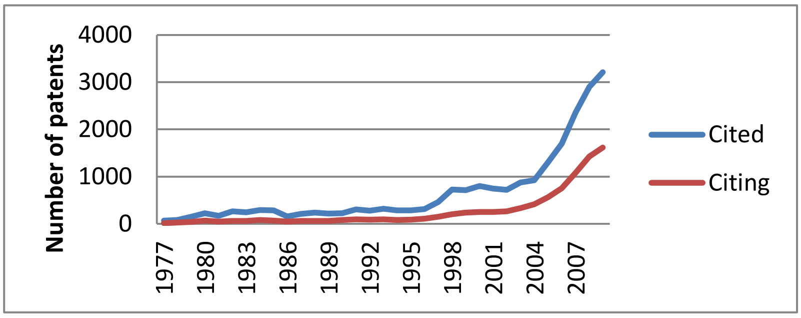


Figure 1. ***Citing and cited patents per year 1977-2009. Source: PATSTAT Online***

Figure 1 shows the flow of PV-patent grants from the European Patent Office (EPO) per year and the patents which are cited by these. In 1977, the first year of measurement, only 15 PV-patents were granted. Until the middle of the 1990's less than 100 PV-patents were granted annually. The fastest growth of new PV-patents occurred after the 1990's and until the end of the period.

Cited patents

c_{pt} is an integer variable [0,∞] that sums the number of times in year t that PV patents have been cited by P_{pt}, using definition 2 (Table 2). t refers to *priority year of the citing patent*, i.e. the year which they were cited, and responds to a number of 13 238 patents. The priority year of the cited patent can be any year prior to t, and is not considered in the study. Likewise the granting authority can be any patent office of the world.

c_{nt} is an integer variable [0,∞] that sums the number of times in year t that n-patents have been cited by P_{pt}, using definition 2 (Table 2). t refers to *priority year of the citing patent*, i.e. the year which they were cited, and responds to a number of 8 700 patents.

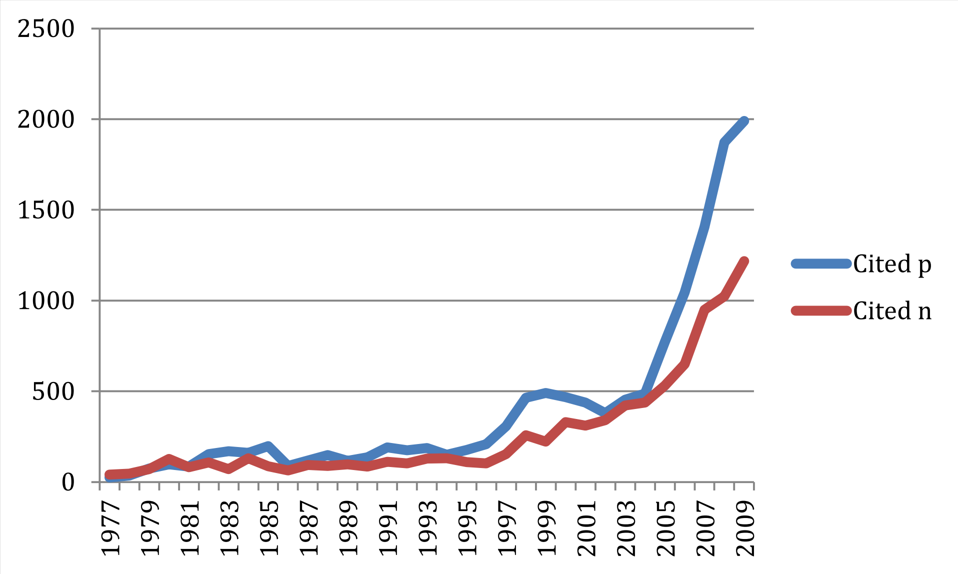


Figure 2. ***Cited PV (p) and non-PV (n) 1977-2009***

Figure 2 show how often the two categories of patents, *p* and *n*, are cited by the European PV-patents. As can be seen, PV patents are more often cited than non-PV patents, but the shares has been fairly constant over time, disregarding the very early years when the PV-share is lower (Figure 3).

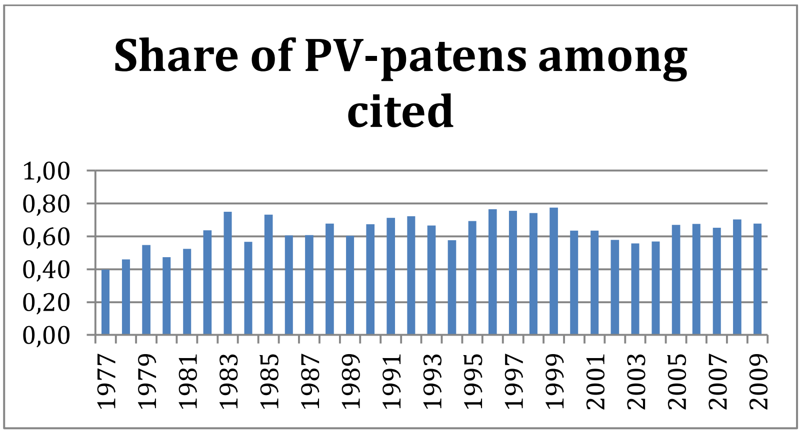


Figure 3. ***Share of PV-patents among cited***

All published patents and patent stocks

We will use Jaffe and Caballero's (1993) popular diffusion-depreciation model, according to which the patents produced in each time period adds to the patent stock through two simultaneous exponential processes, one of obsolescence and one of diffusion (Jaffe and Caballero 1993):

$$P_{i,t} = \sum_{t=0}^{T} x_{i,t}\, e^{-\beta_1 (T-t)}(1 - e^{-\beta_2 (T-t)}) \qquad (19)$$

where $x_{i,t}$is the count of the number of granted patents in category *i* in year *t*. Obsolescence is the process by which knowledge becomes 'outdated' and of no use for new innovation. Diffusion is the process by which new knowledge becomes available and useable to innovators. The magnitude of these effects are determined by the parameters $\beta_1 \geq 0$ and $\beta_2 \geq 0$, where a larger parameter value implies a faster obsolescence respectively diffusion rate. Here β_1 and β_2 are given the values of 0.1 and 0.25 respectively, in accordance with Popp (2013) . With these values the patent reaches is maximum effect about four years after it was granted.

T is the time when the patent $x_{i,t}$ makes the citation, and t is the time at which the *potentially cited patent* was published. A potentially cited patent is a patent that is considered to have some probability of being cited by patent $x_{i,t}$, i.e. patents that contain some knowledge that could be used for developing new innovations. According to this definition of the knowledge stocks, the value of the patent, i.e. its potential to be cited, only depends on its categorization and publication year.

To calculate the patent stocks we use patent data from PATSTAT Online. The set includes all *p* respective *n* patents, *using definition 2* (Table 2), published in PATSTAT Online, which is the vast majority of all existing patents in those categories, for the time period 1904-2008. In total in counts to more than 62 000 *p* patents, and almost 17 million *n* patents (Table 3). The patent stock values are then calculated using the diffusion-depreciation.

For the analysis we only need the patent stock values for the period 1977-2009, since the patent stocks shows the data available for the inventor at the time of the invention. Figure 5 show the patent stock developments for the period 1977-2009 (see also table in appendix C).

(a) (b)

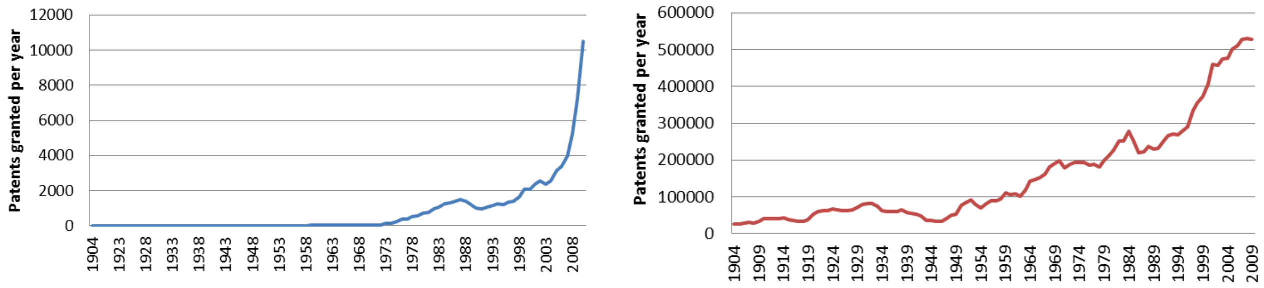


Figure 4. ***Number of PV-patents (a) and non-PV patents (B) granted per year 1904-2009.***

(a) (b)

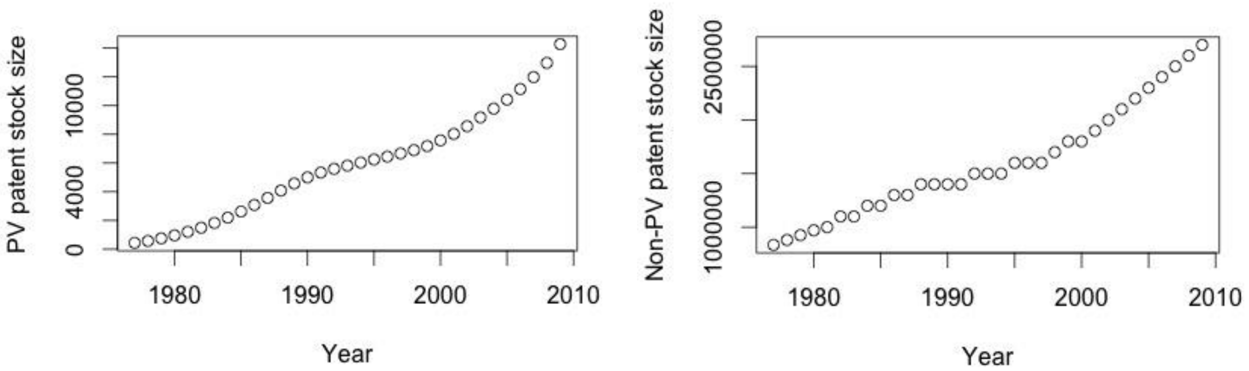


Figure 5. ***Patent stock size för PV (a) and non-PV (b)***

4.4.1 Geographical Origin

Figure 6 shows the origin of the cited patents by citing year, where origin refers to the first granting office (one patent may be granted by more than one office). US patents are the most cited at all points in time, with around 50 percent of the total. EP and Japanese (JP) pantens follow a very similar trend for the whole time period, with a maximum in the early 90's, while US patents shows a mirrored image of these.

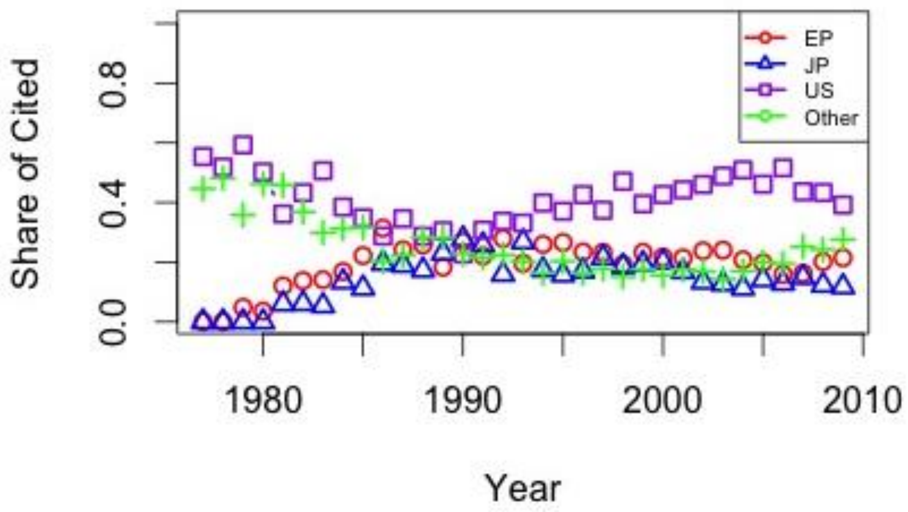


Figure 6. ***Origin of cited patents by citing year***

Figure 7 shows the share of cited patents by origin, now separated for PV and non-PV patents. There is a clear trend that non-PV patents are cited at about the same rate as PV-patents of the same origin at all points in time. The only visible exception to this is for Japanese patents after 1990, where a substantially larger share of the cited patents were PV. The difference we see here might be a result of what Orlando (2002) showed, that geographical distanse isn't a factor for intraindustry spillovers, but it is important for interindustry spillovers. The conclusion we draw by looking at the figure is that origin seems to be an important factor when it come to citations. We know from previous research inventors are much more likely to cite patents from their own country than foreign patents (see e.g. Popp 2013), so the patterns we see might be due to origin of the citing patents. We don't know the first granting office of the citing patents, but it is likely that years that have a large share of US PV and non-PV patents among the cited, is a reflection of a large share of US patent among the citing the same year.

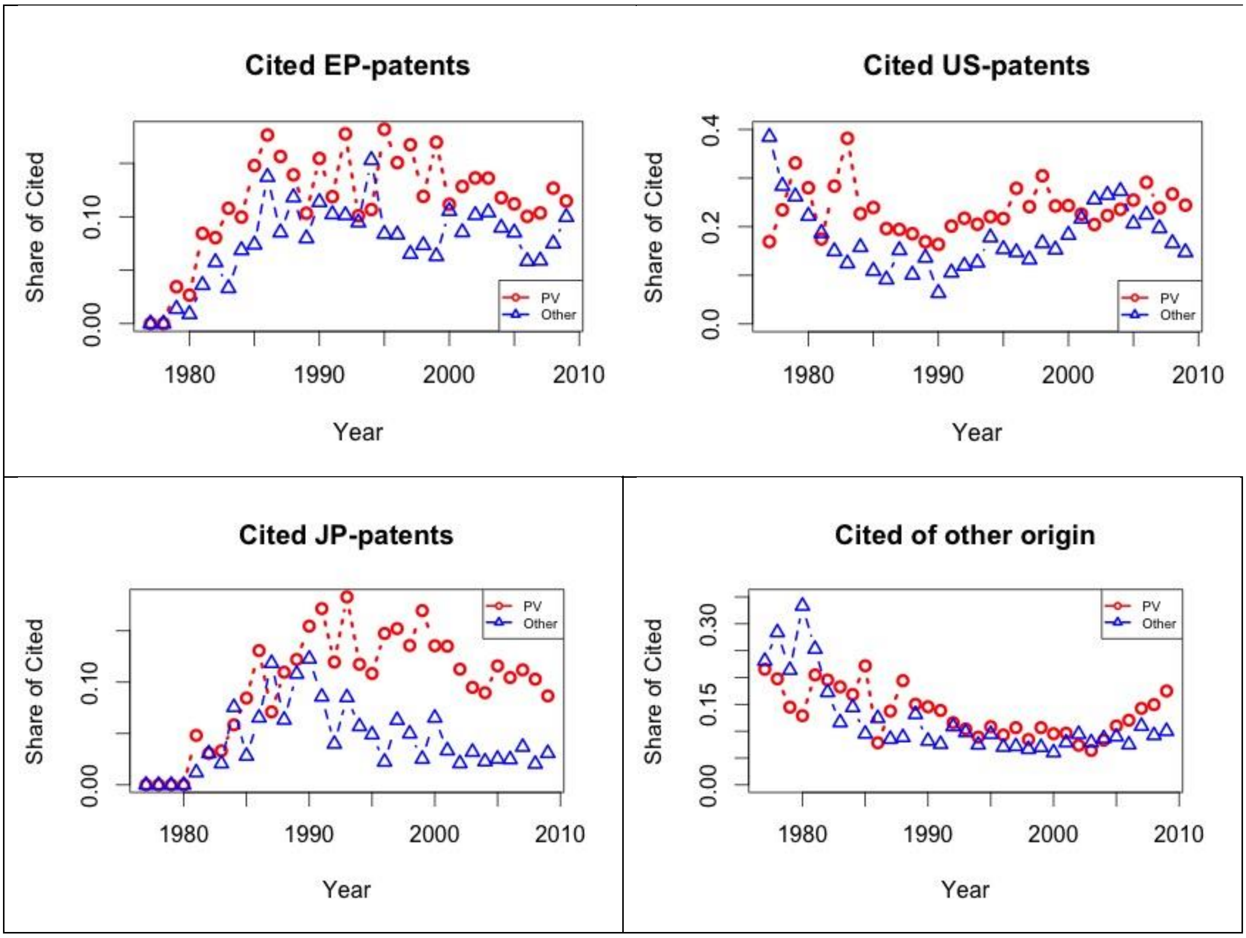


Figure 7. ***Origin of cited patents by citing year***

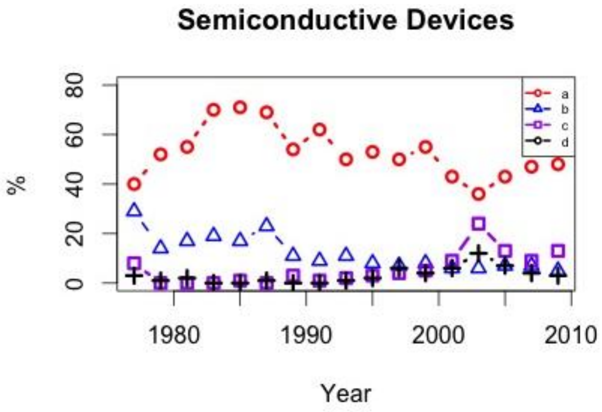
4.4.2 Technological Origin

In this section we look at the most commonly cited CPC-symbols to find out from where the spillovers come, and how this has changed over time. Figure 8 shows the percentage of the cited patents that are given a certain CPC-symbol (see appendix D for the complete data set). It is again important to note that one patent typically receives several CPC-symbols, so the percentage does not add up to 100. The most common technology groups to cite is, not surprisingly, PV technology and the closely related field of semiconductor devices. More than half of the cited patents are given CPC-symbols related to these categories. The citing rates have only changed marginally over time, even though there have been some shifts within these groups.

Solar thermal[5] and solar heat collectors[6] are two other categories of great influence, representing on average 11 percent. We can see large variations between single years, with 1 percent at the lowest, and 20 percent as the biggest, but there is no clear trend over time.

5 *Y02E 10/4: Solar thermal energy*

6 *F24J 2: Use of solar heat, e.g. solar heat collector*

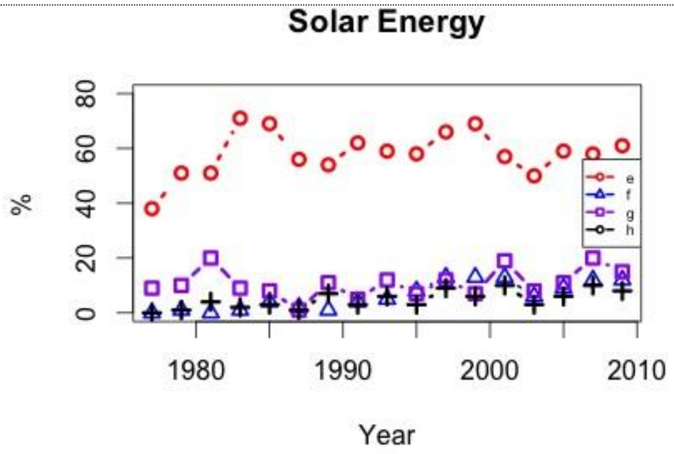


a: Semiconductor devices (H01L 31)

b: Processes or apparatus adapted for the manufacture or treatment of semiconductor or solid state devices or of parts thereof (H01L 21)

c: Solid state devices using organic materials as the active part, or using a combination of organic materials with other materials as the active part; Processes or apparatus specially adapted for the manufacture or treatment of such devices, or of parts thereof (H01L 51)

d: Capacitors (H01G)

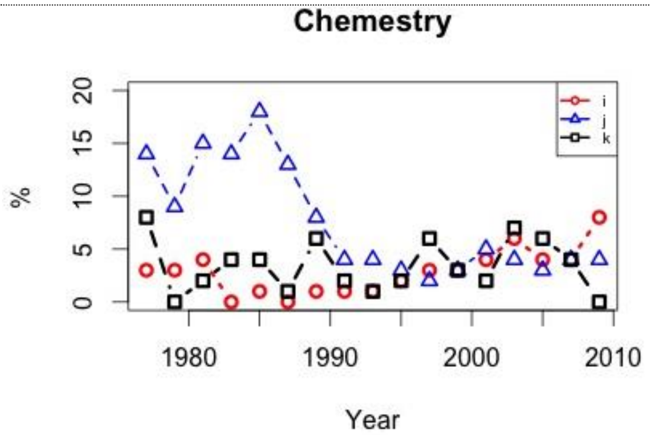


e: PV technology : (Y02E 10/5)

f: PV technology related to buildings (Y02B 10/1)

g: Solar thermal (Y02E 10/4)

h: Solar thermal related to buildings (Y02B 10/2)*



i: Dyes; paints; polishes; natural resins; adhesives; miscellaneous compositions; miscellaneous applications of materials (C09)

j: Coating metallic material, etc. (C23)

k: Glass; mineral or slag wool (C03)

Figure 8. *Cited categories over 1977-2009.*

There is an increase over time for citations of PV[7] and solar thermal[8] related to buildings. In late 1970's and early 1980's we have very few citations of patents from these categories, but from the mid 1990's their share starts growing, and in the end of the period they make up about 10 percent of the citations.

[7] *Y02B 10/1: Integration of renewable energy sources in buildings; Photovoltaic*
[8] *Y02B 10/2: Integration of renewable energy sources in buildings; Solar thermal*

Furthermore, citations relating to nanotechnology[9] and layered products[10] have been increasing over time. Electrography/magnetography[11] receives a large share of the citations before 1990 (17 percent in 1977), but after 1990 the spillovers from this technology group is only marginal.

In chemistry “dyes/paints/polishes etc”, “glass”, and “coating metallic material etc.” are the three most influential groups. The two first show no trend over time, but the latter, and most influential one, shows a clear trend of decreasing importance.

To sum up, we see that the vast majority of the citations come from the solar energy technology (PV- and non-PV) and electricity. In the early development of PV, physics related to electrography was also important, but that influence diminished drastically over time. Chemistry, especially material science, has been important from the beginning up until today, although there are differences over time between subcategories within this group.

9 *B82Y: Specific uses or applications of nano-structures; measurement or analysis of the nano-structures; manufacture of treatment of nano-structures*
10 *B23B: Layered products, i.e. products built-up of strata of flat or non-flat, e.g. cellular or honeycomb, form*
11 G03G: *Electrography; electrophotography; magnetography*

5. Model estimations

In this section we estimate two OLS models, with the purpose to test the theoretical model suggested in section 3. After that, we additionally estimate five logistic mixed effect models, to be able to make use of all the information from the disaggregated data, to see the effect of knowledge spillovers while controlling for patent origin.

5.1 Models

In this subsection we will perform OLS estimations of two simple empirical models, based on the theoretical model presented in section 3. The theoretical model (equation 18) is linearly transformed using logs:

$$ln\left(\frac{c_{pt}}{c_{nt}}\right) = \alpha' = +\varepsilon ln\left(\frac{X_{pt}}{X_{nt}}\right) + e\,, \quad \alpha' = \ln\left(\frac{2}{\sigma} - 1\right). \tag{20}$$

The dependent variable, $\frac{c_{p,t}}{c_{n,t}}$, is the citing rate, and the independent variable, $\frac{X_{p,t}}{X_{n,t}}$, is the relative patent stock size. σ is the spillover coefficient and ε is the substitutability coefficient, which are expected to be ≥ 0 and ≤ 1 in order so fulfill the assumptions of the theoretical model. e is the error term, which is assumed to be normally distributed.

Additionally, we estimate another model, in which we introduce a time dummy, *time*, which takes on the value 1 for all observations after 1999 and 0 for observations before that. The dummy is motivated by the fact that we see big differences in the patent stock growth over time. After 1999 the growth rate takes off substatially, which means we have many more observations from this time in the disaggregated data set. The model is formulated as

$$ln\left(\frac{c_{pt}}{c_{nt}}\right) = a' + \varepsilon ln\left(\frac{X_{pt}}{X_{nt}}\right) + \beta * time + e\,, \qquad \alpha' = \ln\left(\frac{2}{\sigma} - 1\right). \tag{21}$$

5.1.1 Results

Before estimating the models, two outliers representing the years 1980 and 1982 were excluded from the data set. The estimates for the full data set, including these observations can be found in appendix E. They were excluded since they had Cooks D-values that were higher than 4/n, which is a commonly used rule to identify outliers (Jacoby, 2005). The exclusion only changes the estimates marginally.

Table 5 shows the regression outputs from the estimate of the two models. For both models, ε and σ are significant at the 1% level and have predicted values that are within the given

Table 5. ***Model estimates***

Parameter	*Basic model*	*Time dummy model*
Y	$ln\left(\frac{c_{pt}}{c_{nt}}\right)$	$ln\left(\frac{c_{pt}}{c_{nt}}\right)$
ε	0.282***	.397***
	(0.0717)	(.0734)
a’	1.988***	2.742***
	(0.4184)	(.4424)
σ	0.241***	0.121***
Time		-.269***
		(.0874)
Observations	31	31
R-squared	0.348	0.512

*Note: Standard errors in parentheses . *** p<0.01, ** p<0.05, * p<0.10*

bounds >0, <1. The results confirm the hypotheses that there exists a spillover effect (sigma>0) and that there marginal effect of the relative patent stock size is diminishing (epsilon<1). The R^2 values are high for both models, 0.35 for the basic model and 0.51 for the time dummy model, which indicates a good model fit.

Figure 9 shows the relation between relative patent stock size (P_P/P_N) and citation ratio (C_P/C_N) in logarithmic scale, where the red line is the fitted values of the basic model.

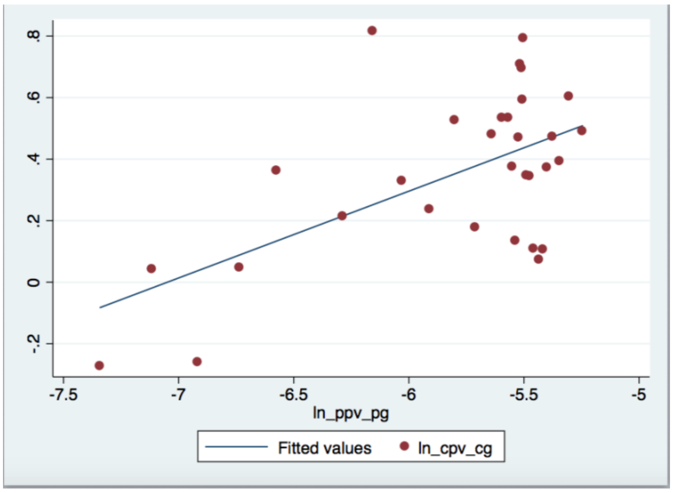


Figure 9. ***Fitted values for the basic model***

We get the value of sigma by solving the eqation $ln\left(\frac{2}{\sigma}-1\right)=\alpha'$. This gives a σ of 0.24 for the basic model and 0.12 for the time dummy-model. The value of ε is 0.28 for the basic model and 0.40 for the time-dummy model. The time dummy coefficient is calulated by exp(-0.27)=0.76.

The predictions of the two models given these estimates are illustraded in Figure 10.

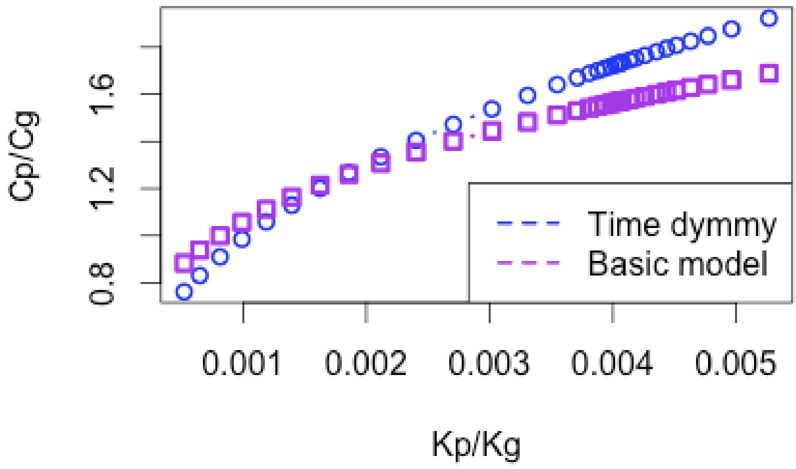


Figure 10. ***Model predictions. (Note: time=1)***

5.1.2 Diagnostics

The assumptions of OLS are:

1. The dependent variable is a *linear function* of the independent variables and the error term
2. The expected value of the error term is zero
3. The residuals are uncorrelated
4. The residuals have constant variance
5. The independent variables and uncorrelated with the error term

If the assumptions are violated, it is not possible to draw inferences from the estimates, and we cannot generalize the results. The results are however valid within the sample. For the hypothesis tests to be valid we also need the residuals to be normally distributed. Furthermore, OLS is sensitive to outliers and multicollinearity (Stata Web Books, ch. 2). In addition to the OLS assumptions we will also look at leverage.

Figure 11 shows the residuals of the two estimations. For a linear model the residuals should be randomly distributed around its zero mean. The plot suggests a slight inverted U-shape relation between the residauls and the fitted values for the basic model. If this is the case, the assumption of a linear relationship between the variables is violated, and another model specification would be a better fit. This could possibly be solved by transforming the variables, or by including other (omitted) variables. The u-shape could also be caused by autocorrelation. There are no clear signs of heteroscedasticity.

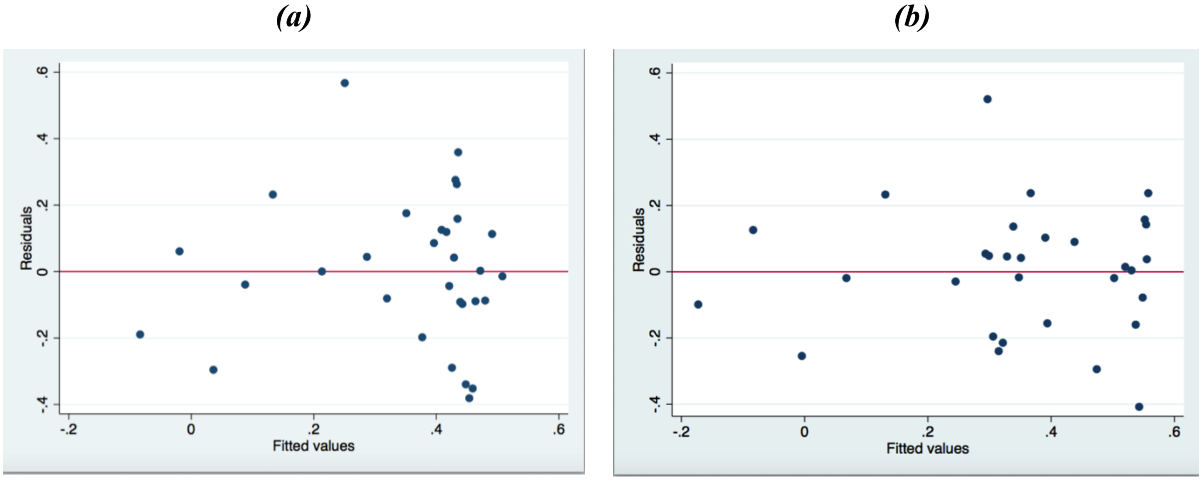


Figure 11. ***Residual plots for the basic model (a) and time dummy model (b).***

For the time dummy model, the inverted u-shape is less visible than in the basic model, and the residuals seem to be more randomly distributed around zero. There are no clear signs of heteroscedasticity or autocorrelation by judgeing from the plot, although it is difficult to tell since there are so few observations on the left side of diagram compared to the right side.

Since it is difficult to tell from the plots if we have autocorrelation or heteroscedasticity among the residuals, we additionally perform two tests, the Whites's test for homoskedasticity and the Breausch-Godfrey test of autocorrelation. In both tests we want to accept the null of zero heteroskedasticity, respectively autocorrelation. The results of the tests are shown in Table 6.

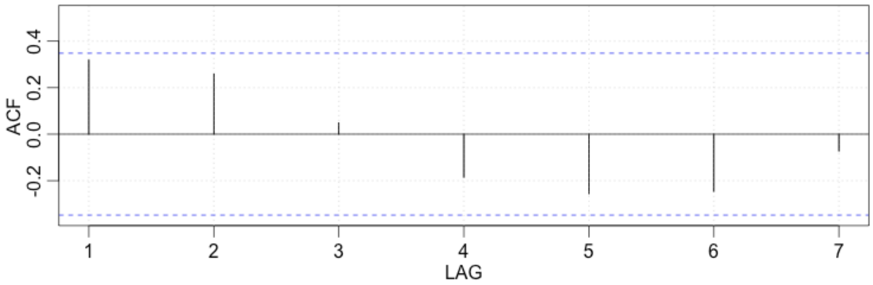
Table 6. ***Tests for heteroskedasticity and autocorrelation***

Test	**White's H_0: Homoskedasticity**		**Breusch-Godfrey H_0: No autocorrelation**	
Model	*Basic*	*Time Dummy*	*Basic*	*Time Dummy*
Chi. square	0.96	2.19	2.817	0.759
P-value	0.6731	0.7007	0.0933*	0.3837

For both models the null of homoscedasticity are accepted at the chosen significance level, so we conclude that there are no signs of heteroskedasticity among the residuals for any of the two models. Neither for the autocorrelation test, can any of the nulls be rejected at 5% significance level for either model. The p-value for the basic model is 0.09, which means that although the null can't be rejected at the 5%-level, it would have been rejected at the 10 %

level, suggesting there might still be some problems caused by autocorrelation. Further, we can see that the p-values of both tests are markedly higher for the time dummy model, suggesting that the model better fits the data with regards to these aspects, than the basic model.

(a)



(b)

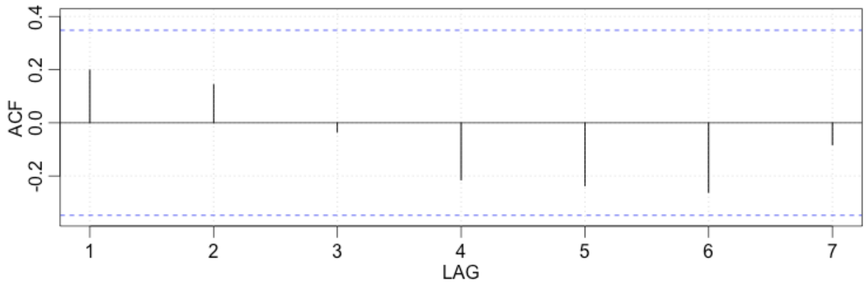
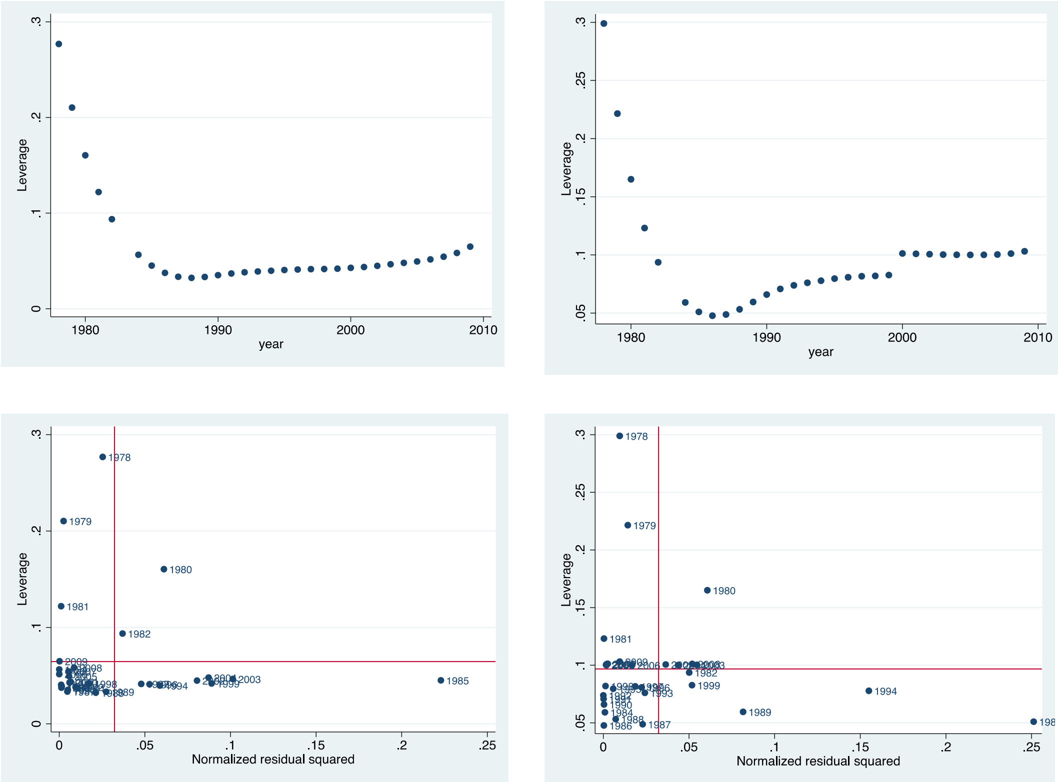


Figure 12. ***AFC for basic model (a) and time dummy model (b)***

Figure 12 shows the autocorrelation functions (ACF) for both models. On the vertical axis the autocorrelation is shown and on the horizontal axis we have the lags. The blue line represents the rejection rule, meaning that if any of the lags are larger than the given threshold value, the null of zero autocorrelation is rejected at 5 % significance level. Here we see that there is no significant autocorrelation among the residuals. Further, we again see that the autocorrelations are smaller for the time dummy model than for the basic model. Given these results we conclude that none of the models suffer from autocorrelation.

Leverage is a measure of an observations influence on the estimate. A leverage point, h_i, is considered high if $h_i > 2 \times mean(h)$. The first row of Figure 14 show the leverage of each observation by year. There is a clear pattern for both models that the early observations have much higher levarage than the later. For the basic model, three observations are considered to have high leverage, whereas for the time dummy-model there are two such observations.

(a) ***(b)***



Mean leverage: h=0.065 *Mean leverage: h:0.097*

Figure 13. ***Leverage for basic model (a) and time dummy model (b).***

As we could see in the earlier plots most of the variation in the data set are concentrated to these early years. The second row of Figure 13 shows the squared normalized residuals on the horizontal axis and the leverage on the vertical. Observations that are outside the two red lines have a combination of high leverage and big residual, which means that they are very influential. We see that for the basic model, the year 1980 and 1982 are among the observations with highest influence. The time dummy only have one observation outside the lines (with exception for a few very close to both borders), and that is for the year 1980. However, when we exclude these two observations we don't find any substatial changes to the results. All coefficiants are still significant and roughly of the same magnitude as when the high-influence observations are included (see Appendix E).

Lastly we look at the normality of the residuals (Figure 14). We can see that the residuals are spread along the diagonal line, indicating an approximately normal distribution. There are some signs of non-normality for the basic model, but the deviations are small.

(a) (b)

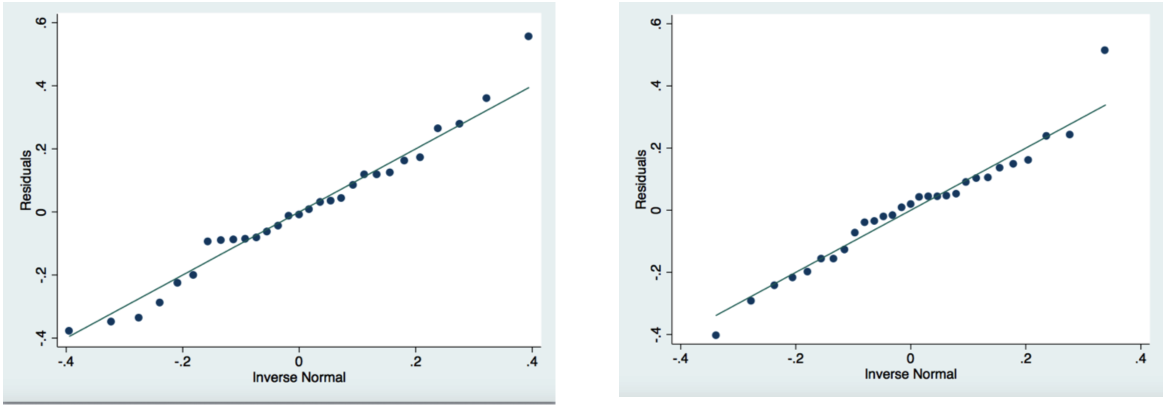


Figure 14. ***Q-Q plots of the residuals of the basic model (a) and time dummy model (b)***

5.1.3 Conclusions

The estimates are significant and give parameter estimates that are within the given bounds of the theoretical model. The R^2 are large for both models, meaning that the independent variables are able to explain a big part of the variation in the dependent variable, which is somewhat surprising given the simplicity of the models.

The basic model gives us the parameter estimates: sigma=0.2, epsilon=0.3. If we insert these values to equation 13 we get:

$$\left[\frac{c_{pt}}{c_{nt}}\right] = \frac{2-0.2}{0.2}\left(\frac{X_{pt}}{X_{nt}}\right)^{0.3} \quad . \tag{22}$$

That means that if the non-PV patent stock is 100 times the size of the PV patent stock, the probability of citing a PV-patent is twice as big as the probability of citing a non-PV patent.

The residual plot for the basic model shows signs of an non-linear relationship between the variables, which means that the first assumption of OLS is violated. Overall we cannot conclude that the assumptions are fulfilled for the basic model, meaning that we are not able to generalize the results.

For the time dummy model we don't find any obvious violations of the OLS assumptions, and the model shows a in general better fit than the basic model. The time-dummy can't be interpreted within the theoretical framework. Time itself should not be an explanatory factor for the citation rate, but probably covaries with some other factors that isn't accounted for in the models.

A possible problem for both models is that the early observations have much higher influence than the late ones. This is due to the fact that the relationship between the independent and the dependent variable is much stronger and more visible for the early observations, whereas the curve flattens out in the late years. This suggests that a linear model might be the wrong fit for this data set, and that other models should be tested. The residual diagnostics and the scatter plots of the basic model show some signs of that a quadratic model specification would better fit the data.

5.2 Logistic mixed effect models

Since the assumptions of the basic OLS model aren't fully fulfulled we continue by fitting a logistic model to the disaggregated data. There are several advantages to this approach. First, there are few distributional assumptions for logistic models (Hair et al. 2010). Second, we can now use the disaggregated data, so there will be no loss of information due to aggregation. The dependent variable in this model, PV, is now a binary variable that takes on the value 1 if the cited patent is PV and 0 otherwise. This will also drastically increase the number of observations in the dataset, from 33 (one per year) to 21 969. By increasing the sample size we can also allow for more explanatory variables, such as origin, which we have data on but coudn't make use of in the OLS-case. The disadvantage of the approach is that the results will be less straight forward to interpret with regards to theory.

The explanatory variable, *K,* will still be aggregated by year, and thus we are combining micro- and macro data. For this reason we use a multi level modeling approach, by introducing a random intercept to the model, which allows to include both individual level data and group level data (Rasbash et al. (2015). We will include dummy variables for origin, since previous studies has shown that origin is an important explanaroty factor of citation.

5.2.1 Models

The logistic random effect model has a binary outcome (p=0 or p=1) and estimates the log odds of the probability that the response variable p will take on the value of 1, given the predictors and random effects. We specify the following five models,

$$Model\ 1:\ ln\left(\frac{P(p_{ij}=1|k_{ij},u_j)}{P(p_{ij}=0|k_{ij},u_j)}\right) = \alpha + \beta_1 k_{ij} + u_j \tag{23}$$

$$Model\ 2:\ ln\left(\frac{P(p_{ij}=1|k_{ij},u_j)}{P(p_{ij}=0|k_{ij},u_j)}\right) = \alpha + \beta_1 k_{ij} + \beta_2 JP_{ij} + u_j \tag{24}$$

$$Model\ 3{:}\ \ ln\left(\frac{P(p_{ij}=1|k_{ij},u_j)}{P(p_{ij}=0|k_{ij},u_j)}\right) = \alpha + \beta_1 k_{ij} + \beta_2 JP_{ij} + \beta_3 US_{ij} + u_j \quad (25)$$

$$Model\ 4{:}\ ln\left(\frac{P(p_{ij}=1|k_{ij},u_j)}{P(p_{ij}=0|k_{ij},u_j)}\right) = \alpha + \beta_1 k_{ij} + \beta_2 JP_{ij} + \beta_3 EP_{ij} + u_j \quad (26)$$

$$Model\ 5{:}\ ln\left(\frac{P(p_{ij}=1|k_{ij},u_j)}{P(p_{ij}=0|k_{ij},u_j)}\right) = \alpha + \beta_1 k_{ij} + \beta_2 JP_{ij} + \beta_3 US_{ij} + \beta_3 EP_{ij} + u_j \quad (27)$$

where $= 1,2, \ldots J$, $i = 1,2, \ldots, n_j$, and the random effect u_j is assumed to be normally distributed with mean 0 and variance σ^2 (Li et al., 2011). j represents the grouping variable (*year*), and i the individual patent. p is a binary variable taking on the value 1 if the patent cites another PV patent, and 0 otherwise and k_{ij} the relative size of the PV knowledge stock. JP_{ij} is a dummy taking on the value 1 if the cited patent is Japanese, US_{ij} is a dummy taking on the value of 1 if the cited patent is from the US and EP_{ij} is a dummy taking on the value 1 if the cited patent is granted by EPO[12]. α is the intercept which represents the reference category for the dummies.

5.2.2 Results

Table 7 show the estimates of model 1-5 All parameter estimates, except for US, are significant on the 1 % level. In this setting the intercept doesn't have a meaningful interpretation. k_{ij} is as expected postive, meaning that when the relative knowledge stock of PV is bigger relative to non-PV, the probability of a PV-patent citing another PV-patent is bigger. Furthermore, if the patent is of Japanese origin, the probality of being cited increases. Also European origin has an positive effect, although not as strong.

Table 7. ***Estimates of model 1-5***

	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects					

[12] European patents not granted by EPO is included in the reference category and not in EP, ie. European patents granted by a national patent office

Intercept	1.81859***	1.39300***	1.40818***	1.32262***	1.30584***
	(0.44060)	(0.43547)	(0.43471)	(0.43481)	(0.43703)
ln_k	0.24947***	0.19386***	0.19410***	0.18508**	0.18388**
	(0.07541)	(0.07453)	(0.07435)	(0.07434)	(0.07447)
JP		0.76416***	0.74997***	0.78443***	0.7945***
		(0.04410)	(0.04656)	(0.04494)	(0.05079)
US			-0.02837		0.01531
			(0.02982)		(0.03583)
EP				0.08315**	0.09324**
				(0.03543)	(0.04257)
Random effect					
σ_u	0.04747	0.04566	0.04548	0.04529	0.04534
AIC	29363.0	29039.0	29040.1	29035.5	29037.3
Log likelihood	-14678.5	-14515.5	-14515.0	-14512.7	-14512.6
Obs= 21 969					
Groups=33					

*Note: Standard errors in parentheses . *** p<0.01, ** p<0.05, * p<0.10*

Table 8 shows the citing odds for model 1, 2 and 4 for three different sizes of k. Looking at model 1 we see that the odds of citing a PV patent only increases marginally (from 6.34 to 6.72) when k goes from 0.001 to 0.1. Given this model, the citing rate is almost constant, regardles of knowledge stock size. This scenario is close to the case where ε approaches 0, and the citing share can be described by a Cobb-Douglas function, where the citing rate only depends on sigma (see section 3).

Looking at the other models, where the origin dummies are included, the size of k is has a substantially bigger effect on the odds. In all cases increasing k from 0.001 to 0.1 doubles the citing odds.

Table 8. ***Citing odds for different model specifications and levels of k***

	Odds		
	K=0.001	**K=0.01**	**K=0.1**
Model 1	6.34	6.48	6.72
Model 2 (jp=1)	2.26	3.54	5.53
Model 2 (jp=0)	1.05	1.65	2.57
Model 4 (Jp=1, Ep=1)	2.48	3.81	5.83
Model 4 (Jp=1, Ep=0)	2.29	3.51	5.37
Model 4 (Jp=0, Ep=0)	1.05	1.60	2.45

The first assumption of the logit model with random intercept is that the dependent variable is independent, conditional on u. In practice this means that the probability that a random patent is citing a PV-patent should not be affected by another patents probability of doing the same. This relates to the model specification. The random intercept is assumed to account for all dependence we see in the data, and if that is not the case the model is wrongly specified (Gibbons et al., 2010).

The second assumption is that u_j is normally distributed with mean 0 and variance σ^2 (Gibbons et al., 2010). The Q-Q plots of the residuals (Figure 17, appendix F) shows that the residuals data are approxomately nomally distributed.

Furthermore there are some sample size considerations. Logistic models demands bigger samples than OLS, and is recommended that the sample size should be at least 400*, with at least 20* observations per explanatory variable. Since we have a sample of almost 22 000 observations, this is not a problem for us (Hair et al. 2010).

5.2.3 Robustness check

When the sample size is very big, as in our case, even minor effects might be detected by hypothesis tests, even though they are too small to have any relevance. To check if the found effects are still significant at smaller sample sizes we reduce the sample size substantially, from 21 969 observations to 1650, which gives us 50 observations per year. We draw a random sample without replacement from the original sample, conditioned on year. This will also give a balanced data set, in contrast with the origial sample which was highly unbalanced, with a large share of the observations concentrated to the last years of the time series.

As Table 9 shows, *ln_k* and *JP* are still significant for all models, whereas *US* and *EP* becomes insignificant for all models. The size of all significant coefficents have increased quite substatually from the previous case. The reason for this is that the early observations, that from previously are shown to be very influential, get a bigger relative weight than when we use the full data set. Also in the smaller sample Model 2 gives the best overall performance, with the lowest AIC value of the models, although the differences between the models are extremely small.

Figure Figure 18 in appendix F shows the Q-Q plots for random effects for all models. Here we find some deviations from normality, but still on an exceptable level.

To be sure we also perform Shapiro-Wilk normality tests for each of the models, with the nollhypothesis that the sample comes from a normal distribution. Table shows that non off the nulls are rejected at 10% significance level, meaning that we can’t find any significant deviations from normality (Appendix E).

Table 9. ***Regression output for model 1-5 with a smaller sample***

	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects					
Intercept	3.3088*** (0.7758)	2.9495*** (0.7665)	2.92360*** (0.76938)	2.7894*** (0.7751)	2.5782 *** (0.7910)
ln_k	0.5012*** (0.1308)	0.4538*** (0.1289)	0.45342*** (0.12895)	0.4330*** (0.1295)	0.4167 *** (0.1300)
JP		0.5395*** (0.1541)	0.56295*** (0.16373)	0.5820*** (0.1582)	0.6994*** (0.1820)
US			0.04751 (0.11190)		0.1742 (0.1341)
EP				0.1625 (0.1377)	0.2807* (0.1649)
Random effects					
Sigma2 u	0.1316	0.1230	0.1233	0.1213	0.1210
AIC	2191.7	2180.9	2182.7	2181.5	2181.8
Log likelihood	-1092.8	-1086.4	-1086.4	-1085.7	-1084.9
Obs= 1650 Groups=32					

*Note: Standard errors in parentheses . *** p<0.01, ** p<0.05, * p<0.10*

5.2.4 Conclusions

The logistic model get significant parameter estimates, showing that the citation rate increases with relative size of the PV patent stock. We could also see that origin was an important factor. The fact that a cited patent came from Japan had higher influence on the citing odds than the patent stock size. This in in line with previous research, showing that origin is a very important factor for understanding citing behaviour. In this case it is probably a reflection of the fact that most of the citing patets are of US origin. The interpretation would then be that sice you are more likely to cite any patent from your own region, only the most relevant patents from other regions are cited.

Furthermore we saw that the results were still significant after a radical decrease in the sample size, with all the effects still pointing in the same direction. The size of the effect changed however, due to the fact that the early observations recieved a higher weight in the smaller sample. Since the data is heavily unbalanced and we have observations with high levereage, the size of the effects should not be given too much emphasis. What is important is that we see that there is an effect, which further confirms that we have spillovers between the sectors, and these are depending on the size of the knowledge stock.

The results of the logistic models gives further evidence that it is necessary to take spillovers into account when modeling technological development. It also suggests that the OLS-models might be suffering from omitted variable bias, for not including origin, which seems to be a very important factor for explaining citing behaviour.

6 Discussion

We tested two OLS models and five logistic mixed effect models to see if we could find inter industry spillovers for PV-patents. The basic OLS model based on Hart's (2016) knowledge production function showed significant results and parameter estimates within the given bounds, but some of the OLS assumptions were violated which means that the results could not be generalized. This was probably due to the fact that we had too few observations, and would have needed a more complex econometric model that could handle the signs of dependence showing in the residuals. We added a time dummy to the basic OLS model, and then the necessary assumptions were fulfilled. Based on this we draw the conclusion that knowledge spillovers from other technological fields has been important for the development of photovoltaic energy. The results are, however, heavily influences by the early observations of the data set, making it difficult to say if the suggested model would give similar results had we a longer time period to study.

The relationship between patent stock size and citation rate was most evident when the PV patents stock was very small, and thus the existing number of PV patents was very low. Already in the late 1980's the trend flattened out, and while earlier PV patents still remained the most cited by European PV patents, their share remained about unchanged at between 55-75 percent. This is in agreement with the theoretical model, where the marginal productivity of knowledge is high when the knowledge stock is small. When the size of the PV knowledge stock reached a certain point the marginal productivity flattened out and instead the spillovers seemed to come in fixed proportions (only the size of sigma matters). What is surprising is how early this point was reached.

The logistic model with disaggregated data confirmed that spillovers have had a significant impact on the development of PV-technology, and that also origin has been a very important factor. In fact, Japanese origin was more important than the relative patent stock size for determining citation. This result might be caused by failure to consider dependencies in the data, due to lack of information about the origin on the citing patent. From earlier studies it is well known that inventors often have a strong preference to cite patents from their own country, and thus considering the origin of the citing and potentially cited patent would likely improve the results.

The results imply that models that neglect inter-industry knowledge spillovers might overstate the power of historical advantages of older technologies. If new technologies can make use of external knowledge they can feed on the technical development in other fields. Given that this is the case, is the assumption of extreme path dependence invalid, which gives a more positive outlook for the marginalized “clean” technologies. However, technological distance clearly was one of the most important determinants for citing. In fact all the cited patents came from a very limited group of closely related technology fields, which makes it possible to argue that given a higher level of aggregation with regards categorization, the spillovers would be close to zero.

In general, a larger share of inventions are being patented today, which might suggests that the average value of a published patent today is lower than it was in earlier times. If this trend is true for for both PV and non-PV patents, this shouldn’t have an effect on the results. However, if this “patent inflation” is more apparent in one of the two patent stocks, this could influence the results by over evaluating the value of one of the patent stocks. Since PV is a young and fast growing technology the incentives to patent inventions are high. The growing profitability of the sector might open up for patenting of less scientifically valuable inventions (see eg. de Rassenfosse & Guellec 2009). We also know that there is a tendency to file patent at EPO instead of in national patent offices. It is likely that in earlier times only very competitive inventions were filed at EPO, so that the average quality level of the EPO patents were higher. Finding patent values was beyond the scope of this study so we can’t rule out that changes in patenting behavior have had an effect on the results. If it have had an effect it is most likely that we without this would see a stronger correlation between patent stock size and citation rate. This is because the PV sector has experienced a higher growth rate than the non PV sector for the whole period and are thus more likely to suffer from inflation.

7 Conclusions

Spillovers can be seen as a pool of opportunities available to exploit by the receiving industry, which if they are utilized can boost the growth of the industry in the initial phases. After intense exploitation, the pool dries out and the growth rate slows down (Clarke et al. 2008). This is a trend that is visible for the development of the PV-technology. The early period shows a very high growth rate of PV-patents in combination with big spillovers from other technology sectors, especially from electrography, magnetography and chemistry/metallurgy. These results are similar to Nemet (2012b), that showed that energy technology has especially benefited from spillovers from fields of chemical, electronics and electrical technologies.

The results of this study indicate that we have inter-industry knowledge spillovers, and that they have had a significant impact on the knowledge production in the PV sector. That means that at least for the PV sector, Acemoglus model of technological development is not applicable. When spillovers are being ignored, the historical advantage of older technologies get overstated, meaning studies will give overly pessimistic outlooks for new technologies.

Further, the results suggest that the spillover rate is positively depending on the relative size of the sectors knowledge stock, but is diminishing. This implies that the Cobb Douglas case of a constant spillover rate is invalid for the PV technology, i.e. ε <1.

The fact that the spillovers are small, taking the relative patent stock sizes into account, and that the spillover rate stabilizes so early makes it possible to argue that both Acemoglus model and the Cobb Douglas function can be useful for studying technological development of mature technologies that have found market applications. For understanding the emergence of new technologies it is however necessary to take spillovers into account. Based on this study it seems that it is also necessary to acknowledge the degree of substitutability between different types/sources of knowledge, ε.

To improve and validate the results of this study a few measures could be taken. One is to estimate the production function on a different technology field, to find whether the results are generalizable for a broader set of technologies.

One potential problem with this study is that all patents are assumed to have the same value. In reality it is unlikely that this is the case, and by finding methods to diversify the assumed values of different patents, the results could be more precise and less noisy. There are mainly two parameters that would be highly interesting to take into account in future research of the subject. The first is the number of citations received, which is a common way to measure the importance of a patent. The second parameter is the origin of the patent. It is well known, and supported by this study, that inventors have a strong tendency to cite patents from their own country or region. For further research it would be highly interesting to locate the origin of the patents, in order to find citation patterns that can make it possible to do a more sophisticated weighting of the patents.

We saw in this study that material science was an important source of spillovers in the early stages when we only had the first generation of PV, and the main focus was to lower the material costs. Later on nanotechnology were an important influence, mirroring the third generation of solar cells. Studying each generation of PV technology separately could thus give interesting insights on the technological development. Also, as previous research has shown, the spillovers seem to be highly correlated with the technological distance between the sectors. In this study we have made a rough division between two sectors, PV and non-PV; it would be interesting to see how well the technological distance is correlated with the spillovers by making a more sophisticated classification including more technological categories.

References

Acemoglu, D. 1998, "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality", *The Quarterly Journal of Economics,* Volume. 113, no. 4, pp. 1055-1089.

Acemoglu, D., 2002. Directed Technical Change. Rev. Econ. Stud. Volume 69, 781–809.

Acemoglu, D., Aghion, P., Bursztyn, L., Hemous, D., 2009. The Environment and Directed Technical Change (Working Paper No. 15451). National Bureau of Economic Research.

Aghion, P., & Howitt, P. W. (2008). *The economics of growth*. MIT press.

Ahmad, S., 1966. On the Theory of Induced Invention. Econ. J. 76, 344.

Arthur, B.W., 2007. The structure of invention. Res. Policy Volume 36, 274–287.

Binswanger, H., 1974. A Microeconomic Approach to Induced Innovation. Econ. J. Volume 84, 940–958.

Clarke et al., L., 2008. On the sources of technological change: What do the models assume? Energy Econ. Volume 30, Pages 409–424.

EPO (2013), "Espacenet - Free access to 80 million patent documents worldwide", From "http://documents.epo.org/projects/babylon/eponet.nsf/0/4E8744EB66E8F944C12577D600598EEF/$File/espacenet_brochure_en.pdf" (Accessed 2014.05.01)

Gibbons, R. D., Hedeker, D., & DuToit, S. (2010). Advances in Analysis of Longitudinal Data. *Annual Review of Clinical Psychology*, *6*, 79–107.

Griliches, Z., 1992. The Search for R&D Spillovers. Scand. J. Econ. Volume 94, 29–47.

Hair, Joseph F. *, 'Multivariate Data Analysis: A Global Perspective'*, Anonymous Translator(7th edn, Upper Saddle River, N.J, Pearson Education, 2010).

Hart, R., 2012. Directed technological change: It's all about knowledge (Working Paper Series No. 2012:02). Department Economics, Swedish University of Agricultural Sciences.

Hart, R., 2016. The knowledge production function (Working Paper RH16:1. Department Economics, Swedish University of Agricultural Sciences.

Heslin Rothenberg Farley & Mesiti 2009, Clean Energy patent growth index 1st Quarter 2009. Cleantech Group . From "http://www.hrfmlaw.com/wpcontent/uploads/2015/05/Clean_Energy_Patent_Growth_Index_1st_Quarter_2009_article_160816.pdf" Accessed 17.01.2016)

Heslin Rothenberg Farley & Mesiti 2010, Clean Energy patent growth index 2nd Quarter 2010. Cleantech Group . From "http://cepgi.typepad.com/files/cepgi-2d-quarter-2010.pdf" Accessed 17.01.2016)

Hicks, J.R., 1932. The Theory of Wages. Macmillan, London.

IEA, 2010. Technology Roadmap - Solar Photovoltaic Energy.

IEA, 2013. PVPS Report: A Snapshot of Global PV (No. IEA-PVPS T1-22:2013).

Jacoby, W. G., 2005. Regression III:Advanced Methods. From: "http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week3/11.Outliers.pdf" (Accessed 11.01.2016)
Kennedy, C., 1964. Induced Bias in Innovation and the Theory of Distribution. Econ. J. Volume 74, 541–547.

Li, B. et al. (2011). Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. *BMC Medical Research Methodology*, *11*, 77.
Liu et al., J.S., 2011. Photovoltaic technology development: A perspective from patent growth analysis. Sol. Energy Mater. Sol. Cells Volume 95, 3130–3136.

Nelson, A.J., 2009. Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. Res. Policy Volume 38, 994–1005.

Nemet, G.F., 2012a. Do important inventions benefit from knowledge originating in other technological domains? Res. Policy Volume 41, 190–200.

Nemet, G.F., 2012b. Inter-technology knowledge spillovers for energy technologies. Energy Econ. Volume 34, 1259–1270.

Nordhaus, W.D., 1973. Some Skeptical Thoughts on the Theory of Induced Innovation. Q. J. Econ. Volume 87, 208–219.

Perlin, J. (2013). *Let it shine: The 6,000-year story of solar energy*.
Pinheiro, J.C. & Bates, D.M. 2000, *Mixed-effects models in S and S-PLUS,* Springer, New York.
Popp et al., D., 2009. Energy, the Environment, and Technological Change, in: Working Paper 14832. National Bureau of Economic Research.

Rasbash, J., Steele, F., Browne, W. and Goldstein, H. (2015) A User's Guide to MLwiN ,Version 2.33, Centre for Multilevel Modelling, University of Bristol
Ruttan, V.W., 2001. Technology, growth, and development: an induced innovation perspective. Oxford University Press, New York.

Stata Web Book, UCLA: Statistical Consulting Group.
From "http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm" (Accessed 17.01.2016)
Trajtenberg, M., Henderson, R., Jaffe, A., 1992. Ivory Tower Versus Corporate Lab: An Empirical Study of Basic Research and Appropriability (Working Paper No. 4146). National Bureau of Economic Research.

Tyagi et al., V.V., 2013. Progress in solar PV technology: Research and achievement. Renew. Sustain. Energy Rev. Volume 20, 443–461.

Wu et al., C.-Y., 2012. Knowledge flows in the solar photovoltaic industry: Insights from patenting by Taiwan, Korea, and China. Res. Policy Volume 41, 534–540.

Appendices

Appendix A: PV-energy

Photovoltaics (PV) is the direct conversion of radiation into electricity. PV-systems contain solar cells that convert light into electricity. Each cell contains layers of semi-conducive materials. When light falls on the cell an electric field across the layers is created, causing the electricity to flow. It is the intensity of the light that determines the amount of electricity that is generated in each cell (Tyagi et al. 2013, 2).

The market for PV energy is still small, only about 0.5 percent of the total world electricity generation is provided by photovoltaics. Global production data for solar energy estimates a production between 18GW and 28GW in 2012 (IEA 2013). Yet, the branch is fast growing, and the annual growth rate of PV installations has been between 40 and 90 percent since year 2000. IEA (2010) estimates PV energy to deliver 5 percent of the world energy demand by 2030, and 11 percent by 2050. Europe is by far the largest market for PV energy, with Germany and Italy as the leading countries, each having more than 5 percent of their total electricity demand covered by PV energy. Asia is currently the fastest growing region for PV, and China is the leading country in solar cell production (IEA 2013).

The technology has experienced fast cost reductions, and the unit cost dropped to one third between 2008 and 2013. In order to increase market shares, the efficiency of the solar cell is an important parameter. The United States is by far the most important country for technological development in solar energy. Between 2002 and 2011 more than 50 percent of the worlds solar patents were assigned to American inventors (including both solar thermal and PV). Second to US was Japan with 22 percent and Germany with 6 percent (Heslin, Rothenberg, Farley & Mesiti P.C. 2010).

The European PV knowledge production has had two periods of fast growth. The first is in the late 1970's until the early 1980's. Even though the first silicon solar cell was created already in 1954, and that the technology had been an important part of both the American and Soviet space program since the 1960's, this was when it first was considered as an interesting option for a bigger market and when PV technology found its way into consumer products (Perlin 2013, Let it shine).

PV technology is generally classified into three generations. The first generation is based on monocrystalline silicones wafers. These are still dominating the market due to the high efficiency of the material, which can reach up to 20 percent (Tyagi et al. 2013). There are still technological improvements in this area, but widespread adaptations of the technology have been hindered by the price and availability of raw silicon. (Heslin Rothenberg Farley & Mesiti, 2009) As the technology have improved, the costs have been more and more dominated by material costs. Polycrystallines are also used as a cheaper alternative, but they are also much lower in efficiency. (Green, 2002) The second generation made up by thin film materials, for example CIS/CIGS, CdS/CdTe and Ampurphus silicone. These require less semi conductive material than the first generations' solar cells but are also less efficient (Heslin Rothenberg Farley & Mesiti, 2009). Another disadvantage is their adverse environmental impact. Despite these flaws the technology is increasing its marketshares (Tyagi et al. 2013). The third generation is still in an early stage of development, and includes more experimental technologies that are less adapted to the market. It includes dye-sensitized solar cells, quantum dots, nano-modified materials (Heslin Rothenberg Farley & Mesiti, 2009). They are generally advantageous regarding costs and environmental impact, but are suffering from the low efficiency (Tyagi et al. 2013).

Appendix B: CPC classification scheme

Table 10. ***CPC classification scheme***

List 1: definition 2 of p (PV)	*List 2: Definition 2 of n (non-PV)*
"Y02E10/5*"	Not PV (list 1)
"Y02T50/69"	Not : A*, D*, E*, G*
"Y02T70/5245"	
"Y02B10/1*"	
"H02N6*"	
"H01L27/142*"	
"H02S*"	
"H02J3/383"	
"H02J3/385"	
"H01L31/02008"	
"H01L31/02021"	
"H01L31/0203"	
"H01L31/02167"	
"H01L31/02168”	
"H01L31/022425"	
"H01L31/022441"	
"H01L31/02245"	
"H01L31/022458"	
"H01L31/0424"	
"H01L31/0485"	
"H01L31/0504"	
"H01L31/0516"	
"H01L31/0522"	
"H01L31/0527"	
"H01L31/068"	
"H01L31/0682"	
"H01L31/0684"	
"H01L31/0687"	
"H01L31/06875"	
"H01L31/0725"	
"H01L31/073"	
"H01L31/0735"	
"H01L31/074"	
"H01L31/0745"	
"H01L31/0747"	
"H01L31/0749"	
"H01L31/076"	
"H01L31/188"	
"H01L31/1884"	
"H01L31/1888"	

*Note: *any of its subcategories*

Appendix C: Patent stock size

Table 11. ***Patent stock size per year***

Year	PV patent stock	Non-PV patent stock
1977	436	834515
1978	571	880693
1979	752	925367
1980	962	925367
1981	1209	1018094
1982	1492	1069946
1983	1824	1123439
1984	2196	1181970
1985	2619	1236968
1986	3080	1281345
1987	3567	1318237
1988	4082	1352324
1989	4574	1382170
1990	5000	1408926
1991	5334	1436998
1992	5593	1467678
1993	5818	1500178
1994	6028	1532580
1995	6244	1566468
1996	6445	1602480
1997	6666	1647110
1998	6896	1700501
1999	7181	1761253
2000	7575	1831915
2001	8027	1917491
2002	8561	2009691
2003	9178	2107051
2004	9786	2205278
2005	10415	2306215
2006	11152	2408064
2007	11984	2511237
2008	12971	2612902
2009	14281	2710138

Source: EPO

Appendix D: Citings per CPC symbol

Table 12. ***Number of citings per CPC symbol***

*Note: *any subcategory. Source: EPO*

CPC-symbol																	
YEAR	Y02E 10/5*	H01L31	Y02B 10/1*	Y02E 10/4*	Y02B 10/2	B82Y	H01G	F24J 2	G03G	H01L 21	H01L 51	E04D	C	C09	C23	C09	C03
1977	*25*	26	0	6	0	0	2	6	11	19	5	0	19	2	9	3	5
1979	*74*	76	1	14	2	3	1	15	2	21	0	1	27	4	13	3	0
1981	*85*	91	0	33	6	0	3	33	7	28	0	0	49	7	25	1	3
1983	*170*	168	2	22	4	0	0	22	7	45	1	0	47	0	33	1	9
1985	*197*	203	10	22	9	1	1	24	23	47	2	1	64	3	51	1	12
1987	*118*	146	4	2	2	6	3	3	5	49	1	0	47	1	27	0	2
1989	*116*	114	3	24	15	5	0	24	1	24	6	5	40	3	18	4	12
1991	*189*	188	11	15	9	4	0	15	5	28	2	4	37	3	11	1	5
1993	*188*	160	15	38	18	13	2	40	7	34	5	7	31	3	13	1	4
1995	*166*	152	22	19	8	11	5	21	1	24	8	13	39	6	10	14	7
1997	*304*	228	60	55	42	6	26	64	6	34	19	22	59	15	11	12	26
1999	*490*	395	94	53	43	16	27	54	7	54	35	27	107	22	23	13	19
2001	*428*	325	94	139	78	22	47	148	2	45	67	40	128	33	40	45	16
2003	*436*	315	55	67	30	52	105	72	14	53	209	8	179	55	38	27	64
2005	*765*	562	118	147	80	54	93	152	6	93	165	42	246	55	40	43	77
2007	*1372*	1102	274	463	228	75	104	481	8	158	206	88	414	99	105	178	92
2009	*1949*	1554	375	497	250	119	98	538	2	176	425	65	750	263	119	347	122

Appendix E: Additional regression outputs

Table 13. ***Regression with all observations***

Model	*Basic all obs*	*Basic without 1980 and 1982*	*Time dummy model without 1980 and 1982*
Y	$ln\left(\frac{c_{pt}}{c_{nt}}\right)$	$ln\left(\frac{c_{pt}}{c_{nt}}\right)$	$ln\left(\frac{c_{pt}}{c_{nt}}\right)$
ε	0.295*** (0.0712)	0.262*** (0.0789)	0.379*** (0.0791)
Constant	2.076*** (0.409)	1.873*** (0.445)	2.694*** (0.857)
Observations	33	29	29
R-squared	0.357	0.289	0.476

*Note: Standard errors in parentheses . *** p<0.01, ** p<0.05, * p<0.10*

Table 14. ***Cooks d***

Year	Ln(Patent stock size)	Ln(Citation ratio)	Cook's D
1983	.87	-6.42	0.219
1977	-.47	-7.56	0.34

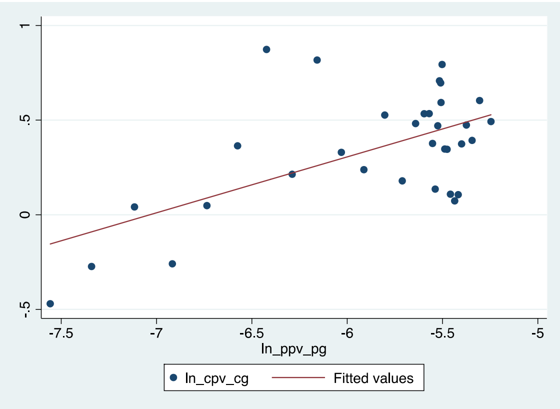


Figure 15. ***Regression plot with all observations.***

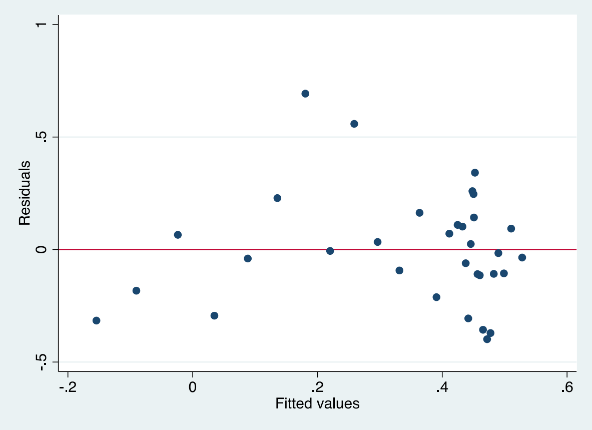


Figure 16. ***Residual plot basic model including outliers***

Appendix F: Q-Q plots

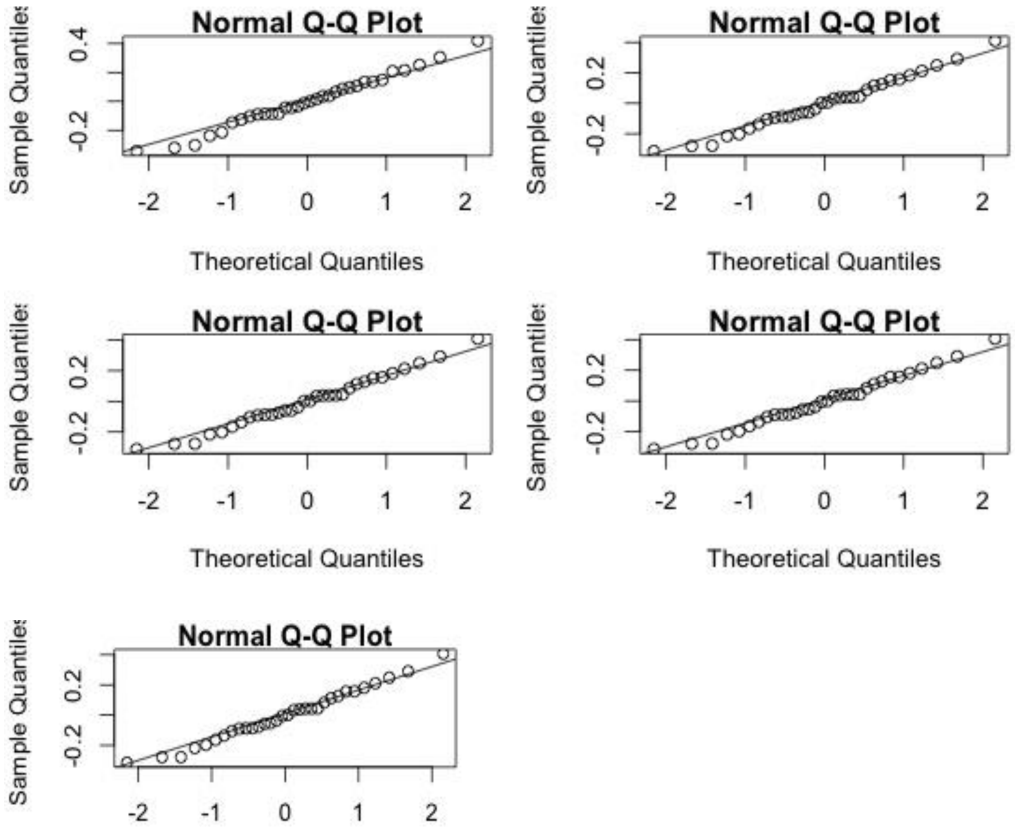


Figure 17. ***QQ-plot for ranodm intercept model 1- 5 full sample size***

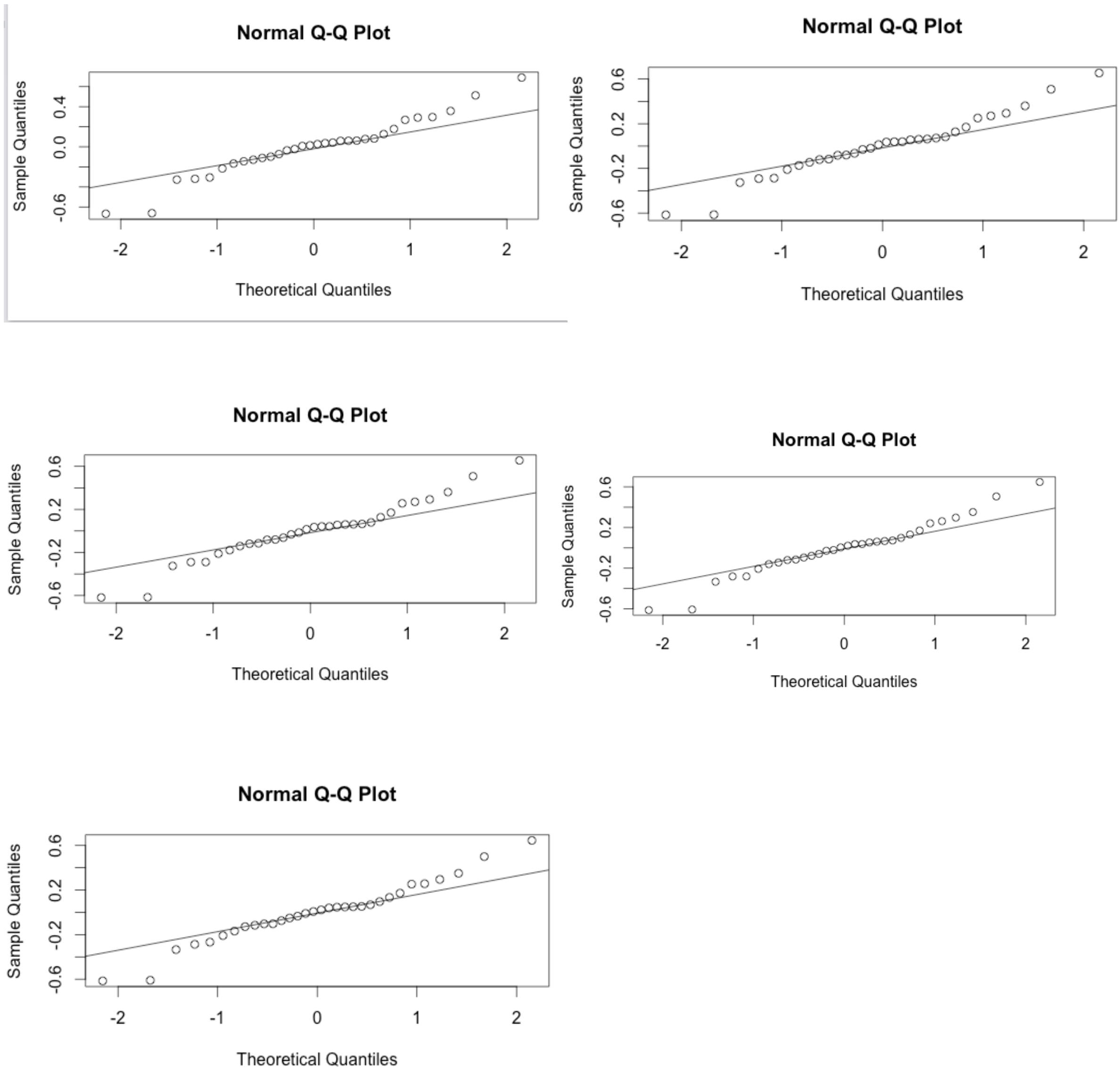


Figure 18. ***q-q plot model 1-5. Smaller sample***

Table 15. ***Shapiro-Wilk normality test for smaller sample size***

Model	W	P-value
1	0.9611	0.2942
2	0.96749	0.4331
3	0.96624	0.4025
4	0.96968	0.4906
5	0.96713	0.4241