Diplomarbeit

Knowledge Spillovers in Austria

A Patent Citation Analysis

ausgeführt zum Zwecke der Erlangung des
akademischen Grades eines Diplom-Ingenieurs

unter der Leitung von

Ao. Univ.-Prof. Mag. DDr. Thomas Dangl

und

Univ.-Ass. Mag. Dr. Michael Weichselbaumer

Institut für Managementwissenschaften (E330)
Arbeitsbereich Wettbewerb und Innovationsmanagement

eingereicht an der

Technischen Universität Wien
Fakultät für Maschinenwesen und Betriebswissenschaften

von

Georg Seliger
0426372

Schadinagasse 8/18
1170 Wien

Wien, am 12.05.2011

Unterschrift

Acknowledgements

First of all, I want to thank Prof. Dangl and Dr. Weichselbaumer for the guidance and professional support throughout the writing of this thesis. Their assistance at solving numerous questions and the provision of new points of view on the topic have been very valuable to me.

Furthermore, I thank my parents, Ernestine and Otto, as well as my brothers, Thomas and Dominik, for their support in countless ways during my studies and the writing of this thesis.

Finally, I thank my girlfriend, Carina, for her everlasting support and motivation under all circumstances.

Abstract

As the foundations of the economy change from tangible to intangible resources, the side effects of knowledge generation become more important. Unlike regular economic goods, knowledge can be distributed and utilized easily once it has been created. While firms generally try to avoid involuntary leakage of knowledge, it is not yet clear which factors influence the occurrence of knowledge spillovers between different economic actors. In this thesis, patent citations are used as a proxy for knowledge spillovers in order to examine how spatial, temporal and industry-related factors influence the probability of knowledge spillovers. Binary response models and three different datasets consisting of Austrian firms' patent citations are used to determine the impact of aforesaid factors on the probability of knowledge spillovers and how this impact changes over time and for different directions of spillovers. The results suggest that time-related variables are only of minor importance, while the affiliation of sender and receiver of a spillover to the same firm or to the same industry is crucial. There is, however, no sign of increased spillovers due to spatial proximity.

Keywords: Austria, knowledge spillovers, patent citations.

Kurzfassung

Durch die steigende Bedeutung von immateriellen Gütern, wie zum Beispiel Wissen, in der heutigen Wirtschaft rücken auch die Nebenwirkungen, die mit neu erschaffenem Wissen einher gehen, in den Vordergrund. Ist erst einmal neues Wissen entdeckt, kann es – im Unterschied zu klassischen wirtschaftlichen Gütern – relativ einfach verteilt und verwendet werden. Während viele Firmen versuchen, neue Erkenntnisse vor der Verwendung durch die Konkurrenz zu schützen, ist immer noch unklar, wodurch das Auftreten von „Wissensflüssen" (*knowledge spillovers*) zwischen einzelnen ökonomischen Akteuren beeinflusst wird. In dieser Diplomarbeit werden Patentzitate als Annäherung für Wissensflüsse verwendet und es werden die Einflüsse von raumbezogenen, zeitbezogenen und industriebezogenen Faktoren untersucht. Mit Hilfe von binomialen Regressionsmodellen und drei verschiedenen Datensätzen, die aus den Patentzitaten österreichischer Firmen bestehen, werden die Auswirkungen der oben genannten Faktoren auf die Wahrscheinlichkeit eines Wissensflusses, sowie deren Veränderung über die Zeit und für verschiedene Flussrichtungen untersucht. Die Ergebnisse zeigen, dass die zeitbezogenen Variablen wenig Einfluss haben, während die Zugehörigkeit des Senders und des Empfängers eines Wissensflusses zur selben Firma, bzw. zur selben Industrie große positive Auswirkungen auf die Wahrscheinlichkeit hat. Es gibt jedoch keine Anzeichen für erhöhte Wissensflüsse in Folge räumlicher Nähe.

Schlüsselwörter: Österreich, Wissensflüsse, Patentzitate.

Table of contents

List of figures

List of tables

Appendix:

Abbreviations and notation

Abbreviations:

AIC:	Akaike's Information Criterion.
AJ:	Spillover model by d'Aspremont and Jacquemin (1988; 1990).
AT:	Austria.
BIC:	Bayesian Information Criterion.
CA:	Canada.
CH:	Switzerland.
DE:	Germany.
EPO:	European Patent Office.
IN_PER1:	Dataset of actual citation pairs of the first period for incoming knowledge flows.
IN_PER2:	Dataset of actual citation pairs of the second period for incoming knowledge flows.
IPC:	International Patent Classification.
IPR:	Intellectual Property Rights.
ISIC:	International Standard Industrial Classification.
JP:	Japan.
KMZ:	Spillover model by Kamien, Muller and Zang (1992).
MAR:	Marshall-Arrow-Romer externalities.
NUTS:	Nomenclature of Statistical Territorial Units (fr. *Nomenclature des unités territoriales statistiques*).
OUT_PER1:	Dataset of actual citation pairs of the first period for outgoing knowledge flows.
PATSTAT:	Patent database issued by the European Patent Office.
PCT:	Patent Cooperation Treaty.
R&D:	Research and Development.
REG_IN_PER1:	Regression dataset of the first period for incoming knowledge flows.

REG_IN_PER2: Regression dataset of the second period for incoming knowledge flows.

REG_OUT_PER1: Regression dataset of the first period for outgoing knowledge flows.

SME: Small and Medium-sized Enterprises.

UK: United Kingdom.

US, USA: United States of America.

USPTO: United States Patent and Trademark Office.

WIPO: World Intellectual Property Organization.

Notation:

a_j: Multiplier for the random variable X_j.

C_i: Number of citations received by patent i.

$\exp(z)$: The exponential of z.

$G(\cdot)$: Linking function.

i: Index variable, indicating e.g. patents, citations and independent regression variables.

j: Index variable, indicating e.g. patents, different groups of patents and iterations of subregressions.

J: Total number of e.g. different groups of patents and iterations.

k: Total number of e.g. patents and independent regression variables.

K: Level of economically valuable technological knowledge.

$\dot{K}$: Change of the level of economically valuable technological knowledge per time unit.

l: Total number of cited patents.

$L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})$: Likelihood of $\boldsymbol{\beta}$, dependent on the values of $\boldsymbol{y}$ and $\boldsymbol{X}$.

$\hat{L}(M)$: Estimated likelihood of model M.

$\ln(z)$: Natural logarithm of z ($z > 0$).

M_α: Regression model with only the intercept included.

M_β: Regression model with independent variables and intercept included.

n:	Index variable, indicating the observation number; number of firms in the KMZ model.
N_j:	Total number of patents in a given group j.
N_t:	Total number of patents issued in a given product class in year t.
P:	Patents.
$P_1, \dots, P_i, \dots, P_k$:	Citing patents.
$\Pr(y = 1 \mid x_1, \dots, x_k)$:	Probability of the event $y = 1$ dependent on the values of $x_1, \dots, x_k$.
$\widehat{\Pr}(y = 1)$:	Estimated probability of the event $y = 1$.
$Q_1, \dots, Q_j, \dots, Q_l$:	Cited patents.
R:	Resources invested into inventive activities.
R^2:	The coefficient of determination.
R^2_{McF}:	The coefficient of determination according to McFadden.
$SE_{i,j}$:	Standard error of the i-th independent regression variable of the j-th subregression.
$SE_{i,total}$:	Standard error of the i-th independent regression variable of the overall regression.
t:	Year variable.
u:	Unobservable stochastic process.
v, v_i:	Unobservable stochastic processes and disturbances.
$\mathrm{Var}(X_j)$:	The variance of the random variable X_j ($j = 1, \dots, J$).
VI:	Patent value index.
w_j:	Weight assigned to the j-th group of patents.
WPC_t:	Index of weighted patent counts for a given product class in year t.
$\boldsymbol{X}$:	Matrix of independent regression variables of the entire sample.
$x_1, \dots, x_i, \dots, x_k$:	Independent regression variables.
$\boldsymbol{x_n}$:	Row vector of independent regression variables of the n-th observation.
X_i:	Observable variables influencing Z_i.
y:	Dependent regression variable ($y \in \{0,1\}$ in a binary response model).

$\boldsymbol{y}$:	Vector of dependent regression variables of the entire sample.
z:	A real number.
Z_i:	Indicators of the benefits of invention and innovation.
α:	Nonlinearity factor.
$\boldsymbol{\beta}$:	Vector of coefficients of the independent regression variables.
$\beta_0, \beta_1, \dots, \beta_i, \dots, \beta_k$:	Coefficients of the independent regression variables.
$\hat{\beta}_{i,j}$:	Estimated coefficient of the i-th independent regression variable of the j-th subregression.
$\hat{\beta}_{i,total}$:	Estimated coefficient of the i-th independent regression variable of the overall regression.
γ:	Parameter denoting the degree of substitutability of products in an oligopoly ($\gamma \in [0,1]$).
δ:	A change in the value of an independent variable x_i.
$\Lambda(\cdot)$:	Standard logistic cumulative distribution function.
σ:	Parameter denoting the degree of spillovers between firms in an oligopoly ($\sigma \in [0,1]$).
σ_{AJ}:	Spillover parameter in the d'Aspremont-Jacquemin model.
$\sigma_{AJ,max}$:	Maximum value of the spillover parameter in the d'Aspremont-Jacquemin model.
σ_{KMZ}:	Spillover parameter in the Kamien-Muller-Zang model.
$\sigma_{KMZ,max}$:	Maximum value of the spillover parameter in the Kamien-Muller-Zang model.
$\Phi(\cdot)$:	Standard normal cumulative distribution function.

1 Introduction

The past decades have revealed a serious change in the perception of the determinants of economic activity. The focus of economic analysis has shifted to the examination of innovation and technological progress. Starting with Schumpeter, who coined the term "creative destruction" in order to describe the innovative processes in an economy, a large body of research has been accumulated on this topic. Especially during the last twenty years, various innovation models, such as the disruptive innovation model (Christensen, 1997), the dominant design theory (Utterback, 1994) and the open innovation model (Chesbrough, 2003), have been developed in order to explain the processes of innovation. Hitherto, most models are based mainly on case studies of the past and lack the ability to make predictions as well as the applicability on a general scale. Nevertheless, all models agree upon the foundation of innovation: the creation of knowledge.

1.1 Problem formulation

The creation of knowledge in firms can happen in various ways. The most straightforward one is the establishment of a lab for Research and Development (R&D) which may conduct basic as well as applied research. Another way, especially for process innovations, is to consider propositions that were submitted to the employee suggestion system in order to make the internal work flows more efficient. All of these mechanisms result in valuable knowledge that can be exploited by the firm. However, these improvements come at a certain cost: Setting up a R&D lab, recruiting scientists and introducing an employee suggestion system requires many financial as well as organizational resources. Yet, once the knowledge has been created, it is often the case that other firms, which were not involved in the creation, gain access to the knowledge at practically no costs. In economic literature, this phenomenon is described as knowledge spillovers.

Knowledge spillovers refer to all kinds of externalities that make knowledge created by one firm available to its competitors as well as to other firms. Since knowledge spillovers provide the recipients with valuable information on new technologies at very little or no costs, they form a serious influence on the economic performance of the involved companies. Although knowledge spillovers are regarded as positive externalities (as opposed to negative ones, such as environmental pollution and global warming), many companies perceive knowledge spillovers as a negative impact on their business, especially when it is their knowledge that is exploited by others. The firm constituting the source of the spillover is usually not pleased when other companies utilize its knowledge without paying for it. Although there are several ways to prevent others from using one's invention (e.g. intellectual property rights, such as patents and utility models), there is no perfect protection possible. This is also due to the informal nature of the channels through which knowledge spillovers happen, e.g. scientific conferences and networks among scientists.

For the same reason, it is very difficult to measure knowledge spillovers appropriately for use in quantitative research. Knowledge spillovers tend to be disembodied, leaving almost no trace of their existence which is usable for econometric analyses. Nevertheless, due to their economic importance, there is a major interest in finding empirical evidence of knowledge spillovers. Up until now, no well-established method of accurately measuring this phenomenon has been developed. There exist, however, a few approaches that are intended to find out which factors influence the probability of the occurrence of knowledge spillovers. One of them proposes the use of patent citations as a proxy for knowledge flows between inventors.

1.2 Objectives

The patent citation approach is used in this thesis in order to examine the knowledge spillovers in which Austrian companies are involved. There are already a few studies for other countries using this approach. The goal of this analysis is to learn more about the factors that influence knowledge spillovers in Austria and find out how large their influence is. On a more precise level, this thesis tries to find answers to the following questions regarding Austrian firms:

- How is the appearance of knowledge spillovers connected to the spatial proximity of the source and the receiver of the spillover? Are there any signs of an increased probability of spillovers due to sharing a common language?
- Do some industries stimulate the occurrence of knowledge spillovers? Is there a difference between spillovers within an industry or across industries?
- How fast does knowledge diffuse and when does it become obsolete? Is there an interrelationship between the probability of the occurrence of a spillover and the temporal distance between the findings of bits of knowledge?
- Have the above factors changed over time? What is the difference between knowledge spillovers flowing into Austria and knowledge spillovers flowing out of Austria?

The intention of this thesis is to shed some light on these issues and provide an insight into the determining influences on knowledge spillovers in Austria.

1.3 Outline of the thesis

This thesis is organized as follows: The subsequent two chapters provide a theoretical introduction into the topic. Chapter two focuses on theoretical research on the concept of knowledge spillovers and provides definitions, classifications as well as criticism on the existing framework. Additionally, the concept of absorptive capacity is explained as well as the basic game theoretical models in order to put the topic into a wider context.

Chapter three, on the other hand, concentrates on the use of patent data research. Therefore, the basic characteristics of patents are summarized. Furthermore, a short literature review provides insight into the possible areas of application of patent data research and presents the results of previous articles. In this context, the results of a few earlier patent citation studies are presented.

In the fourth chapter, the data used in the econometrical analyses is described. The first part of this chapter focuses on a detailed explanation of the processing steps that were necessary to obtain the datasets. The second part provides preliminary results of the data as well as some conclusions on the citing behavior.

The fifth chapter features a description of the econometrical model as well as of the relevant explanatory variables. The data is then used for the computation of three different binary response models. The results of these regressions as well as the interpretation and the comparisons between the regressions of the different datasets can also be found in this chapter.

Finally, the sixth chapter presents the conclusions from the results obtained in the regression and provides an outlook on future research topics.

2 Knowledge spillovers

For a long period of time, economic theory treated technological progress as an exogenous variable which could not be explained by the neoclassic growth models. This insufficiency was addressed by the developers of the endogenous growth theory who tried to describe technological change within the model. Technological progress is based on the creation of knowledge, either by finding “new” knowledge by the means of intensive research or by linking “old” knowledge in new combinations. However, in many cases, the newly created knowledge is somehow revealed to the public or other economic actors, even (and maybe more so) if the creator wants to keep it secret. Although there are certain instruments to protect the rights of intangible property (e.g. via patents), knowledge is likely to flow to a firm’s competitors and provide them with valuable information at practically no costs.

This chapter of the thesis presents some of the results of the theoretical research on knowledge spillovers so far, including a short description of the basic game theoretical spillover models.

2.1 Definition

While the first endogenous growth models were not presented until the end of the 1980s, it was recognized much earlier that the creation of technological knowledge has side effects that influence not only the knowledge generating firm, but also its competitors and even firms outside of its industry (Griliches, 1979). These effects are referred to as “knowledge spillovers”, or synonymously “technological spillovers” and “Research and Development (R&D) spillovers”. An exact definition is provided by Grossman and Helpman (1991, p. 16):

> *“By technological spillovers, we mean that (1) firms can acquire information created by others without paying for that information in a market transaction, and (2) the creators (or current owners) of the information have no effective recourse, under prevailing laws, if other firms utilize information so acquired.”*

The existence of spillovers is rooted in the nature of technological knowledge as a non-rival and partially non-excludable good, as explained in the following section.

2.1.1 Characteristics of knowledge

Technological change constitutes one of the foundations of the endogenous growth model. In contrast to neoclassic growth models, Romer (1990) proposes that technological progress happens mainly due to commercially motivated innovative activities. However, knowledge in general and technology, as a form of knowledge, in particular differ from regular private goods in two important characteristics: they are non-rival and partially non-excludable. Thus, knowledge can be referred to as a kind of public good.

Rivalry determines the degree to which a good can be utilized by two or more parties. Thus, a good is fully rival if usage by one party precludes everyone else from using it. Technological knowledge is fully non-rival in that its usage by one party has no effect on the use of other parties. Knowledge can be utilized by an unrestricted number of parties, even at the same time, without experiencing a depreciation of its value. In contrast, a conventional economic good, e.g. a production machine, can be operated only by one party at a time (Cornes and Sandler, 1996). According to Romer (1990), non-rival goods are not affected by any physical boundaries and can basically be accumulated without limits.

Excludability is a measure for the degree to which the owner of a good can prevent others from using it. While the owner of a production machine will have no problem in excluding everyone else from using it, the developer of a new technology will face some difficulties when trying to prevent others from exploiting her ideas. Thus, knowledge is a partially non-excludable good. However, the legal system (in particular the intellectual property rights) plays an important role in determining the degree of the excludability. While some countries provide a comprehensive Intellectual Property Rights (IPR) framework in order to protect the creators of new knowledge, other countries restrict their rights in order to promote competition. Furthermore, the issue of enforcing property rights also influences the excludability of a good (Grossman and Helpman, 1991).

2.1.2 Absorptive capacity

The two characteristics described above are responsible for the occurrence of knowledge spillovers. However, so as to utilize incoming spillovers, a firm has to develop an “absorptive capacity”. This means that a company has to already possess a certain level of knowledge on a given topic in order to fully appropriate the value of knowledge that spills over from other firms. Cohen and Levinthal (1990) find that absorptive capacity is subject to path dependence because once a firm decides not to invest in R&D on a certain research topic, it loses the ability to efficiently exploit new knowledge regarding this issue. Thus, due to the cumulative nature of the absorptive capacity, the firm is “locked out”.

Cohen and Levinthal (1990) further state that a firm’s own R&D is the major source for its absorptive capacity (see Figure 2.1). Moreover, the aggregation of absorptive capacity is a self-reinforcing process because the higher the absorptive capacity, the more knowledge can be absorbed from outside the company, which itself increases the absorptive capacity. This self-reinforcement increases the degree of path dependence for the absorptive capacity. Furthermore, it is not enough for firms to focus on a specific topic of knowledge. Although this is very important in order to assimilate new knowledge on that matter, it is crucial that some of the knowledge is diverse and spreads over different areas of interest so that the new knowledge can be exploited in creative ways.

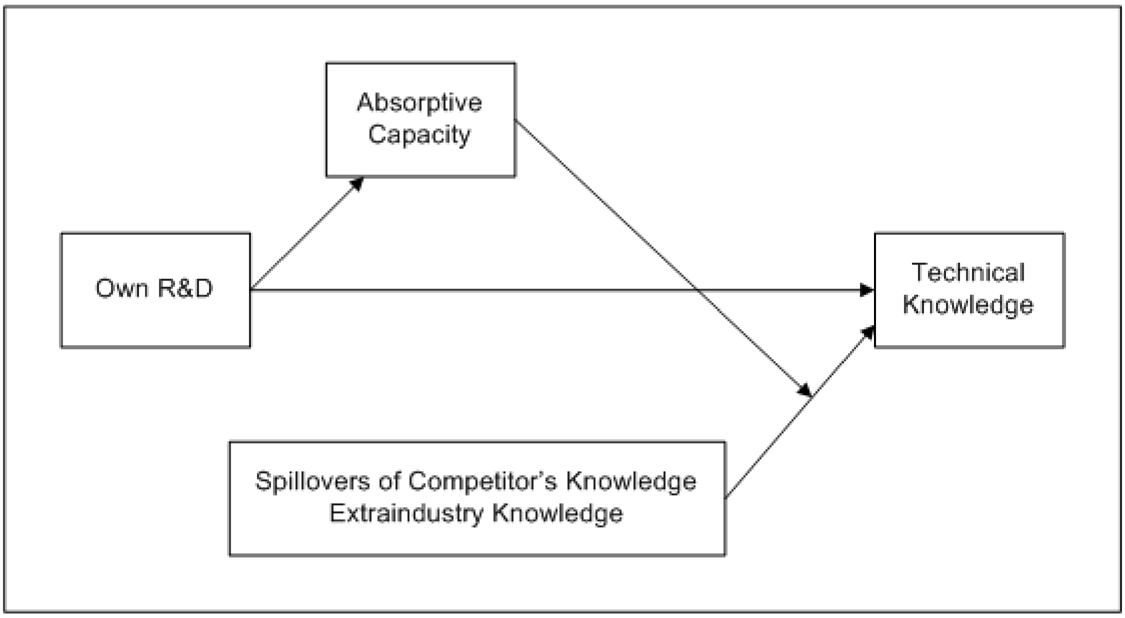

Figure 2.1 - Model of sources of a firm's technical knowledge (Source: Cohen and Levinthal, 1990, p. 141)

2.1.3 Two kinds of spillovers

Griliches (1979; 1992) distinguishes between two distinct kinds of spillovers, which are often mixed up or not differentiated properly. The first one refers to R&D intensive production inputs, which firms of one industry buy from firms of another industry. This capital equipment can lead to a productivity increase in the purchasing firm which is not attributable to the firm's own R&D, but to the research conducted by the selling firm. However, in many cases this impact is not fully reflected in the price of the equipment. Hence, the return of the selling firm's R&D is not fully appropriated by the firm, but instead spills over to the firm that purchases the equipment. In Griliches' opinion, this kind of spillover does not represent a real knowledge transfer, but rather a measurement error. Therefore, this kind of spillover is referred to as "rent spillover". One way to overcome this problem is to use a hedonic pricing model in order to quantify the improvements of the quality of the product.

As an example of this first kind of spillover, Griliches (1979; 1992) mentions the development of the computer industry. During the last decades, the computer industry has experienced enormous productivity growth rates and introduced new products with ever shortening innovation cycles. Not only companies from the computer industry have profited from these improvements, but also firms which introduced computers in order to handle their own production processes more efficiently. However, since the computer industry has been very competitive, prices do not reflect the real quality value of the products. Thus, rewards from

R&D performed by companies from the computer industry are appropriated by firms which purchase products from them.

The second kind of spillover, as described by Griliches (1979; 1992), refers to “real” knowledge spillovers, which cannot be explained by a measurement error. These knowledge spillovers happen when employees of one firm use ideas or solutions from another firm, which can be part of either the same or a different industry. It is possible as well that knowledge created by public research institutions, such as universities, spills over to private companies. These knowledge spillovers can proceed through various channels (e.g. informal networks among researchers of different companies, reverse engineering, hiring of a competitor’s employee, etc.) and can happen voluntarily as well as involuntarily (De Bondt, 1997).

2.1.4 Spillovers as externalities

In economic literature, knowledge spillovers are seen as a type of externality. However, up until now, two opposite views on externalities, especially on knowledge spillovers, have been developed and both of them seem to have their supporters in economic research. The main difference between the two models concerns the question whether knowledge spillovers are more likely to happen in a localized area with many firms from the same industry or in an urbanized area where firms from various industries are located (Huber, 2007). The Jacobian model proposes that areas with a diverse set of industries are most likely to grow due to cross-fertilization between different industries. Thus, cities and urban areas are the ideal place for innovations since they bring together people from different cultural as well as work background and foster inter-industry knowledge spillovers. Furthermore, the Jacobian model of externalities suggests that competition is more conducive to innovation than a monopoly because competition incites firms to innovate in order not to perish (Glaeser et al., 1992).

The foundation for the second theory on externalities was laid by Alfred Marshall. The theory of Marshallian externalities proposes that the localization of firms from the same industry has positive effects on their economic development. As listed in Krugman (1991), there are three externalities due to localization:

- Economies of specialization
- Labor market economies
- Knowledge spillovers

The first two entries of the list are referred to as “pecuniary externalities”, since they allow co-localized firms to access supplies and labor, respectively, at a lower price than their competitors that are located further away. Knowledge spillovers, on the other hand, are not part of market transactions and, hence, cannot be accounted for monetarily (Breschi and

Lissoni, 2001a). Intra-industry knowledge spillovers that occur due to localization of firms have become known as Marshall-Arrow-Romer (MAR) externalities due to the influential contributions on this topic by these three scholars. The MAR model proposes that the probability of a knowledge transfer is higher if the involved firms are part of the same industry and if they are located near each other. Therefore, the MAR model favors regional specialization. In further contrast to the Jacobian model, the MAR framework states that a monopoly provides more incentives for innovations, since an innovating firm is aware of the existence of spillovers and would prefer if it was the only one that benefited from its invention (Glaeser et al., 1992).

2.1.5 Localized knowledge spillovers and tacit knowledge

The concept of localized knowledge spillovers, introduced as MAR externalities in the previous section, has attracted much academic research recently and is, therefore, worth taking a closer look at. According to Breschi and Lissoni (2001b, p. 257), localized knowledge spillovers are *"[…] 'knowledge externalities bound in space', which allow companies operating nearby key knowledge sources to introduce innovations at a faster rate than rival firms located elsewhere."* Localized knowledge spillovers are often linked, implicitly or explicitly, to the concept of tacit knowledge. "Tacitness" refers to technological knowledge as being highly contextual and difficult to codify. Thus, tacit knowledge is communicated more easily by face-to-face contact and personal relationships, which in turn is facilitated by spatial proximity. Hence, many scholars propose that knowledge spillovers are localized. Furthermore, localized knowledge spillovers are regarded as one of the reasons why some regional areas appear to be more innovative than others (Breschi and Lissoni, 2001a).

Huber (2007) sums up some of the preconditions that have to be met in order for knowledge to diffuse efficiently. Local institutions and cultural aspects that encourage trust, entrepreneurship, social networks and a sense of belonging to the local community facilitate regional spillovers. Malmberg and Maskell (2002) mention two other factors that promote the transfer of knowledge between nearby firms of the same industry: observability and comparability. Observability means that collocated firms have fewer problems to inform themselves about the undertakings of their rivals because of spatial proximity. Companies in a localized area can observe the operations of their competitors without much effort and thus absorb their knowledge. Furthermore, comparability means that collocated firms can easily find out about the strengths and weaknesses of their nearby competitors since they operate under the same conditions and use this information to their advantage.

2.2 Spillovers in game theoretical models

Starting at the end of 1980s, the concept of knowledge spillovers appeared in various game theoretical models, especially in economic research dealing with R&D cooperations. The foundation for much subsequent research on that matter was laid by d'Aspremont and Jacquemin (1988; 1990) who consider a two-staged duopoly game on R&D cooperation, introducing a parameter that reflects the degree of spillovers. In order to make spillovers quantifiable for game theoretical purposes, two different approaches on how to account for spillovers have evolved in the literature: the output-oriented and the input-oriented approach. The first one models the effects of spillovers on the level of R&D output (e.g. cost reductions) from which both the innovating firm and (partly) its competitors benefit. The latter one considers spillovers of R&D inputs (e.g. R&D expenditures) that also influence a company's rivals.

In the following subsections a simple example of each approach will be discussed and their results will be compared.

2.2.1 Output-oriented model

The classic example for an output-oriented model can be found in the papers of d'Aspremont and Jacquemin (1988; 1990), henceforth AJ. They consider a symmetric two-staged duopoly with homogenous goods and linear demand as well as linear unit costs. In the first stage, both firms choose their unit cost reductions for which they have to pay by conducting R&D. The costs for R&D are quadratic, which should reflect the diminishing returns to R&D. However, the unit costs of one firm are not only affected by their own cost reduction, but also by the cost reduction chosen by the opponent. The degree to which the rival's cost reductions spill over is denoted by a continuous parameter, σ, which ranges from zero (no spillover) to one (complete spillover). In the second stage of the game, both firms take part in a Cournot competition.

AJ examine four different situations: (i) no cooperation between the firms, (ii) cooperation during the first stage in order to maximize the sum of the combined profits, (iii) cooperation on both stages, and (iv) the optimum for a social planner. A comparison between the first two cases yields that for sufficiently high amounts of spillover ($\sigma > 0.5$), the level of R&D increases when both firms cooperate in R&D, while the incentive for individual innovation decreases. This reflects an internalization of the external effects of R&D. In the third case, in which the two firms cooperate on both stages, the R&D level is at its highest, while the production output reaches a minimum (for $\sigma > 0.41$).

As apparent from these results, the AJ model is highly dependent on the value of the spillover parameter. The predictions of the model change when a critical value of σ is exceeded. For instance, in the noncooperative case, an increase in σ only leads to an increase in R&D (respectively cost reductions) if $\sigma < 0.5$. Otherwise, an increase of the spillover level will reduce the amount of R&D. Since equilibrium quantity and consumer surplus rise with the cost reductions, high spillover levels reduce the output and the consumer surplus (Martin, 2002). Despite these inconsistent predictions, this model has been very influential and was, inter alia, expanded to the case of an oligopoly (Suzumura, 1992) and asymmetric R&D investments (Salant and Shaffer, 1998). Furthermore, the AJ model has been corrected for reasons of stability (Henriques, 1990).

2.2.2 Input-oriented model

The second approach to account for spillovers in game theory is to model them as spillovers of R&D input. Kamien, Muller and Zang (1992), henceforth KMZ, present a model with n firms in which each company benefits to a certain degree from the R&D expenditures of its rivals. The amount of spillover is again denoted by a parameter, σ, that ranges from zero to one. Furthermore, their model introduces a continuous parameter, γ, to implicate the level of substitutability of the goods that each firm produces. For $\gamma = 1$ the products are homogenous and perfectly substitutable, while for $\gamma = 0$ each firm has a monopoly on its product. Again, the model consists of two stages: firms choose their R&D expenditures in the first stage and participate in a Cournot competition in the second stage.

The KMZ model is evaluated in four different situations: (i) the firms do not cooperate at all, (ii) the firms cooperate during the first stage in order to maximize the sum of their combined profits, (iii) Research Joint Venture (RJV) competition, and (iv) RJV cartelization. Cases (iii) and (iv) are similar to cases (i) and (ii), respectively, with $\sigma = 1$. In the model, a RJV is formed *"[…] if firms pool their R&D efforts so as to fully internalize the spillover effects"* (Kamien, Muller and Zang, 1992, p. 1297). Out of these four scenarios, the authors find the RJV cartelization the most desirable, since it leads to the highest profits for the companies and the lowest product prices. RJV competition, on the other hand, results in the highest product prices and the lowest investments in R&D and is, therefore, the worst option.

2.2.3 Comparison of output-oriented and input-oriented model

While the models presented in the previous two sections are intended to describe the same situation, they come to different conclusions in most of the cases. Amir (2000) sums up the differences and draws comparisons between the two models. In doing so, he uses the KMZ model with two firms ($n = 2$) and homogenous goods ($\gamma = 1$). The first obvious observation is that for the same degree of spillovers both models yield different results for produced quantities and prices. The only exception is the case without spillovers ($\sigma = 0$). Furthermore, the author finds that the nature of the joint returns to scale in the AJ model is dependent on the level of the spillover parameter. Large spillovers result in increasing returns to scale, while small spillovers yield decreasing returns to scale. This outcome is in contrast to the common conception that high spillovers are not very desirable by firms. The KMZ model, on the other hand, predicts decreasing returns to scale for all levels of σ.

Another interesting finding of Amir (2000) is that the spillover parameters of both models can be linked via a simple equation in order to achieve the same outcome for both R&D processes with the same inputs. For equivalence in the duopoly case, he derives

$$\sigma_{KMZ} = \sigma_{AJ}(\sigma_{AJ} + 2). \tag{2.1}$$

Thus, if σ_{AJ} and σ_{KMZ} satisfy equation (2.1), a given amount of R&D expenditures will lead to the same cost reductions in both the AJ and the KMZ model. However, equation (2.1) imposes a restriction on the maximum value of σ_{AJ}. $\sigma_{KMZ} = \sigma_{KMZ,max} = 1$ yields the maximum for the spillover parameter in the AJ model, $\sigma_{AJ,max} = 0.41$. Moreover, a spillover level exceeding $\sigma_{AJ,max}$ in the AJ model will lead to $\sigma_{KMZ} > 1$, which is beyond the model specifications.

Finally, Amir (2000) examines the validity and appropriateness of the two models. In his opinion, each model applies to different situations. While the AJ model seems to be very appropriate to describe particular events, the KMZ model applies to a more general environment. This is also reflected in the different predictions obtained by each model, which result in different policy recommendations. Furthermore, Amir (2000) expresses some concern about the validity of the AJ model, especially for higher levels of spillover. For instance, under certain circumstances it is possible that a firm benefits more from a R&D dollar spent by a competitor than from a R&D dollar spent by itself. In contrast, the KMZ model does not show these kinds of shortcomings. Nevertheless, many scholars are convinced that the AJ model is very valuable as starting point for spillover analyses.

2.3 Criticism on the spillover theory

The research on knowledge spillovers has created some important insights into the field of the diffusion of knowledge. However, there are some voices in the academic community that raise questions concerning the direction in which this branch of economic research is headed.

2.3.1 General criticism

First of all, Breschi and Lissoni (2001a) are not satisfied with the general use of the term "knowledge spillover". In their opinion, this term has been overloaded with a variety of meanings and used to subsume phenomena of knowledge flows that do not fit the original intention (compare to section 2.1.3). Therefore, they recommend to stick to the traditional interpretation of “knowledge externality”. Additionally, the authors propose to reassess the conceptual framework in order to introduce more elaborate categories so as to do justice to the complexity of the topic.

Another aspect of the criticism concerns the actual process of knowledge spillovers. Breschi and Lissoni (2001a) argue that – despite many empirical studies on the topic – knowledge spillovers remain merely a “black box” with ambiguous content. Furthermore, they claim that upon opening this black box, one will make three discoveries: Firstly, some of the knowledge flows that initially appear to be “pure” knowledge externalities turn out to be, in fact, pecuniary externalities mediated by economic mechanisms, such as the market for technologies and club or network agreements. Secondly, what at first appears to be involuntary leakage might indeed be part of a regulated knowledge flow that is managed with deliberate appropriation purpose. Finally, they argue that much of the knowledge flowing between firms does not serve as a nucleus for new innovations, but rather as acceleration of the development phase of new products and processes.

2.3.2 Criticism on the concept of localized knowledge spillovers

The issue of localized knowledge spillovers (see section 2.1.5) has attracted much of the research effort in the field of economic geography. However, as described in the literature surveys by Howells (2002) and Breschi and Lissoni (2001a; 2001b), there are some discrepancies in many of the reviewed studies. First of all, some scholars use too broad definitions of both geographical and technological areas so that one cannot draw any conclusions on the localization or the industry dependency of knowledge flows. Another point of criticism is the equalization of academic and industrial R&D in some studies. Breschi and Lissoni (2001a) consider academic research to be more basic than industrial research and therefore think that there is a larger time lag until the results of academic research enter the market in form of innovative products or processes. Furthermore, the authors criticize the tendency to interpret ambiguous results as proof of the existence of localized knowledge spillovers.

In connection with the topic of localized knowledge spillovers, there is also critique on the concept of tacit knowledge. Breschi and Lissoni (2001b) postulate that scientific or technological knowledge is not as tacit as it might appear at first sight. In their opinion, only the exchange of knowledge may be tacit, but the knowledge itself is not. They argue that while technological knowledge is highly specific, it can be codified in the jargon of the technological community and thus transmitted also over larger distances. This jargon, however, can be fairly different from the one of the broader social community, in which the firm and its workers are set. Furthermore, it is very difficult for nonmembers of the specific community to gain access to the "codebook" of the jargon. This can only be achieved by practical experience and learning-by-doing (albeit at very high costs). Due to these deliberations, Breschi and Lissoni (2001b) consider the exchange of tacit knowledge not as sufficient for explaining localized knowledge spillovers.

3 Patent data in economic research

The nature and characteristics of knowledge spillovers implicitly suggest a certain level of resistance towards measurement and traceability. Since many occasions at which spillovers happen are rather informal, there is little evidence or useful scientific data on the actual process of the knowledge flows. Furthermore, it is not possible to design an experiment which replicates the mechanisms of spillovers. Recently, however, the use of patent citation data has become a popular instrument to trace the evolvement of spillovers.

This chapter will feature a short description of the characteristics of patents and an overview of the general use of patent statistics in econometrical models. Furthermore, special attention will be drawn on previous studies that use patent citation data to measure knowledge spillovers.

3.1 Basic characteristics of patents

According to the World Intellectual Property Organization (WIPO), a patent is *"[…] a document, issued, upon application, by a government office […], which describes an invention and creates a legal situation in which the patented invention can only be exploited (manufactured, sold, used, imported) with the authorization of the owner of the patent"* (WIPO, 2004, p. 17).

In order to receive a granted patent right, the invention has to fulfill some criteria, which consist mainly of novelty, industrial applicability, a certain "inventive step" and the full disclosure of the invention in the patent application. Additionally, the subject of the invention must not lie within the group of non-patentable subjects (e.g. scientific theories, plants, animals and methods of treatment for humans). The inventor and the applicant of a patent are two separate aspects of a patent application. Both positions can be filled by the same person, but it is also possible that they differ, for example at so-called service inventions, at which the employer of the inventor serves as applicant (WIPO, 2004).

A patent enables its owner to exclude others from exploiting her invention. Hence, she gains a monopoly for a certain period of time (typically twenty years starting at the date of the application). Furthermore, if the patent is applied for at the European Patent Office (EPO) or under the Patent Cooperation Treaty (PCT), the validation of the patent is restricted to the countries that were declared in the patent application. On the other hand, if the patent is applied for at a national patent office, it is valid only in the respective country. The holder of the patent can also choose to sell or license her invention to others, in case she does not want to or does not have the capabilities to capitalize the invention herself. However, if there is a violation of the patent right, the owner has to take action by herself and file charges against the violator at court. There is no public authority that prosecutes patent violations by itself.

Citations constitute an important part of the patent document. The citations made in the patent application connect the patent to both preceding patents (“prior art”) and non-patent literature. The purpose of citations is to delimit the claims of the patent right in order not to violate the rights of patents on which the citing patent is based or topically adjacent patents. Citations can be added to the patent document by the applicant as well as by the patent examiner during the search or examination phase. Another – but less prominent – possibility to add a citation is during the opposition proceeding.

A patent documents an invention in every detail and makes it available to the public via the published patent specification. Thus, if it is a corporate patent, all the information about the invention is made accessible to a company’s rivals and competitors. Therefore, some companies prefer not to patent their inventions in order to keep important information a secret. Another strategy, which is used mainly by large corporations, is preemptive patenting. Companies that use this strategy try to receive a patent right not only for the actual invention, but also for devices that differ from the invention only in small parts and for devices that reach the same result as the actual invention with a different process. Thus, this strategy makes it very difficult for competitors of the firm to imitate the product without violating one of the various patent rights. Both examples support the fact that the meaning of patents has evolved from a legal right to a strategic tool for companies.

3.2 The use of patent statistics in economic research

Over the last decades patent data has become a widely used source for a variety of economic studies and analyses of the innovative behavior of firms and countries. The rapid increase in the utilization of this data is due to two factors: (i) The documentation of a patent covers a wide range of characteristics of both the invention and the inventor/applicant, and (ii) the data is easily available. According to Pavitt (1985), the first studies using patent statistics were conducted in the 1960s and examined the relationship between the size of a firm and the rate and direction of inventive activities.

Furthermore, Pavitt (1985) identifies three main directions in analytical patent statistics research. The first one tries to find and prove a causal connection between economic indices of firms (e.g. R&D expenditures) and their patenting activities. The second direction focuses on the impact of different Intellectual Property Rights (IPR) policies on the patenting behavior in different countries or sectors and draws comparisons between them. Finally, the last direction exploits the citations given in the patent documents by using bibliographic methods in order to find connections and knowledge flows between different technological and industrial fields. A few sample studies for each of the three directions are presented in the next subsections, following a short summary of the fundamental problems of patent data research.

3.2.1 Issues in patent data research

The main question that arises when discussing the problems of patent data analysis concerns the target variable: What aspects of economic activities can be measured with patent statistics at all? (Pavitt, 1985; Griliches, 1990). Many authors use patent data to measure the innovative output of firms. However, patents do not reflect the whole level of the innovative output, since some inventions may not be patentable while others are deliberately not patented. Furthermore, patents differ greatly in the magnitude of their value. While a large percentage of patents are more or less commercially worthless, only a small fraction of patents comprise a considerable economic value. Nevertheless, aside from economic value, patents can also serve as a means to mislead a company's rivals or to improve the protection against possible imitators (Griliches, 1990; Langinier, 2005).

The underlying problem is described in a simple model by Pakes and Griliches (1984). They define K as the level of economically valuable technological knowledge and thus $\dot{K} = dK/dt$ as the change of this level per time unit. $\dot{K}$ is the variable in which many economic researchers are interested because *"[…] it is the measure of innovative output"* (Griliches, 1990, p. 1670); however, $\dot{K}$ is unobservable. In the model depicted in Figure 3.1, $\dot{K}$ is one of the influences on Z_i $(i = 1,..,k)$, which represent several indicators of the benefits of invention and innovation (e.g. the stock market value of the firm and the productivity of traditional factors of production). Further influences on Z_i are other observable variables, X_i, and unobservable disturbances, v_i.

However, the more interesting part of the model is the lower section of Figure 3.1. $\dot{K}$ is dependent on the resources invested into inventive activities, R, and a stochastic process, u. The former contains an observable measure (e.g. past R&D expenditures or the number of research scientists), while the latter captures the uncertainty inherent to the innovation process as well as other informal sources of $\dot{K}$. Patents, P, are a flawed indicator of the change in valuable knowledge, since they are influenced also by a stochastic process, v, which depicts the randomness of the patenting process that occurs because not all inventions are patented.

These two relationships, the "knowledge production function" (containing R, u, and $\dot{K}$) and the "indicator function" (containing $\dot{K}$, v, and P), reveal the problem of patent data research: If one examines the relationship between research investments and patents, it is not possible to distinguish between the effects of the knowledge production function and the indicator function. Hence, without additional information, the impacts of u and v cannot be separated. This is especially troublesome since only u has an influence on Z_i. Furthermore, if patents are indeed reliable indicators of economically valuable knowledge, the magnitude and the variance of v have to be sufficiently small. Thus, if the last requirement is not met, it is possible that R is the better proxy for $\dot{K}$, even though u is not accounted for.

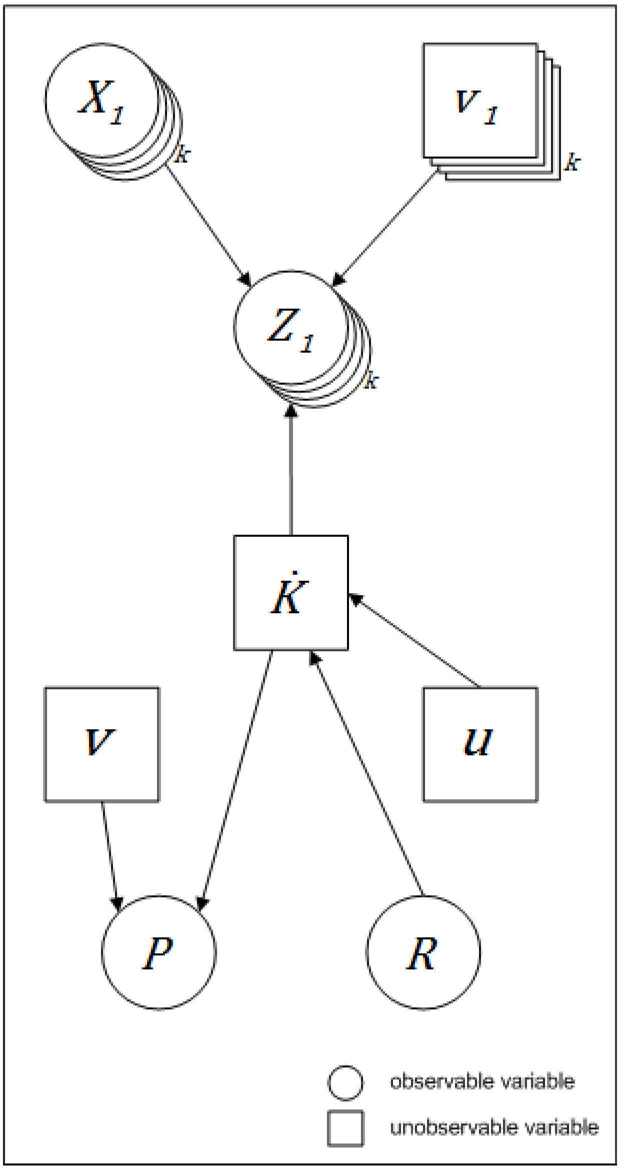

Figure 3.1 - Simplified path analysis diagram of the knowledge production model (Source: Pakes and Griliches, 1984, p. 56)

Aside from this fundamental problem, Pavitt (1985) offers a classification of sources of biases regarding patent statistics research: Firstly, there are differences in the perception of the value of patent protection among different countries. This concerns topics like the economic benefits of patent protection as well as the necessary effort to gain this protection (e.g. costs and duration of the patent examination as well as enforcement procedures). The second bias identified by Pavitt is the varying significance that is attributed to patent protection in different industries and technologies. Finally, a bias is created due to fact that the propensity to patent differs from company to company.

A survey conducted by Mansfield (1986) tried to examine the latter two biases. He gathered firm data on how many innovations would not have been introduced if there was no patent protection available and accumulated this data for different industries. He finds out that firms in industries, in which the patent system is regarded as less important, patent a relatively high amount of their inventions anyway (at least 50%). Nevertheless, this percentage is significantly smaller than in industries where patent protection is of high importance. Additionally, Mansfield (1986) finds a positive correlation between a firm's size and its propensity to patent inventions that are patentable.

3.2.2 Patent statistics as economic indicators

Many authors have tried to find causal relationships between patent data and all sorts of economic indices. The largest amount of work, however, has been put into finding a connection between the R&D intensity of a firm and its number of patents. Griliches (1990) cites several studies which pursue this path of research, one of them being Bound et al. (1984). Their study examines the relationship between the size of a firm (measured in sales) and the patent per R&D dollar ratio. The results suggest that small firms work more efficient than large firms (the patent per R&D dollar ratio decreases with higher sales until it reaches a certain level), yet there is no evidence that there exists something like "diminishing returns" for large corporations. The authors state, however, that the result of small firms being more efficient may be biased by the selection of companies used in the analysis. Other factors that account for the higher patent per R&D ratio of smaller firms are the larger amount of informal R&D in these firms which is not represented in the data and the higher dependence on patents of these firms in order to survive (Griliches, 1990).

The next branch of research deals with determining the value of patents. This proves to be a difficult endeavor, since there is no market at which patents are traded. Nevertheless, Griliches (1990) mentions several ways to assess the value of a patent. One of them is to send out questionnaires to the owners of patents and simply ask them about past revenues and future potential of their patent rights. However, this method has its flaws, since owners of a less successful patent may be more reluctant to return their questionnaire than holders of a prosperous patent. Furthermore, the mean and the median of the distribution deviate very strongly due to the high amount of patents with little or no economic value.

Another method to define the value of a patent is using data on the patent renewal fees. In Europe (and since the 1980s also in the USA) it is necessary to pay a fee in order to keep the patent right valid. The frequency of the payment ranges from once a year (e.g. in Austria and Germany as well as for EPO applications) to once in three and a half years (e.g. in the

USA). The data on patent renewal fees provides useful information on the private value of a patent to its owner, since she will only pay the fee if the benefits of keeping the patent right in force exceed the costs of paying the fee. This information can be used in order to reduce the noise in studies where simple patent counts are used as a measure of innovative output. Lanjouw, Pakes and Putnam (1998) present an approach to use patent renewal data as well as patent application data (representing the number of countries in which the applicant has applied for patent protection) in order to weight ordinary patent counts. According to these authors, a comparison of patent counts between different groups is likely to be deceptive if the mean values of the groups are varying. The various groups can consist (for example) of sets of patents from different periods of time, industries or home countries of the applicant. They start their analysis by constructing a patent value index VI as

$$VI = \sum_{j=1}^{J} w_j N_j, \tag{3.1}$$

where J is the number of different groups, N_j is the number of patents in group j, and w_j is the weight assigned to that group. Lanjouw, Pakes and Putnam (1998) base their model on the assumption that the value of patent protection is proportional to the average value of the patents. The value of patent protection is estimated by using patent renewal or application data. They then regress the value of patent protection on the number of patents in the different groups in order to get estimates of the weights. Thus, these weights represent the w_j needed for determining VI except for a factor of proportionality. The authors have simulated this procedure on their own data and gained some valuable insights, for instance that an average pharmaceutical patent elapsing at age four is more than five times as valuable as a similar patent lapsing at age three. Moreover, the authors present an overview of the distribution of the value of patents across different countries by comparing France, Germany, Japan, the UK and the USA.

Another way to assess the value of a patent is using the number of citations the patent receives. This method can be used to improve the explanatory power of simple patent counts. The implicit assumption is that the higher the number of citations a given patent receives, the more important is it or the greater is its value, respectively. Trajtenberg (1990) proposes a simple weighting scheme in order to calculate an index of weighted patent counts (WPC) for a given product class in a given year t as

$$WPC_t = \sum_{i=1}^{N_t} (1 + C_i), \tag{3.2}$$

where N_t is the number of patents issued in the product class in year t, and C_i is the number of citations a given patent i has received. Furthermore, the author also introduces a nonlinear weighting scheme where C_i in equation (3.2) is replaced by C_i^{α}. α is a factor dependent on the increase in social surplus as well as on the total social surplus. These schemes are then applied to a set of data from the computer tomography industry and it turns out that the citation weighted patent count index is indeed a good indicator of the value of innovation, while simple patent counts are a good indicator of the input of innovative activities, like R&D expenditures.

Another approach in patent data research is to connect patent statistics to the stock market value of firms. In an attempt to do so, Pakes (1985) derives a model for the relationship between the number of patents, the R&D expenditures and the stock market rate of return of a firm. He concludes that "unexpected" changes in the number of patents and R&D expenditures are significantly correlated with variations in the market value of the firm. However, he cannot confirm that small movements in the number of patents and R&D expenses have an impact on the stock market value. Moreover, by reviewing several other studies on this topic and filling the models with own data, Griliches (1990) comes to the conclusion that testing detailed hypotheses on the information content of patent data by using stock market value data is not very promising. He even compares it to *"[…] looking for our particular needle in a very large haystack"* (Griliches, 1990, p. 1688). Furthermore, another study on the same topic finds that while R&D assets are valued fairly well by the financial markets, patent counts add only a little amount of information for the explanation of market value variations (Hall, 2000).

In a further attempt to find a connection between patent data and the market value of a firm, Hall, Jaffe and Trajtenberg (2005) make use of patent citation data. They use the ratios of R&D to asset stocks, patents to R&D, and citations to patents to estimate Tobin's q. They find that an additional citation leads to a 3% increase of Tobin's q. However, the main parameter for explaining the market value is the R&D stock. Another finding of the study is that there are severe differences between different industrial sectors, e.g. in the drugs sector the impact of the citations to patents ratio is more than 50% higher than the effect averaged for all industrial sectors. Nevertheless, the authors conclude that patent citations are only usable for innovations that were introduced several years ago, since there is a truncation bias due to the fact that patents are cited many years or even decades after they were published.

3.2.3 Patent statistics as policy indicators

Apart from studies on the firm or industry level presented in the previous section, patent statistics can also be used to examine the appropriateness of IPR policies. The paper of

Lanjouw, Pakes and Putnam (1998) features this aspect as well by combining it with their findings about the use of patent renewal and application data. One approach they describe is to calculate the "implicit subsidy" rate of the patent system by calculating the ratio of the total private value of patent protection to related R&D expenditures. Another direction that is suggested in the paper is to use renewal data to measure the sensitivity of the revenue of patenting in dependence to various patent law changes that have occurred or will occur. Furthermore, Lanjouw, Pakes and Putnam (1998) present the result of their simulation of various changes in protection length, renewal fees and legal situation for a group of German computer patents.

Another study examines the impact of changes in the federal technology policy regarding national laboratories in the USA (Jaffe and Lerner, 2001). The authors of the study investigated the changes in patenting behavior of national laboratories due to several policy reforms in the 1980s. The most important reforms were the compulsory installation of a technology-transfer office at all federal laboratories and the right for academic and nonprofit institutions to retain title to patents funded by federal R&D. They find that the reforms have been successful since the patents per R&D Dollar ratio of national laboratories has equaled the one of American universities gradually. Furthermore, the overall quality of patents has not suffered from the increase in the number of patents. Thus, the authors conclude that the common negative public image of national laboratories in the USA is not appropriate nowadays.

The development of university patenting has also been a subject of patent data research. Henderson, Jaffe and Trajtenberg (1998) have analyzed the development of university patenting in the USA between 1965 and 1988, and have found a significant increase in the number of university patents. The higher number of patents stems from the policy changes mentioned in the previous paragraph as well as from an increase in industry funded university research. However, the authors cannot separate the impact of these influences from each other since all of them appeared roughly at the same time. Moreover, the results suggest that the increase in the number of patents is accompanied by a decrease in the quality or importance (measured by the number of received citations) of the patents.

3.2.4 Patent citation analysis

Patent citations provide a rich source of data and can be utilized not only to assess the value of a patent, but also to track down knowledge spillovers that have occurred between the inventors/applicants of the citing and the cited patent. One of the first attempts to use patent citations in this manner was made by Jaffe, Trajtenberg and Henderson (1993) in order to find out more about the localization of spillovers. Their research design involves the construction of a control sample and a "control frequency" which represents a reference value, against which the examined citations are compared. The authors scrutinize two cohorts of patents from different application years to find out that spillovers as indicated by patent citations appear to be localized, whereas the localization effect fades over time. However, technological proximity between citing and cited patent does not have an effect on the localization. Furthermore, there seems to be only little difference between university and corporate patents with respect to localization.

The basic model for measuring knowledge spillovers with patent citations is developed by Jaffe and Trajtenberg (1996; 1999). They examine a large set of citation pairs, which consist of both the potentially citing and the potentially cited patent. The relevant characteristics are the grant years of the patents, the location of the applicants and the technological field of the potentially cited patent. Due to the magnitude of the data set, the fact that the same combination of characteristics may appear numerous times and the categorical nature of the regression variables, the observations are combined into cells with the same characteristics. The dependent variable is thus replaced by the ratio of the number of pairs in one cell for which a citation occurs to the total number of pairs in the same cell.

The econometric model of Jaffe and Trajtenberg (1996; 1999) consists of two exponential processes: one depicts the rate at which knowledge diffuses, while the other one constitutes the rate of obsolescence. Both processes aim at diametrically opposite directions. The diffusion process increases the probability of the occurrence of a citation as time passes by, since the knowledge spreads to new spatial as well as technological areas. However, the process of obsolescence leads to a decrease of the citation likelihood, since a patent might lose its functional relevance as time goes by. The results of the regression indicate that spillovers are localized and that there is a greater proximity between inventors in the USA and Japan, compared to the one between the USA and Europe. A study that uses a similar approach, but focuses on the spillovers in Taiwan and Korea can be found in Hu and Jaffe (2003).

Though it seems reasonable to use patent citations as a measure of knowledge spillovers, it is necessary to check the validation and verification of the theory. Jaffe, Trajtenberg and Fogarty (2000; 2002) conducted a survey among inventors whose patents received at least one citation as well as among inventors whose patents cited other ones in order to find out more on this topic. The survey among the latter group included questions about three cited patents, of which one was a “placebo”. This placebo patent was a patent from the same patent class and issued in the same year as one of the actually cited patents, however, it had not been cited by that particular inventor. The assessment of the inventors of cited patents aimed at inventors whose patents were among either of the two actually cited patents from the other survey.

The results of the survey suggest that about 38% of the citing inventors were familiar to the cited inventions (placebos excluded) before or during the development of their own device. One third of the inventors came across the cited patent after finishing their own invention. Another interesting result is that 18% of the inventors of the citing patent had direct communication with the developer of the cited patent, while another 18% declared that there was some sort of indirect communication, for instance via word of mouth or by reading the patent document. Approximately 40% learned about the prior patent during their own patent application. Although about half of the citations do not seem to represent any kind of spillover, Jaffe, Trajtenberg and Fogarty (2000; 2002) conclude that citations can be seen as a signal of spillovers, albeit a noisy one. In a similar effort, Duguet and MacGarvie (2005) assess the validity of using patent citations as a proxy of knowledge flows by matching citation count data to French firms' responses to an innovation survey. They find that citations are significantly correlated to the acquisition and dissemination behavior of firms regarding new technologies. They emphasize the positive correlation between citations issued by the surveyed firms and learning through R&D cooperations as well as licensing of foreign technology.

A more critic evaluation of the method of using patent citations as an indicator of spillovers can be found in the paper of Alcácer and Gittelman (2006). Their analysis is based on a change in the US patent law from 2001, which made it possible to distinguish citations that were added by the inventor from those added by the patent authority examiner. They find out that the exclusion of examiner added citations from a pooled sample can lead to a change of the estimates in some cases. However, the bias depends on the tested hypothesis. Another – rather surprising – finding is that examiners are more likely to add "self citations" (citations where the applicant of both the citing and cited patent are the same person or entity) than the applicants themselves. In a related paper, Criscuolo and Verspagen (2008) further scrutinize the differences between citations added by the inventor and those added by the examiner. Their findings using EPO patent data suggest that there are indeed differences between the two kinds of citations. Inventor-added citations are more localized and occur more often within the same industry than examiner-added citations.

In a recent study, Bacchiocchi and Montobbio (2010) examine if there are any substantial differences between USPTO (United States Patent and Trademark Office) and EPO patent data. They use the approach of Jaffe and Trajtenberg (1999), but analyze different sets of citation data stemming from applications of the two patent offices. The authors find that there is, in fact, a patent office bias. Furthermore, while some of the results of Jaffe and Trajtenberg (1999) are confirmed, Bacchiocchi and Montobbio (2010) find that – in contrast to the previous outcomes – the USA is not the most open and innovative country according to EPO data. Moreover, different rates of knowledge diffusion in different industries are confirmed by the result of the study.

Despite the objections mentioned above, the method of using patent citations as a proxy for knowledge spillovers has grown quite popular in Europe, resulting in various studies, some of which utilize a slightly different model. Maurseth and Verspagen (2002) investigate whether geographical distance, national borders and language differences have a negative impact on knowledge spillovers in Europe. The remarkable aspect of this paper is that the patents were aggregated on a combination of NUTS1 and NUTS2 regions[1] in order to measure the geographic distance more precisely. The results show that speaking the same language has a positive effect on knowledge spillovers, even if the citing and the cited region do not belong to the same country. However, there is an even bigger positive influence stemming from both regions being in the same country. The distance has, as suspected, a negative influence on the probability of the occurrence of a citation.

Lukach and Plasmans (2002) have conducted an analysis of the spillover behavior of Belgian firms based on data from the EPO as well as from the USPTO. They use a binary response model in order to investigate the citations stemming from eight different industries. Based on their results, the authors classify the industries as “open” or “closed”, depending on the degree of inter-firm and inter-industry citations. They find that most industries are characterized by a rather closed behavior, since the slope coefficient for inter-industry citations is negative. The only exceptions are the sectors “Instruments” and “Computer and Office Machines”, which have a significantly positive coefficient slope for inter-industry citations. Furthermore, the authors recommend a subsidy policy based on their results in order to generate incentives for more R&D cooperations in industries with low knowledge spillovers.

[1] The “Nomenclature of Statistical Territorial Units” (NUTS; fr. *Nomenclature des unités territoriales statistiques*) is a hierarchical system in order to divide the member countries of the European Union into smaller regions. There are three levels (NUTS1, NUTS2 and NUTS3) which differ in the degree of hierarchy, with NUTS1 being the roughest and NUTS3 being the most precise level (see EUROSTAT, 2011, for more information).

4 Data and preliminary results

After describing the theoretical background in which this thesis is embedded, the following two chapters focus on the econometric analysis of actual patent data. The goal of the thesis is to find out which factors influence the probability of the occurrence of a patent citation, which in turn allows conclusions on the appearance of knowledge spillovers. This study concentrates on Austrian patent data on the firm level and examines citations in two directions: (i) Austrian corporate patents citing patents from any other country (including inter-Austrian citations), and (ii) citations issued in any patent referencing to an Austrian corporate patent. This allows us to draw conclusions on knowledge spillovers flowing from Austria to other countries and vice versa as well. Furthermore, for case (i), two periods of time are examined in order to find out more about potential changes in the citing behavior and the development of knowledge spillovers over time.

First of all, this chapter presents a summary of the concept of patent citation analysis. Furthermore, it focuses on the description of the used data and the methods of data processing as well as the presentation of preliminary results.

4.1 The intention of patent citation analysis

The foundations of patent citation analysis have already been discussed in section 3.2.4 along with the survey of previous studies on this topic. Nevertheless, it seems necessary to describe the underlying intention of this sort of analysis in detail in order to provide an understanding for the subsequent econometric models.

4.1.1 Using patent citations for the measurement of knowledge flows

Spillovers are generally regarded as disembodied flows of knowledge, offering very limited possibilities of quantification. However, *"[…] knowledge flows do sometimes leave a paper trail, in the form of citations in patents"* (Jaffe, Trajtenberg and Henderson, 1993, p. 578). As described in a previous section, citations constitute a very important aspect of a patent documentation, since they serve not only as a mean to give credit to a preceding invention (the "antecedent"), but also as a limit of the scope of protection of the new patent right (the "descendant"). Apart from the legal interpretation of patent citations, there is also an economic one. As Jaffe and Trajtenberg (1999) point out, patents represent a proxy for "bits of knowledge". Therefore, it seems straightforward to see patent citations as a proxy for the usage of existing bits of knowledge in the creation of a new bit of knowledge. Thus, a citation indicates that the invention of the descendant was facilitated to some degree by knowing about the antecedent.

However, one has to keep in mind that citations are “just” a proxy for knowledge flows and do by far not serve as a perfect measure for spillovers. For instance, there are variations over time and space in the propensity to patent as well as in the propensity to cite. This means that the creation of a bit of knowledge does not necessarily have to result in a patent and the usage of previously available bits of knowledge does not have to lead to a patent citation. Nevertheless, patent citation data provides a good starting point for the examination of knowledge spillovers, since it contains detailed information on the inventor and applicant of a new device as well as on the device itself. This allows econometric analyses on topics such as the spatial, technological and temporal distance between the citing and the cited patent.

4.1.2 Issues in patent citation analysis

Jaffe, Trajtenberg and Henderson (1993) insist that it is very important to distinguish between the different implications of patent citations and citations in academic articles. While both serve the purpose of giving credit to a source of inspiration, there is a huge difference in the cost of the citations. The cost of citing a source in an academic piece of work is approximately zero. It can be even negative if a long list of reference is used to apparently improve the significance of the research. However, a patent citation does result in a cost for the applicant of a patent since it reduces the scope of protection of the patent right. Thus, whenever a descendant references to an antecedent, the owner of the descendant patent loses some part of the value of her patent. Hence, it is unlikely that “gratuitous” citations as a favor for the owner of the antecedent dilute the patent citation data.

Nevertheless, there are other aspects of patent citation data that have to be considered carefully. Jaffe, Trajtenberg and Henderson (1993) point out that a possible link between two inventions can be attributed to one of three groups: (i) spillovers that result in a citation, (ii) citations that occur without the existence of knowledge spillovers, and (iii) spillovers that do not lead to a citation. The same authors further state that only the first group is relevant for a patent citation analysis. However, it cannot be avoided that citations, which actually belong to the second group, are part of the data set (e.g. when citations are added to the patent application by the examiner at the patent office). These citations are the main reasons for the noise in the measurement of spillovers. Group (iii) is likely to be the largest of the three groups since many inventions are not patented at all and thus the involved knowledge spillovers cannot be accounted for by citations. Jaffe, Trajtenberg and Henderson (1993) explicitly mention the basic research sector which, in contrast to the applied research sector, has a low patent rate and uses other mechanisms of communication in order to spread the results.

4.2 Overview of the data

The aim of this thesis is to examine knowledge spillovers in which Austrian firms act either as the sender or the receiver (or both) of the knowledge flow. Thus, different datasets were constructed. This section is dedicated to a detailed description of the data used and the way it has been prepared for the econometric analyses.

4.2.1 Data source

The data used in the analyses stems from the PATSTAT database issued by the European Patent Office (EPO) (PATSTAT, 2006). The PATSTAT database features tabulated information of patents applied for at patent offices all over the world. A new version of the PATSTAT database is released twice a year. The version used for this thesis was published in April of 2006 and marked one of the first editions of this database. A significant amount of work had to be put into filtering and cleaning the data in order to make it usable for an econometric analysis. The exact approach is described below in section 4.2.2.1. Another issue is that the personal data of the inventors and applicants is very incomplete, especially for patents applied for at national patent offices (in most cases the only personal information available is the name of the inventor/applicant).

The PATSTAT database was set up using a *MySQL* database system in order to account for its enormous size. The data is available on three DVDs and already organized in a relational data model (see Figure 4.1). Thus, the necessary information for one observation is split up into several tables and has to be rejoined in order to provide one dataset with the essential variables. As mentioned above, the patent data from national patent offices suffers from a lack of completeness of the personal records. Therefore, in order to allow an analysis of the geographical distribution of patent citations, only patents applied for at the EPO and the United States Patent and Trademark Office (USPTO) were used for the patent citation analysis. These data sources provide a reference to the country of origin of the applicant in most of the cases. Thus, the analyses in this thesis use only patent citations for which the citing and the cited patent were issued either by the EPO or the USPTO and for which there is information on the country of origin of the applicant available. The data used for the analyses consists of all granted patents (USPTO) and all patent applications (EPO) that were involved in a citation relationship that included an Austrian patent on at least one side of the citation pair in the relevant periods of time.

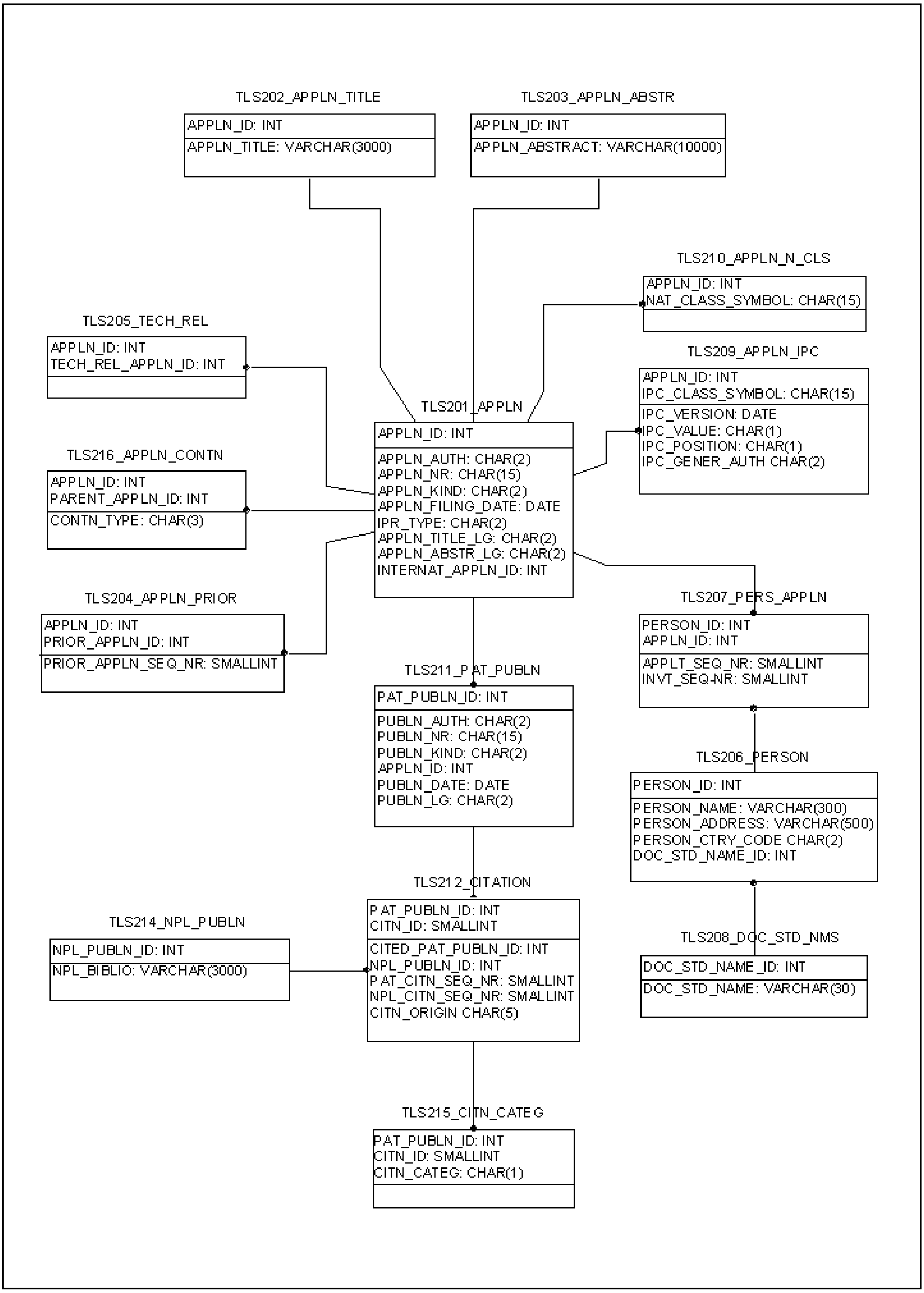

Figure 4.1 - Physical model of the PATSTAT database (Source: EPO, 2006, p. 85)

As depicted in Figure 4.1, a record for a patent application has many attributes. They include information about the application (e.g. application title and abstract, International Patent Classification (IPC), date of application), the applicant and the inventor (e.g. name, address, country of origin) as well as about citations to other patents and non-patent literature. This makes the database a rich source for statistical analyses. The physical model also takes into account that one patent application can have several applicants as well as several inventors. Furthermore, it is important to recognize the link between a patent application and a patent publication. One application can result in several publications, for instance if the search report conducted by the patent examiner is published separately. The application consists of the filing of all necessary documents at the patent office, while the publication of a patent is issued by the patent office after the examination of the invention and the (possible) grant of a patent right. Patent citations are linked via the publication ID because citations can be issued in applications as well as in search reports and related reports of minor importance published by the patent office.

4.2.2 Data processing

As previously mentioned, the raw data of the PATSTAT database is not useful for statistical analyses, since it has to be cleaned from duplicate records and “dummy” patents at first. Dummy patents have been introduced to the database in order to include citations in which the cited patent is not yet available in the database. However, since almost all of the attributes of dummy patents are either empty or filled with a default value, they cannot be used in a patent citation analysis. Duplicate entries of the same patent can occur due to a variety of reasons, for instance when a patent is applied for at several national patent offices without using the option of a single application for several countries at the EPO or under the Patent Cooperation Treaty (PCT). The main problem with these multiple records is that each entry has its own application ID. Thus, duplicates have to be identified on the basis of other attributes, for instance via a joint analysis of the name of the applicant, application filing date and the title of the application.

4.2.2.1 Preparation of citation pairs

Apart from the removal of duplicate and dummy patents, there were many other steps necessary in order to prepare the dataset (see Table 4.1 for a detailed description of the relevant data processing actions). The goal of the preparation was to obtain a set of “citation pairs”, which is comprised of both the citing and the cited patent. Firstly, all applications from Austrian citizens were extracted from the database. Furthermore, only applications issued at the EPO or USPTO were included, since they allow the recognition of the country of origin of the applicant. It is important to note that there are some differences in the citation behavior depending on the patent office. For USPTO applications, it is the duty of the applicant to reveal any prior art that has influenced or eased the invention of the particular device and of which she is aware (“duty of candor”). However, this is not the case for EPO applications, resulting in many citations being added by the patent examiner. This leads to a higher rate of citations per patent for patents applied for in the United States.

Previous studies offer a twofold picture: The results from Lukach and Plasmans (2002) suggest that it is possible to combine citations from EPO and USPTO applications and even use “cross citations” referencing from an EPO to an USPTO application and vice versa without making a large error. Bacchiocchi and Montobbio (2010), on the other hand, find a “home-bias” by comparing regression results of dataset with EPO and USPTO patents. In this thesis, it was decided to use both EPO and USPTO data in the same dataset in order to allow also for cross citations and to increase the number of observations. Nevertheless, in the preliminary analysis, the differences between EPO and USPTO data are analyzed as well.

Table 4.1 - Data processing

	Austrian patents
1	Select all applicants from table *TLS206_PERSON* with Austrian country of origin.
2	Join them with table *TLS201_APPLN* via *TLS207_PERS_APPLN* in order to gain Austrian applications.
3	Add the application title by joining the results with *TLS202_APPLN_TITLE.*
4	Select applications applied for at the EPO and USPTO.
5	Select corporate applications by searching for keywords in the applicant's name.
6	Filter out double entries, dummy patents and utility models.
7	Assign the new name and ID of the applicant to each application according to the classification of firms.
8	Join the results with table *TLS209_APPLN_IPC* in order to assign IPC codes to each application.
9	Assign field of technology and industry according to the IPC classes of each patent.
10	Select the most appropriate industry for every application which was assigned two or more different industries.
11	Join the results with table *TLS211_PAT_PUBLN* in order to assign publication IDs to each application.
	The citation link
12	Join the Austrian publication IDs with table *TLS212_CITATION* on field *pat_publn_id* to yield the citations where the Austrian patent is citing.
13	Join the Austrian publication IDs with table *TLS212_CITATION* on field *cited_pat_publn_id* to yield the citations where the Austrian patent is cited.

Table 4.1 - Data processing (continued)

	For each set of citation counterparts separately
14	Join the publication ID of the citation counterpart with table *TLS211_PAT_PUBLN* to yield the corresponding application IDs.
15	Join with *TLS201_APPLN, TLS202_APPLN_TITLE, TLS206_PERS_APPLN* and *TLS207_PERSON.*
16	Select applications applied for at the EPO and USPTO which have a non-empty field *person_ctry_code.*
17	Select the first applicant (or inventor, where no applicant is available) for each patent.
18	For Austrian applications: Add the new name and ID of the applicant to each application according to the classification of firms. All other applications maintain their name and receive a new *applicant_id* = 0
19	Assign IPC code and industry to each application just like in steps 8 to 10.
	Citation pairs
20	Rejoin the applications via the citation link. This yields two datasets: one, where all citing patents are from Austria (cf. step 12), and one, where all cited patents are from Austria (cf. step 13).
21	New field *time_lag*: equal to the difference between the year of application of the citing and the cited patent.
22	New field *same_firm*: equal to 1 wherever *citing_applicant_id = cited_applicant_id*; otherwise equal to 0.
23	New field *same_industry*: equal to 1 wherever *citing_industry = cited_industry*; otherwise equal to 0.
24	Select the relevant citation pairs for each period according to the application filing date.

Since utility models are also covered in PATSTAT, they had to be removed because they were not considered in the analysis. Afterwards, all patents from Austrian firms that either had cited another patent or had been cited by another patent as well as their citation pair counterpart(s) were identified. On the basis of these citation pairs, three separate datasets were constructed:

- Citation pairs in which the citing patent was applied for by an Austrian firm between January 1993 and December 1998.[2]

[2] As a result of the differentiation between the application and the publication of a patent, there is also an application filing date and a publication date. The difference between application and publication date depends on the efficiency of the patent office as well as the granting process and can amount to several years. For this analysis, the application filing date was used as relevant point of time because it is closer to the date of the actual invention. As Hall, Jaffe and Trajtenberg (2002, p. 410) put it:

> *"Indeed, the mode of operation at the Patent Office underwent significant changes in the past decades, thereby introducing a great deal of randomness (which has nothing to do with the actual timing of the inventions) into any patent time series dated by grant year. Thus, and whenever possible, the application date should be used as the relevant time placer for patents."*

- Citation pairs in which the citing patent was applied for by an Austrian firm between January 1999 and December 2004.
- Citation pairs in which the cited patent was applied for by an Austrian firm between January 1993 and December 1998.

The choice of the listed datasets was based on several deliberations: First of all, a period of six years ensures that there are enough observations in order to conduct an econometric analysis. The two periods of time (1993-1998 and 1999-2004) were chosen due to pragmatic reasons. The utilized PATSTAT database includes patent applications up until the end of 2005, but especially for the most recent years not all applications have been accounted for (in particular those that have not been published until a certain deadline). Therefore, as the application filing date approximates the end of 2005 there are fewer and fewer records in the database. However, in order to provide also relatively up-to-date results, 2004 was specified as the end of the second period.

Another consideration in the determination of the periods of time regards the third case of the list above. A patent can be – depending on the importance and "quality" of the underlying invention – cited for many years to come. Thus, there is a truncation bias inherent in the choice of the sample for the third analysis because citations issued from patents published after 2005 to Austrian patents of the selected period are not taken into account. The longest possible time lag in this case is twelve years (a patent from 2005 citing an Austrian patent from 1993). However, previous studies showed that the distribution of the time lag peaks after approximately five years and declines steadily afterwards (Jaffe and Trajtenberg, 1999; Lukach and Plasmans, 2002). Thus, in order to keep the balance between topicality and minimizing the bias, the above mentioned period of time was chosen.

The decision to examine three separate samples was based on two considerations. Firstly, it seems very reasonable to scrutinize the situation in which Austrian firms' patents are citing other patents as well as the situation in which Austrian firms' patents are cited. The former case provides hints of knowledge spillovers flowing into Austria, while the latter is a proxy for spillovers flowing out of Austria. In order to compare these two cases, it was necessary to choose samples which cover the same period of time. Secondly, it might be interesting to find out how the knowledge flow changes over time. Therefore, a third analysis is performed which again examines the patent citing behavior of Austrian firms, but deals with a later period of time. Due to the aforementioned truncation bias it is not practical to investigate the case in which Austrian firms' patents are cited by other patents for a subsequent period of time.

4.2.2.2 Obtaining firm level data

Knowledge spillovers can appear on many levels and are strongly influenced by personal relationships. Nevertheless, in this thesis we are interested in how knowledge spillovers affect Austrian firms. The counterpart of the Austrian firms in the citation pairs, however, can be a corporate entity as well as an individual. This is due to the below mentioned identification problem and the consideration that employees of firms can also include ideas and knowledge gained from private persons into their work. Thus, for the analysis it was necessary to aggregate the data and filter out the patents which had been applied for by Austrian firms. However, this proved to be a more complex task than expected. In the PATSTAT database, there is no attribute that either identifies a patent as applied for by a corporation or labels an applicant as company. Therefore, Austrian firms had to be recognized solely on the basis of their name provided in the record.

Hence, the database was scanned for applicant names that contained key words like “Gesellschaft” and “Genossenschaft” or abbreviations like “AG”, “KG” and “GmbH” as well as their variations. This scan created a list of patents that appeared to be fairly accurate, as a series of random checks suggested. However, due to the fact that the personal data had been entered manually into the database (and thus offered partly adventurous spelling), an extra step was necessary to aggregate patents belonging to the same firm. This was done manually due to the manifoldness of possible firm names.

Furthermore, the aspect of organizational structure was also addressed in this step. Large companies and holdings, such as the Voest Alpine AG, have applied for patents under the name of the parent company as well as under the name of their several subdivisions and subsidiary companies (e.g. Voest Alpine Stahl AG, Voest Alpine Industrieanlagenbau AG, Voest Alpine Bergtechnik AG, etc.). Since the examined period of time ranges over twelve years and thus many reorganizations as well as mergers and acquisitions have taken place during that period, it was decided to consider subsidiary companies as well as pre- and post-merger corporations as separate entities for the analysis. This reflects the situation that knowledge spillovers can also appear between firms belonging to the same parent company (but located in different places). This implicates that subdivisions of Austrian firms in other countries are treated as foreign companies. Therefore, a “self citation” (where the applicant of the citing and the cited patent is the same entity) can only appear if both the citing and the cited firm are from Austria.

4.2.2.3 Using the IPC code to assign industry classes

A very interesting aspect of knowledge spillovers is the extent to which they flow across industries. One would expect that flows within an industry (intra-industry) are higher than spillovers across the borders of industries (inter-industry). In order to measure the degree to which spillovers flow over industry boundaries, one has to determine the range of each industry first. This can be difficult, since such a definition of industries will never be complete and is always subject to exceptions. It does not seem reasonable to assign a certain industry to a firm since many companies produce a variety of goods that do not fall in the same category. Thus, for this analysis, an industry is assigned to each patent, according to the International Patent Classification (IPC) code it has received from the publishing patent office. The IPC is a very detailed classification system that assigns a technological category to patents and is updated in certain intervals. The IPC is hierarchically organized and consists of five levels. An example for an IPC code is given in Figure 4.2, which also provides a hint of the level of complexity of the entire classification system.

The intention of using IPC codes to assign an industry to a patent is straightforward: Since the IPC code is attributed to the patent by the patent examiner during the search and examination phase, it reflects the insights that the examiner gained during the investigation of the patent. The patent examiner is – aside from the inventor of the patent – the person who is most qualified to decide in which field of technology or industry a patent fits best. Thus, by assigning an IPC code to a patent, the examiner provides valuable information regarding the industry classification of the patent. However, this decision is not always unambiguous because in many cases two or more IPC codes are assigned to a patent. But since the classification takes place on a very detailed level, different IPC codes can nonetheless refer to the same industry.

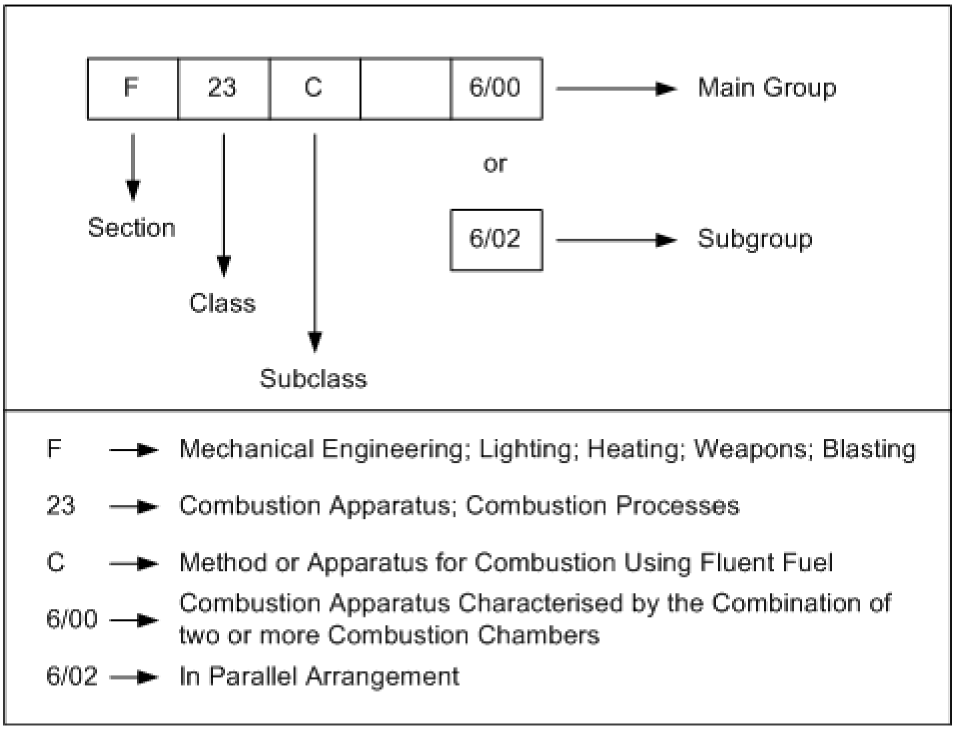

Figure 4.2 - Breakdown of class F23C 6/02 as example for IPC codes (Source: Based on WIPO, 2011, p. 5)

Table 4.2 - Technology and industry classification (Source: Field of technology based on Schmoch, 2008)

Field of technology according to Schmoch (2008)	Assigned industry	Industry No.
Electrical machinery and apparatus, electrical energy	Electrical engineering	1
Audio-visual technology	Electronics	2
Telecommunications		
Information technology		
Semiconductors		
Optics	Instruments	3
Analysis, measurement, control technology		
Medical technology		
Nuclear engineering	Other machinery	4
Space technology, weapons		
Organic fine chemistry	Chemistry	5
Macromolecular chemistry, polymers		
Chemical and petrol industry, basic materials chemistry		
Pharmaceuticals, cosmetics	Pharmaceuticals and biotechnology	6
Biotechnology		
Agriculture, food chemistry	Agricultural chemistry and machinery	7
Agricultural and food processing, machinery and apparatus		
Surface technology, coating	Material engineering	8
Materials, metallurgy		
Chemical engineering	Chemical and environmental engineering	9
Environmental technology		
Materials processing, textiles, paper	Materials processing and handling	10
Handling, printing		
Machine tools	Mechanical engineering	11
Engines, pumps, turbines		
Thermal processes and apparatus		
Mechanical elements		
Transport	Transport	12
Consumer goods and equipment	Consumer goods	13
Civil engineering, building, mining	Civil Engineering	14

The remaining issue is the transformation from IPC code to industry. For this purpose, several concordance tables have been compiled by different scholars. For instance, the MERIT concordance table, as one of the first approaches, offers a conversion from IPC to the International Standard Industry Classification (ISIC) code (Verspagen, van Moergastel and Slabbers, 1994). However, in this thesis, a different concordance table is used, since both the IPC and the ISIC codes of the aforementioned paper are out of date. Schmoch (2008) uses IPC codes on the class and subclass level (in rare cases also on the group level) and assigns them to a certain field of technology. In Schmoch's paper, concordance tables for two different versions of IPC codes are presented. In this thesis, the one regarding the February 2005 version is used, since its IPC version matches with the IPC version of the PATSTAT database. The used concordance table features a list of thirty technologies that are grouped in seven main categories. However, in order to balance between the detailedness of the categorization, the number of observations for each category and the ease of interpretation, Schmoch's thirty technologies were arranged into fourteen groups that resemble the main industries (see Table 4.2).

Following the concordance table of Schmoch (2008) and Table 4.2, fields of technology and industry were assigned to each patent. As mentioned above, it is possible that one patent has several IPC codes. While some patent offices provide a label for the "main class" of a patent, this is not the case for the patents used in this thesis. According to EPO (2006), the order of IPC codes for a patent may be, for instance, alphabetical, thus offering no hint on the relevance of the various IPC codes. Therefore, in a first step, each IPC code of a given patent was connected to the corresponding industry. It occurred that one patent was allotted to two or more different industries. In these cases, it was checked manually, which industry was suited best for the given patent, according to the name of the patent and the applicant as well as (in a few cases) the abstract of the patent. In most of the cases, however, there existed a well-defined connection between the IPC code(s) of a patent and a certain industry.

4.3 Preliminary results

The previous sections of this chapter have shown how the datasets of citation pairs were obtained from the PATSTAT database. In this subchapter, some preliminary results and descriptive statistics for each dataset will be presented. This shall provide the reader with an understanding for the subsequent regression analyses of the data, which will use slightly modified datasets. For the descriptive analysis, Microsoft Excel 2007© was used. For the sake of clarification and ease of reading, the three datasets will be abbreviated henceforth in the following way:

- **IN_PER1**: Citation pairs in which the *citing* patent was applied for by an Austrian firm between January 1993 and December 1998, indicating knowledge spillovers flowing *into* Austria during period 1.
- **IN_PER2**: Citation pairs in which the *citing* patent was applied for by an Austrian firm between January 1999 and December 2004, indicating knowledge spillovers flowing *into* Austria during period 2.

- **OUT_PER1**: Citation pairs in which the *cited* patent was applied for by an Austrian firm between January 1993 and December 1998, indicating knowledge spillovers flowing *out of* Austria during period 1.

4.3.1 Basic statistics for the citation pair datasets

Table 4.3 and Table 4.4 present summaries of the main characteristics of datasets IN_PER1 and IN_PER2. As can be seen from these tables, IN_PER1 has about thrice as many observations as IN_PER2 although the length of the contemplated periods of time is the same. While the number of patents and citations stemming from EPO applications are on roughly the same level, there is a significant drop in the ones stemming from USPTO applications. However, this difference can be also attributable to missing data in the PATSTAT database, since the second period lies at the end of the time span covered in PATSTAT. Another interesting aspect is that there are on average six citations per USPTO patent, while there are only two citations per EPO patent in both periods. This indicates the different behavior towards patent citations in the United States and in Europe. It is surprising that there is such a large difference in the average time lag between IN_PER1 and IN_PER2. This indicates that the rate of obsolescence seems to have decreased in the second period of time. Moreover, one can see that in the second period there is also a difference in the average time lag between EPO and USPTO patents. The maximum time lag (53 years) for IN_PER1 is an outlier, since the second highest time lag is 27 years.

Table 4.3 - Basic statistics of dataset IN_PER1

	Entire dataset	Citing patent applied for at the EPO	Citing patent applied for at the USPTO
Number of citation pairs	10430	2780	7650
Number of citing patents	2638	1374	1264
Number of cited patents	8290	1931*	6359*
Average number of citations per citing patent	3.95	2.02	6.05
Average time lag (years)	8.61	8.42	8.69
Maximum time lag (years)	53	25	53
% of intra-industry citations	76.95%	80.90%	75.52%
% of non-Austrian cited patents	83.43%	85.79%	82.58%
% of self citations among intra-Austrian citations	81.08%	72.91%	83.50%
Number of firms holding citing patents	513	400	268

* refers to the number of *cited* patents applied for at the EPO resp. USPTO

Furthermore, while the percentage of intra-industry spillovers has remained on approximately the same level, we can observe a significant increase in the (already very large) percentage of citations of non-Austrian patents. This suggests a further internationalization of the knowledge flows. Another interesting figure is the percentage of self citations. Due to the previously mentioned narrow definition of a firm, a self citation is only possible in citation pairs in which both the citing and the cited patent are held by an Austrian firm. Despite this restrictive definition of the borders of a firm, we find for both periods a relatively high amount of self citations, even though it has decreased in the second period. The number of firms owning a citing patent has stayed on a similar level for the entire dataset. However, on a more specific level, the diversity of firms has risen for EPO patents, while it declined for USPTO patents (probably also due to lack of data).

Table 4.4 - Basic statistics of dataset IN_PER2

	Entire dataset	Citing patent applied for at the EPO	Citing patent applied for at the USPTO
Number of citation pairs	3582	3142	440
Number of citing patents	1666	1591	75
Number of cited patents	3369	689*	2680*
Average number of citations per citing patent	2.15	1.97	5.87
Average time lag (years)	11.37	11.64	9.41
Maximum time lag (years)	30	30	26
% of intra-industry citations	79.06%	79.12%	78.64%
% of non-Austrian cited patents	90.12%	90.42%	87.95%
% of self citations among intra-Austrian citations	68.08%	69.10%	62.26%
Number of firms holding citing patents	535	516	52

* refers to the number of *cited* patents applied for at the EPO resp. USPTO

Table 4.5 offers a summary of the basic statistics for OUT_PER1. The number of observations is the smallest of the three datasets which can also be attributed to the aforementioned bias that citations stemming from patents published after 2005 are not included in the PATSTAT database (see section 4.2.2.1). This is also reflected in the average time lag, which lies significantly below the one of the other two datasets. The degree of intra-industry citations is on the same level as for IN_PER1 and IN_PER2, while the percentage of non-Austrian citation pair counterparts is well below the values of the other two datasets. It is also remarkable that there is a variation between EPO and USPTO patents for this value of more than 8%. The degree of self citations is comparable to the one of IN_PER1.

Table 4.5 - Basic statistics of dataset OUT_PER1

	Entire dataset	Cited patent applied for at the EPO	Cited patent applied for at the USPTO
Number of citation pairs	2289	713	1576
Number of cited patents	1059	417	642
Number of citing patents	2025	851*	1174*
Average number of citations per cited patent	2.16	1.71	2.45
Average time lag (years)	3.21	3.11	3.25
Maximum time lag (years)	11	10	11
% of intra-industry citations	78.29%	81.63%	76.78%
% of non-Austrian citing patents	74.14%	68.16%	76.84%
% of self citations among intra-Austrian citations	80.57%	81.50%	80.00%
Number of firms holding cited patents	304	188	184
		* refers to the number of *citing* patents applied for at the EPO resp. USPTO	

4.3.2 The most active firms

An interesting aspect of patent data concerns the variety of firms that patent their inventions. Since the PATSTAT database offers patent data on the level of persons and firms, respectively, it is possible to determine the number of patents and citations for each firm. Table 4.6 lists the top ten Austrian firms regarding their number of patents and issued citations of dataset IN_PER1. Furthermore, the distinct shares for applications at the EPO and at the USPTO are presented in order to find out if there are any severe differences in the patenting behavior of the firms.

First of all, Table 4.6 shows that there are no big outliers in the data, which means that there are no firms which patent excessively in order to protect their inventions. The upper part of the table gives a hint on the composition of the Austrian industry of the examined period: From the ten most active patenting firms, six belong to the mechanical engineering and automotive sector. The other ones belong to the chemical and pharmaceutical industry, which usually have a very high rate of patenting due to the nature of their inventions. Another interesting characteristic is that some firms seem to focus their patenting efforts on either the EPO or the USPTO (e.g. Voest Alpine Industriemaschinenbau's share of patents at the USPTO is almost thrice as high as at the EPO, although the number of patents in this dataset is roughly the same for both patent offices). Furthermore, Table 4.6 suggests that some firms cite more than others. For instance, Immuno and Lenzing have approximately twice the share of total citations compared to their share of patents.

Table 4.6 - List of the firms with the most citing patents and issued citations in dataset IN_PER1

	Patents			
	Firm	Overall (in %)	EPO (in %)	USPTO (in %)
1	Plasser Bahnbaumaschinen	5.04	4.37	5.78
2	Voest Alpine Industrieanlagenbau	4.09	2.26	6.09
3	AVL List	2.84	2.47	3.24
4	Blum	2.73	1.97	3.56
5	Immuno	2.65	0.73	4.75
6	Andritz	2.62	1.75	3.56
7	Novartis Pharma	2.62	5.02	0.00
8	Steyr-Daimler-Puch	2.62	2.26	3.01
9	Lenzing	2.31	0.58	4.19
10	DSM Fine Chemicals	2.08	2.26	1.90
	Citations			
	Firm	Overall (in %)	EPO (in %)	USPTO (in %)
1	Plasser Bahnbaumaschinen	5.70	4.39	6.18
2	Immuno	5.40	0.90	7.03
3	Lenzing	4.89	0.54	6.47
4	Voest Alpine Industrieanlagenbau	4.39	2.59	5.05
5	Andritz	3.56	1.98	4.13
6	Blum	3.14	1.73	3.65
7	AVL List	2.82	2.66	2.88
8	Steyr-Daimler-Puch	2.42	1.98	2.58
9	Biochemie	1.92	0.47	2.44
10	HTM Sport- und Freizeitgeräte	1.92	1.87	1.93

A similar index of firms for the second dataset, IN_PER2, can be found in Table 4.7. While some firms have managed to stay in the list of the top ten patenting companies also in the second period, others have dropped out and new ones have entered the list. Most of these new firms operate in different sectors compared to the "classic" ones mentioned above: Tridonic is a lighting and electronic company while HTM (short for Head Tyrolia Mares) is a sports equipment firm with many ski and ski binding patents. Innova, on the other hand, is the patent collecting company of Doppelmayr, the famous Austrian ropeway manufacturer. The magnitude of the shares is reduced, since IN_PER2 features fewer patents but a higher number of firms. Furthermore, due to the lack of USPTO patent data, many firms appear to have no share of patents and citations originating from USPTO applications.

Table 4.7 - List of the firms with the most citing patents and issued citations in dataset IN_PER2

	Patents			
	Firm	Overall (in %)	EPO (in %)	USPTO (in %)
1	Plasser Bahnbaumaschinen	3.72	3.71	4.00
2	Novartis Pharma	2.64	2.77	0.00
3	AVL List	2.52	2.51	2.67
4	Andritz	2.28	2.07	6.67
5	Cytec Surface Specialties	2.16	2.26	0.00
6	Tridonic	2.16	2.07	4.00
7	HTM Sport- und Freizeitgeräte	2.10	2.20	0.00
8	Voest Alpine Industrieanlagenbau	1.44	1.26	5.33
9	Jenbacher	1.26	1.32	0.00
10	Innova	1.26	1.32	0.00
	Citations			
	Firm	Overall (in %)	EPO (in %)	USPTO (in %)
1	Plasser Bahnbaumaschinen	3.38	3.44	2.95
2	AVL List	3.24	3.41	2.05
3	Tridonic	2.88	2.16	7.95
4	Andritz	2.68	1.85	8.64
5	Novartis Pharma	2.21	2.51	0.00
6	Cytec Surface Specialties	1.95	2.23	0.00
7	HTM Sport- und Freizeitgeräte	1.93	2.20	0.00
8	Voest Alpine Industrieanlagenbau	1.70	1.18	5.45
9	Jenbacher	1.62	1.85	0.00
10	TCG Unitech	1.54	1.53	1.59

Finally, Table 4.8 features the same kind of list for dataset OUT_PER2. The list of firms is very similar to the one from Table 4.6, although with a slightly different ranking. Moreover, the magnitude of the shares of patents and citations is also on roughly the same level. Plasser Bahnbaumaschinen is the unchallenged leader in all datasets. A striking aspect of the three presented tables is that some “big players” of the Austrian corporate landscape (e.g. Siemens) are not part of the lists. This can have several reasons: For instance, these firms may prefer to patent only a small fraction of their inventions or file their applications at the national patent offices. Furthermore, it is also possible that their patents do not contain many citations or only citations to non-EPO and non-USPTO patents. Apart from that, the Austrian corporate structure with many Small and Medium-sized Enterprises (SMEs) is reflected very well by the list of patent owning firms.

Table 4.8 - List of the firms with the most cited patents and received citations in dataset OUT_PER1

	Patents			
	Firm	Overall (in %)	EPO (in %)	USPTO (in %)
1	Plasser Bahnbaumaschinen	5.85	4.32	6.85
2	Voest Alpine Industrieanlagenbau	4.44	3.12	5.30
3	Blum	4.25	3.60	4.67
4	Steyr-Daimler-Puch	3.31	3.36	3.27
5	AVL List	3.21	2.64	3.58
6	Immuno	2.83	1.20	3.89
7	Lenzing	2.83	0.72	4.21
8	Andritz	2.27	1.20	2.96
9	Schablonentechnik Kufstein	2.27	3.12	1.71
10	Grass	2.17	0.24	3.43
	Citations			
	Firm	Overall (in %)	EPO (in %)	USPTO (in %)
1	Plasser Bahnbaumaschinen	6.03	3.65	7.11
2	Voest Alpine Industrieanlagenbau	5.16	5.75	4.89
3	Blum	5.02	4.21	5.39
4	AVL List	5.02	3.65	5.65
5	Lenzing	4.02	3.09	4.44
6	Steyr-Daimler-Puch	3.45	3.37	3.49
7	Immuno	3.32	1.12	4.31
8	Grass	2.45	0.28	3.43
9	Engel Maschinenbau	2.14	0.98	2.66
10	Andritz	1.88	1.68	1.97

4.3.3 Influence of the time lag

An interesting characteristic of patent citations is the time lag between the citing and the cited patent. As mentioned before, the application filing date of a patent is used in this thesis to approximate the date of invention. Figure 4.3 depicts the share of citations as a function of the time lag for the first dataset (the outlier with a time lag of 53 years was omitted for this diagram). The graph illustrates very well the diffusion and obsolescence processes that determine the flow of knowledge: At first, the citation frequency rises steeply until it reaches its maximum after four to five years. During this period of time, the diffusion process is prevalent. Afterwards, however, the citation frequency starts to decline steadily, albeit at a smaller rate than during the rise before. During this phase, the obsolescence process is dominant as other, newer patents gain more importance and the knowledge embodied in the old patents becomes obsolete. For IN_PER1, there is no big difference between the sources of the patent data, even though the maximum for EPO patents is slightly higher and occurs at a larger time lag than for the overall data.

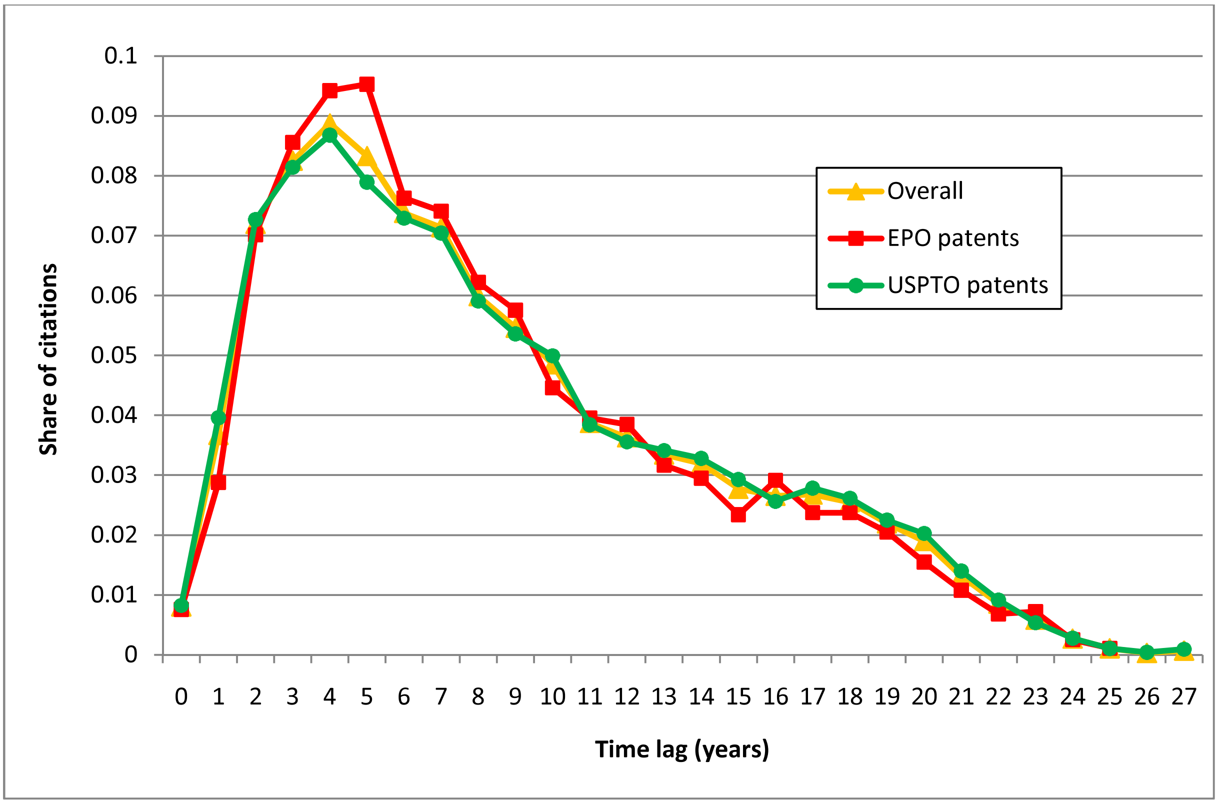

Figure 4.3 - Share of citations as a function of the time lag for dataset IN_PER1

Figure 4.4 shows a similar chart of the share of citations for dataset IN_PER2. Since there are fewer observations available for the second period, the course of the graph is not as smooth as for the first period. Especially the lack of data from USPTO patents is obvious because this part of the data is subject to very strong variations. Nevertheless, one can see that the maximum of citations occurs after six years and thus later than in the first period. Furthermore, the value of the maximum is below the one of IN_PER1. On the other hand, the tail is longer and fatter in the second period, suggesting a smaller difference between the rates of obsolescence and diffusion.

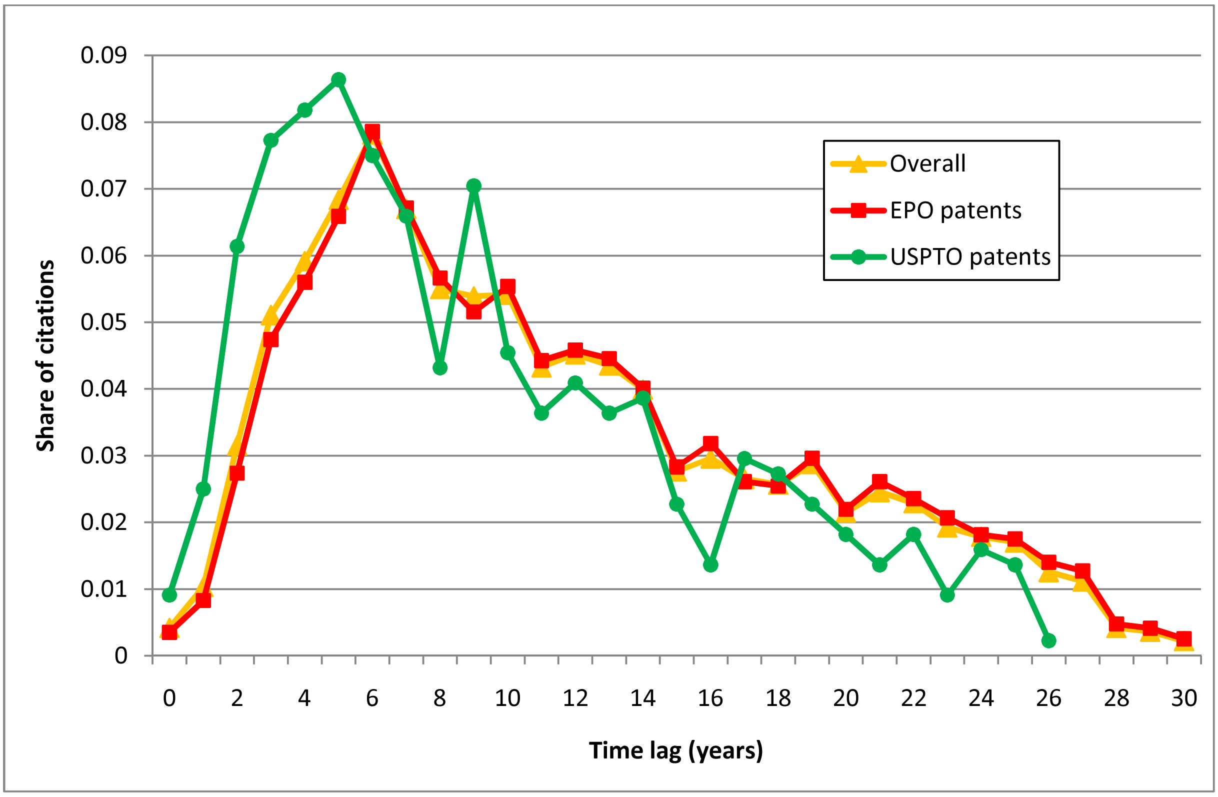

Figure 4.4 - Share of citations as a function of the time lag for dataset IN_PER2

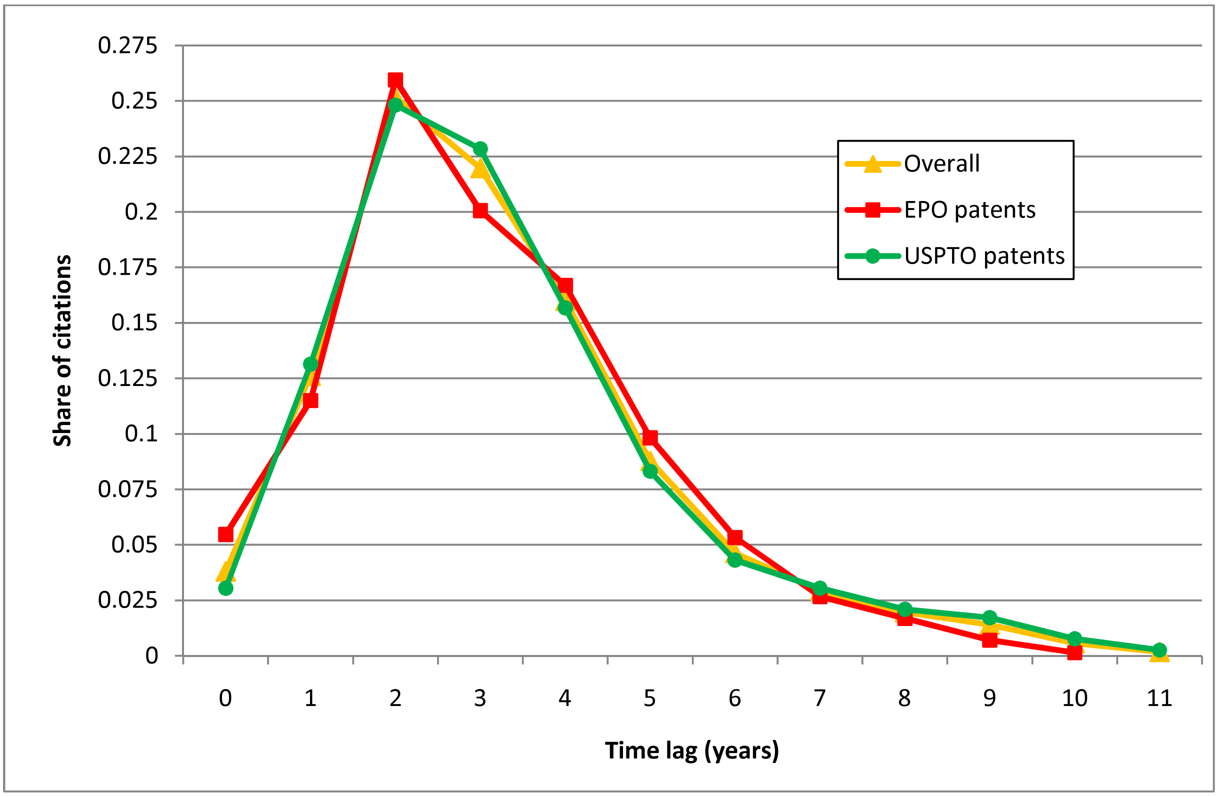

Figure 4.5 - Share of citations as a function of the time lag for dataset OUT_PER1

The percentage of citations dependent on the time lag for OUT_PER1 is depicted in Figure 4.5. The influence of the truncation bias is apparent. Due to the limitation of the maximum value of the time lag, the right side of the graph is underrepresented. Another consequence of this bias is that the maximum of the citation frequency occurs earlier and has a much higher value compared to the other two datasets. In order to reduce this bias, one would have to use patent data from an earlier period of time. Nevertheless, the typical structure of the chart remains the same for all three datasets. Furthermore, for OUT_PER1 the data from EPO and USPTO offer very similar results.

4.3.4 Geographical considerations

As mentioned in the chapter on the theory of knowledge spillovers, the spatial localization of spillovers is of high interest for economic research (see section 2.1.5). Much effort has been put into the examination of this topic. The version of the PATSTAT database used for this thesis provides information on the origin of the inventor or applicant only (if at all) on a country level, which is too inaccurate to examine the extent of the localization of knowledge spillovers. Nevertheless, it might be interesting to scrutinize where the applicants of patents, that are cited by or citing patents from Austrian firms, come from in order to find out among which countries a strong flow of knowledge exists. Furthermore, it seems useful to examine if there is a difference in the structure of the time lag among citations to patents from different countries so as to make a statement about the speed of the knowledge flow depending on the spatial proximity of two applicants. For the ease of understanding, “country of origin of a patent” will be used henceforth in order to indicate the country of origin of the applicant of a patent.

4.3.4.1 Geographic distribution of patents and citations

Table 4.9 features a list of countries with the highest share of cited patents as well as received citations in dataset IN_PER1. Dissenting to the localization hypothesis, the country with both the most cited patents and the highest share of received citations are – by far – the USA. While Austria is only fourth on the list regarding the share of cited patents, it is in second place when it comes to the share of citations. Furthermore, the high rank of Japan and Germany in Table 4.9 reinforces the common perception of these countries as being very innovative and on the edge of new technology. The other entries of the list are – with the exception of Canada – all European countries, suggesting a strong flow of knowledge within Europe.

A similar list for the second period (see Table 4.10) indicates a few changes. The USA is still on top of the list with an even increased share in both categories. While the share of German and Japanese cited patents and citations has approximately remained the same, there has been a decrease in both categories for Austria. Especially the decline in the share of received citations of Austrian patents of more than 6% is remarkable. This drop is not attributable to the reduced number of citing USPTO patents, since (if considered separately) both EPO and USPTO citations show a significant decrease of several percent. The rest of Table 4.10 shows a result similar to the one of Table 4.9, the only changes being the swap of Canada and Italy and Sweden entering as number ten of the list.

Table 4.9 - Geographic distribution of cited patents and citations for dataset IN_PER1

Country of origin of the applicant of the cited patent	Share of patents (in %)	Share of citations (in %)
USA	38.75	36.50
Germany	15.02	14.52
Japan	13.05	11.93
Austria	12.05	16.57
France	4.54	4.45
Switzerland	3.58	3.32
United Kingdom	2.69	2.68
Italy	2.39	2.40
Canada	1.56	1.43
Netherlands	1.04	1.01

Table 4.10 - Geographic distribution of cited patents and citations for dataset IN_PER2

Country of origin of the applicant of the cited patent	Share of patents (in %)	Share of citations (in %)
USA	42.68	42.32
Germany	14.93	14.74
Japan	13.45	13.34
Austria	8.79	9.88
France	4.60	4.55
Switzerland	3.00	2.90
United Kingdom	2.46	2.37
Canada	2.02	2.01
Italy	1.99	1.93
Sweden	1.37	1.28

Lastly, the distribution of countries of origin of citing patents from dataset OUT_PER1 is presented in Table 4.11. While the top four countries are the same as in the previous two cases, the magnitude of the particular shares has changed severely. US applicants account “only” for roughly 27% of the citing patents and issued citations, while Austria is in second place in both categories and can almost match the US share of citations. The German share in both categories is also higher than in the other datasets, while the Japanese one is slightly below the previous values. The other countries listed in Table 4.11 are the same as for IN_PER1, albeit sometimes in different positions. This indicates a bidirectional knowledge spillover relationship between Austria and other countries.

Table 4.11 - Geographic distribution of citing patents and citations for dataset OUT_PER1

Country of origin of the applicant of the citing patent	Share of patents (in %)	Share of citations (in %)
USA	27.70	26.74
Austria	22.32	25.86
Germany	17.28	16.43
Japan	11.51	10.48
France	4.59	4.28
Switzerland	3.16	3.10
Italy	2.96	2.75
United Kingdom	1.78	1.66
Netherlands	1.63	1.79
Canada	1.38	1.27

4.3.4.2 The relationship between country of origin and time lag

As mentioned before, there is a certain interest in the relationship between the country of origin of the citing patent’s applicant and the time lag of the citation pair. In order to facilitate the further analysis, most of the countries have been aggregated according to geographical deliberations. Thus, each cited patent (for OUT_PER1: each citing patent) is assigned to one of the following six groups of countries:

- Austria (AT)
- Germany and Switzerland (DE & CH)
- All other European countries (Europe)
- Japan (JP)
- USA and Canada (US & CA)
- All other countries (Other)

Germany and Switzerland are merged to one group because they are the only other German speaking and neighboring countries of Austria. For the other European countries, there was no other significant distinguishing feature detected, hence they are regarded as one group. The USA and Canada are viewed as a separate group due to their spatial proximity and their share of patents. The latter argument applies also to Japan. The share of patents and citations for other countries is so low that they are referred to as one group.

For dataset IN_PER1, a detailed diagram showing the connection between the time lag and the country of origin of the cited patent can be found in Figure 4.6 (the outlier with a time lag of 53 years was again omitted for this graph). The basic appearance of the six curves is the same as in the overall curve in Figure 4.3: a steep rise until the share of citations reaches its peak and a slow decline afterwards. However, as we can see from the graph, the time lag at which the peak occurs varies depending on the country of the cited patent. Thus, for cited Austrian patents, the share peaks at a time lag of two years, while for North American or German and Swiss patents, the peak does not appear until the time lag reaches four years. This indicates that spatial proximity does have an influence on the rate of diffusion of new knowledge. Austrian firms seem to profit earlier from the knowledge of other Austrian firms or inventors than from the knowledge generated outside of Austria.

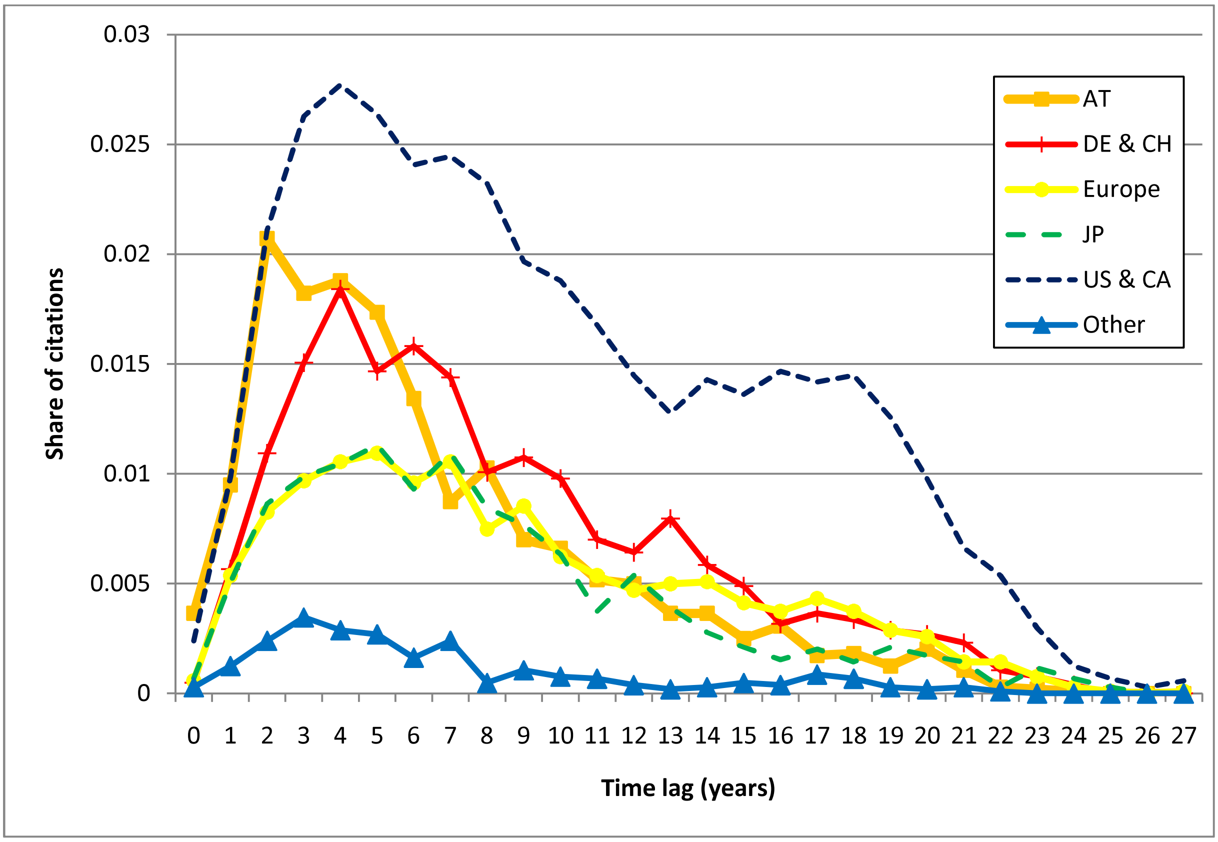

Figure 4.6 - Share of citations as a function of time lag and country of origin of the cited patent for dataset IN_PER1

Furthermore, for other European countries and Japan, there is no distinct peak of shares of citations identifiable in Figure 4.6 since the curve is rather flat for several consecutive time lags. Additionally, it is interesting to see that all curves decline rather strictly, except for the graph of the USA and Canada, which features a six year interval (between a time lag of 13 and 19 years) in which the share of citations is approximately constant. This suggests that patents from the USA and Canada contain knowledge which does not become obsolete as fast as the knowledge from other patents.

Figure 4.7 sums up the relationship between time lag, country of origin of the cited patent and share of citations for IN_PER2. Due to the dominance of cited patents from the United States and Canada, a separate scale (the right-hand side) was introduced for these patents in order to facilitate the interpretation of the chart. The maximum of the curve of cited Austrian patents occurs again very early, at two to three years. For North American patents, the peak is more distinct than in the previous period and happens at a larger time lag (approximately six years). Moreover, there is a range of the time lag in the right half of the diagram for which the share of citations of US and Canadian patents oscillates around a constant level of shares. This indicates again a slower rate of obsolescence of knowledge of North American patents. Citations to patents originating from Japan peak at a time lag of seven years, suggesting an even slower rate of knowledge diffusion. The peak for German and Swiss patents occurs at an unsuspected large time lag (about six to eight years), thus showing no signs of a faster diffusion due to the use of a common language.

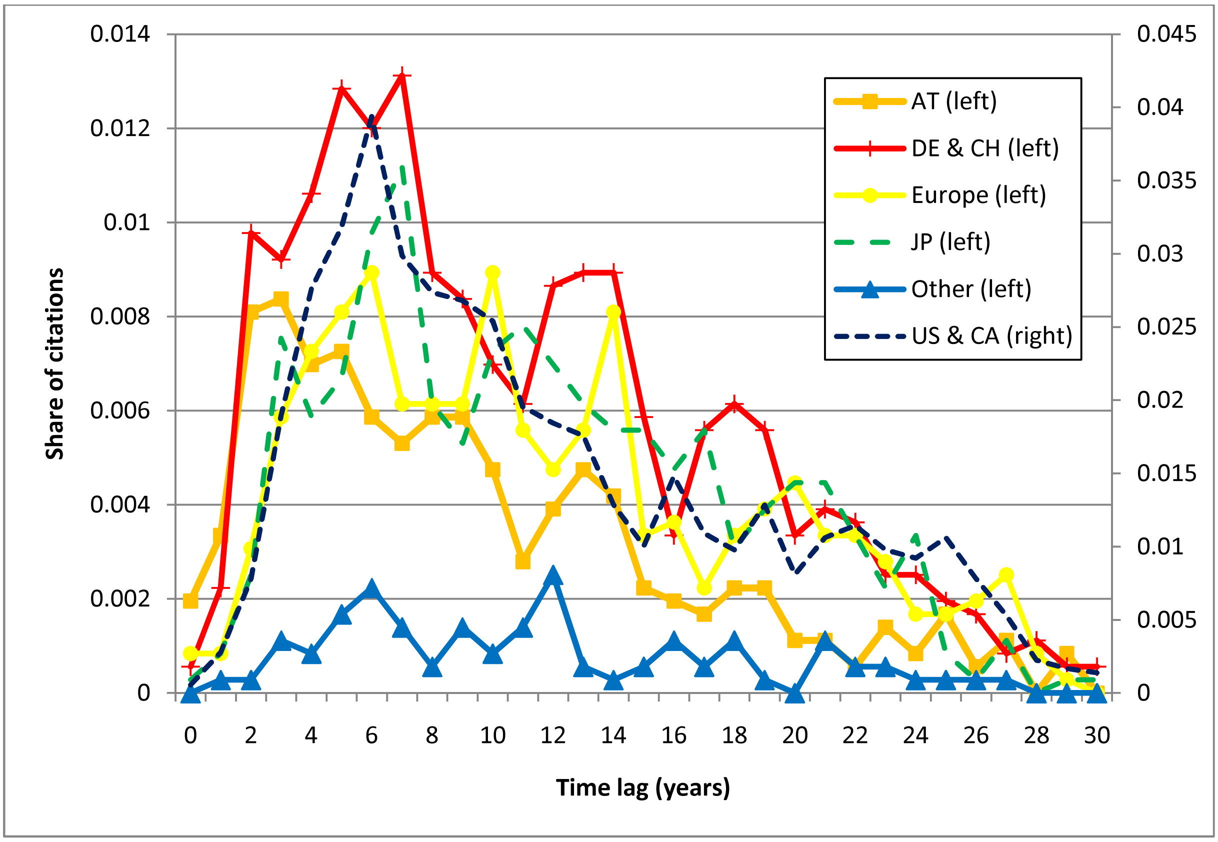

Figure 4.7 - Share of citations as a function of time lag and country of origin of the cited patent for dataset IN_PER2

Figure 4.8 presents the same kind of graph for the third and last dataset, OUT_PER1. Although the range of observations is limited, there are some trends obvious from the diagram. First of all, the Austrian curve is again the first to peak, suggesting once more a connection between spatial proximity and the speed of diffusion of knowledge. The curve representing patents from Germany and Switzerland is peaking also at a time lag of two years, signifying a fast rate of diffusion for outgoing knowledge spillovers due to the geographical proximity between said countries and Austria. This is in contrast to the results for incoming knowledge flows. The maximum of shares for citing North American patents, on the other hand, occurs at a time lag of three years. Contrary to the first two cases, the curve for patents from the USA and Canada shows no signs of a fat tail at all. Although this may be due to the lack of data for higher time lags, it is also possible that knowledge spillovers flowing out of Austria happen mostly with a shorter delay compared to spillovers flowing into Austria.

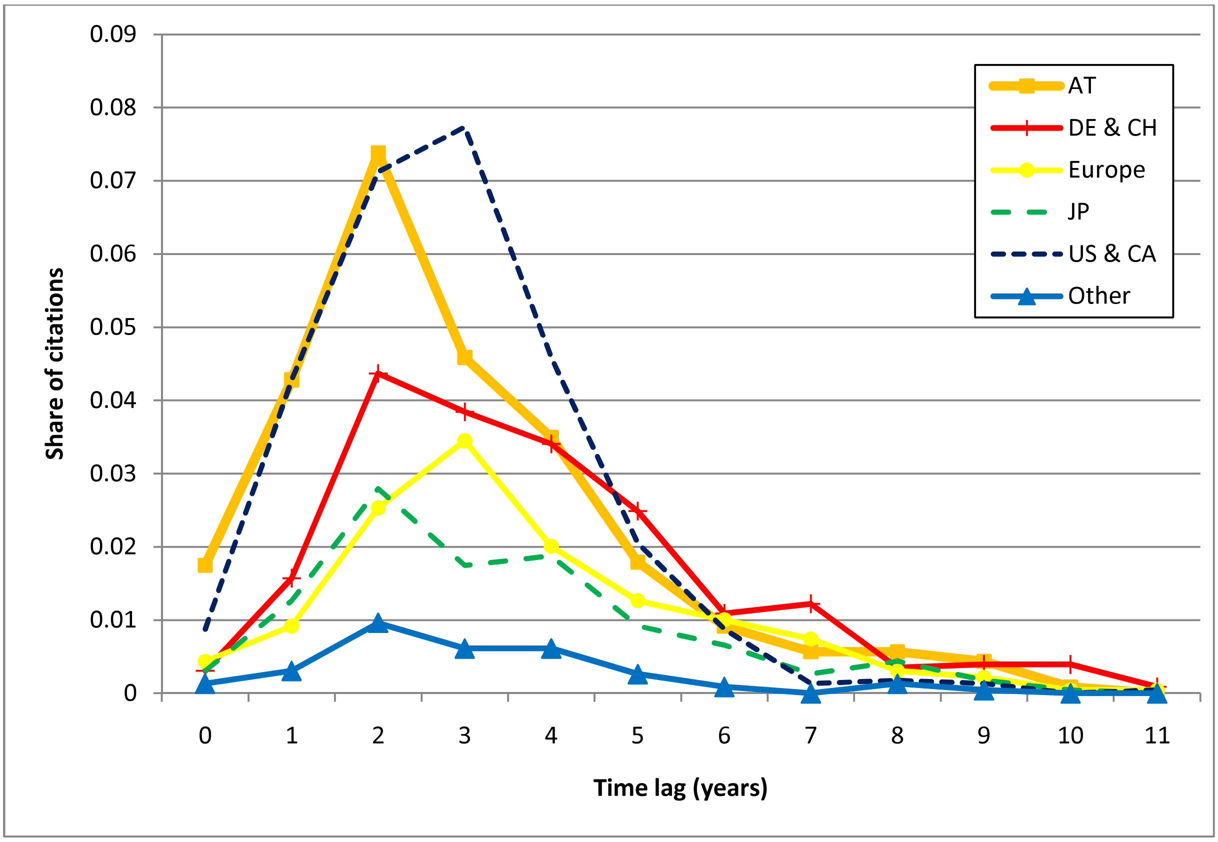

Figure 4.8 - Share of citations as a function of time lag and country of origin of the citing patent for dataset OUT_PER1

4.3.5 Industry related aspects

The final section of the descriptive analysis of the three datasets will approach questions on the relevance of the affiliation of a patent to a certain industry. For all graphs in this section, the numerical codes from Table 4.2 were used as shortcuts for the various industries. First of all, the distribution of patents and citations over the range of industries will be discussed. Afterwards, the focus will be laid on finding out between which industries there is a distinctive citation relationship.

4.3.5.1 Distribution of patents and citations among industries

For the first dataset, IN_PER1, a graph of the share of citing patents and issued citations for each of the fourteen industries can be found in Figure 4.9. Firstly, industries 4 (other machinery, includes e.g. weapons and nuclear engineering) and 7 (agricultural chemistry and machinery) appear to have by far the lowest share of both patents and citations. Although most of the citing patents are from mechanical engineering (11), these patents do not issue as many citations as the ones from industry 10 (materials processing and handling). The largest differences between the share of patents and the share of citations occur in industry 3 (instruments) and 12 (transport). Nevertheless, Figure 4.9 clearly displays the focus of the Austrian firms of the considered period: Industries related to machinery and plants are on top, followed by civil engineering (14) and chemistry (5). The electronic industry (2), on the other hand, has not been as popular in Austria according to the patent data.

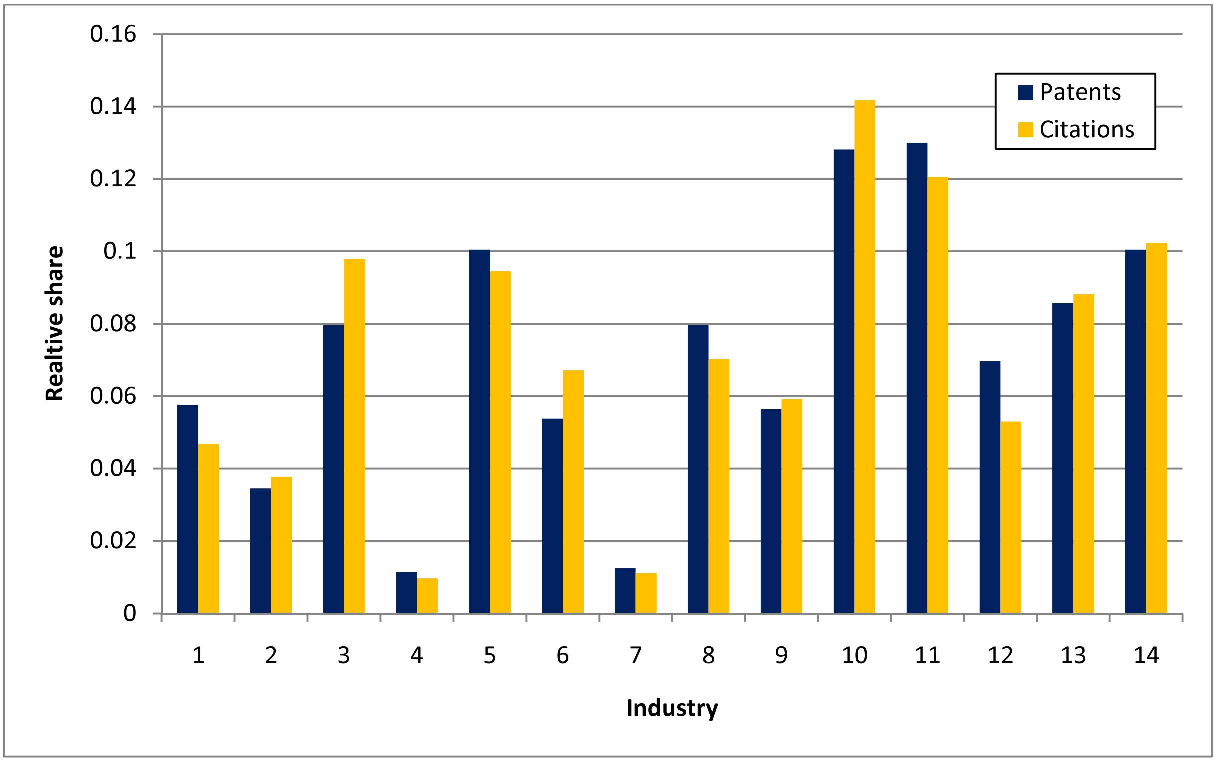

Figure 4.9 - Distribution of citing patents and issued citations per industry for dataset IN_PER1

Figure 4.10 presents a similar chart for dataset IN_PER2. Here, the center of the Austrian corporate landscape is even more obvious. The mechanical engineering industry accounts for almost one fifth of the issued citations. Nevertheless, some things have changed compared to the first period. The group of industries with the lowest share of patents and citations is joined by pharmaceuticals and biotechnology (6). The dividing line between the chemical and the pharmaceutical industry can be very blurry sometimes and thus some innovations might be part of both industries. In this case however, there has not been a trade-off between the shares of the two industries since the percentages of the chemical industry for the second period have also dropped by almost 5% each. The shares of industry 10 are reduced and on roughly the same level as industry 3 and 14. The instruments industry has already been very prominent in the first period, since it covers a broad collection of technologies, ranging from optical to measurement and control devices as well as some medical instruments. All in all, the chart of period two shows a more concentrated distribution of patents and citations compared to Figure 4.9, where the shares are more evenly spread.

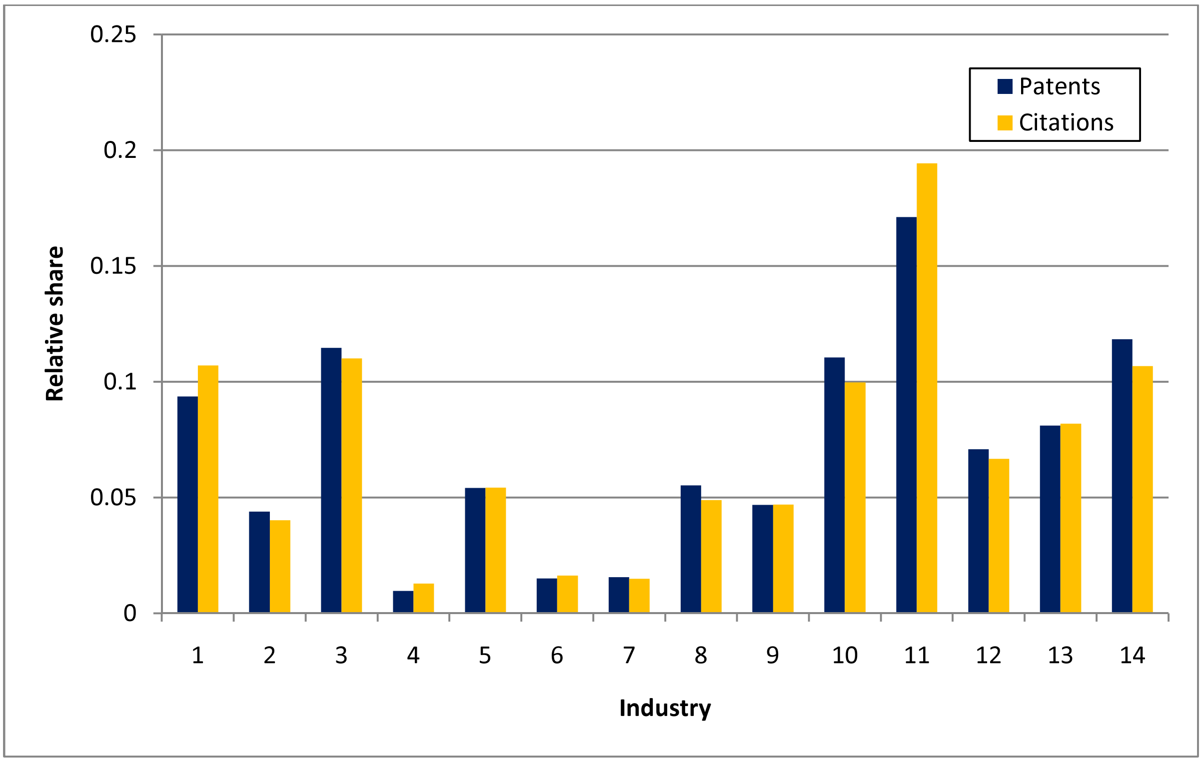

Figure 4.10 - Distribution of citing patents and issued citations per industry for dataset IN_PER2

For dataset OUT_PER1, the distribution of cited patents and received citations across industries is depicted in Figure 4.11. Once again, the leading industry in both categories is mechanical engineering, while industries 4 and 7 lag behind. Furthermore, the consumer goods industry (13) received the second most citations, while accounting only for the fourth highest share of patents. Nevertheless, in most of the industries, the shares of cited patents and received citations are on the same level. Civil engineering as well as materials processing and handling maintain their status as very important industries also in the dataset concerned with outgoing knowledge flows. Moreover, we find that industries 8 (material engineering) and 9 (chemical and environmental engineering) retain a place in the midfield throughout all three datasets. According to the patent data, the electronics industry is – compared to other industries – not very prominent in Austria, even though the chosen definition of the industry covers many new technologies that were emerging at a high rate during the examined periods of time.

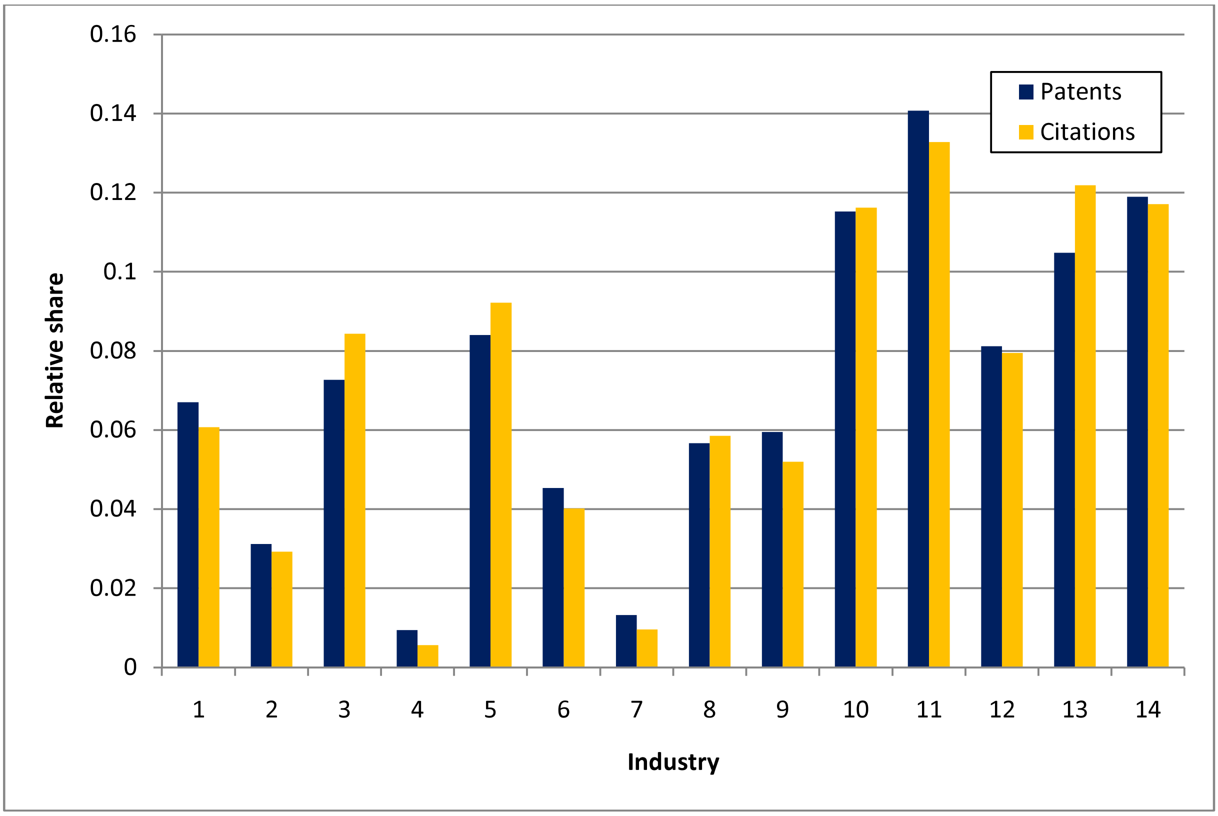

Figure 4.11 - Distribution of cited patents and received citations per industry for dataset OUT_PER1

4.3.5.2 Industry-citation map

In order to find out more about the citation behavior across different industries, it is useful to take a look at an "industry-citation map". This sort of graph resembles a three-dimensional map, in which the x-axis and the y-axis indicate the industry of the citing and the cited patent, respectively. The z-axis displays the share of citations that is allotted to each combination of industries. The map is not symmetrical since the share of citations from patents of a given industry to another industry is not necessarily the same as vice versa.

The industry-citation map for dataset IN_PER1 is depicted in Figure 4.12. It is obvious from the first sight that the main diagonal is very dominant. This means that the majority of citations are of intra-industry nature (76.95% of all citations). Furthermore, we see that the highest values of intra-industry citations occur for patents from industry 10 (materials processing and handling) and 11 (mechanical engineering), which are also the industries with the highest overall share of citing patent and issued citations (see Figure 4.9). Apart from the main diagonal, the map appears to be rather flat. However, one can still make a few interesting observations: The highest values of inter-industry citations occur between industries 5 (chemistry) and 6 (pharmaceuticals and biotechnology) in both directions (0.85%) and from patents of industry 10 to patents of industry 11 (0.80%). Furthermore, out of the 196 possible combinations of industries, there are 41 categories that contain no observation at all.

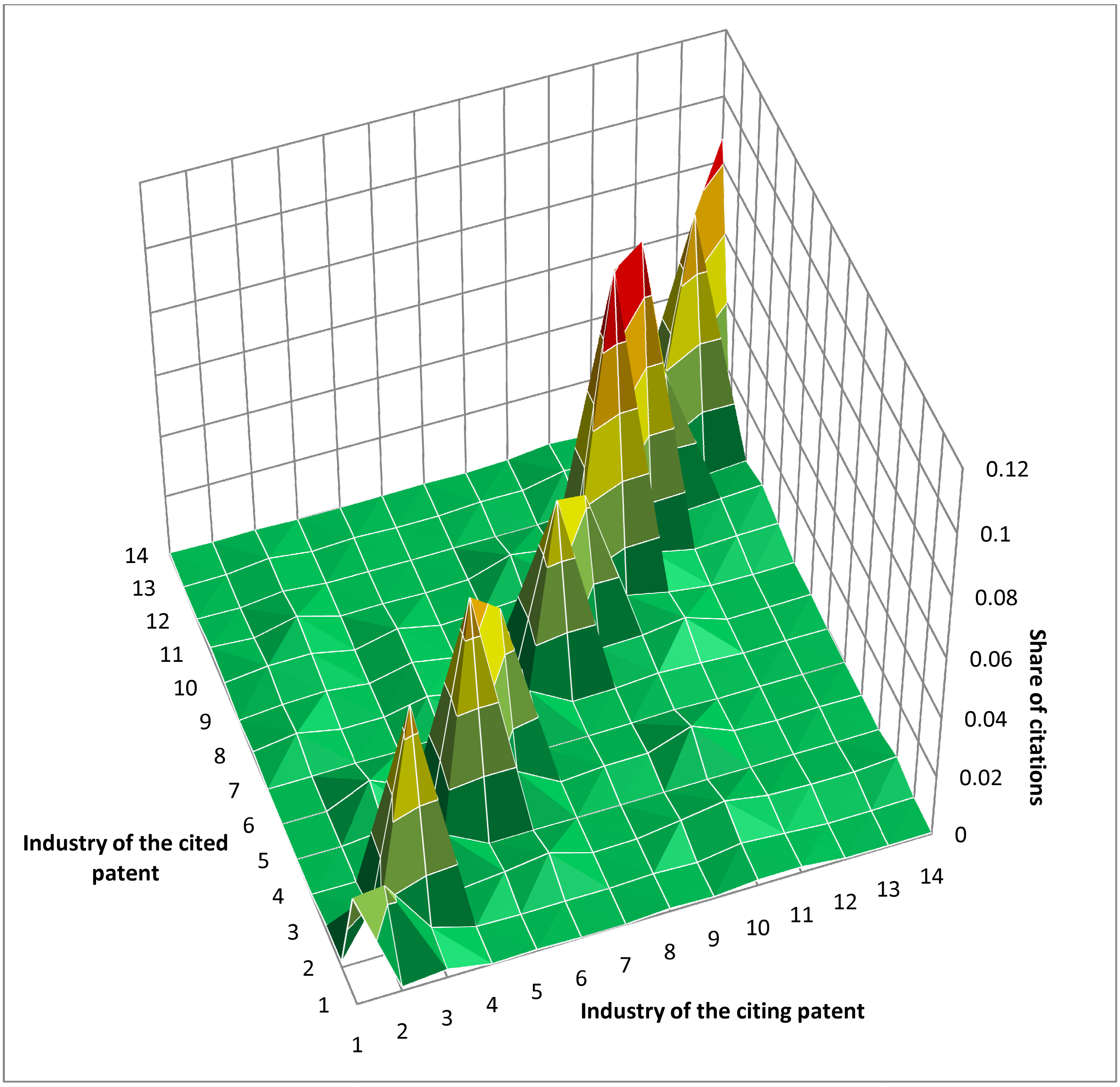

Figure 4.12 - Industry-citation map for dataset IN_PER1

Figure 4.13 displays the industry-citation map for the second dataset, IN_PER2. Again, the main diagonal indicating intra-industry citations is dominant. However, within this diagonal, industry 11 has with almost 16% by far the highest share of citations. Industries 4 to 9 show a rather low citation rate. The highest shares of inter-industry citations occur for citing patents from industry 3 (instruments): 0.89% when the cited patent belongs to industry 2 (electronics) and 0.7% to patents from the electrical engineering industry (1). This seems reasonable, since technologies from these two industries are very relevant for the measurement and control devices of industry 3. Another maximum value of inter-industry citations appears for citations from industry 11 to industry 10 (0.73%). Apart from these local maxima, the rest of the map is even flatter than for the first dataset. This suggests lower knowledge spillovers between different industries. This argument is even strengthened by the fact that almost a quarter (48 of 196) of the possible combinations of industries is empty and 79.06% of the citations are located on the main diagonal.

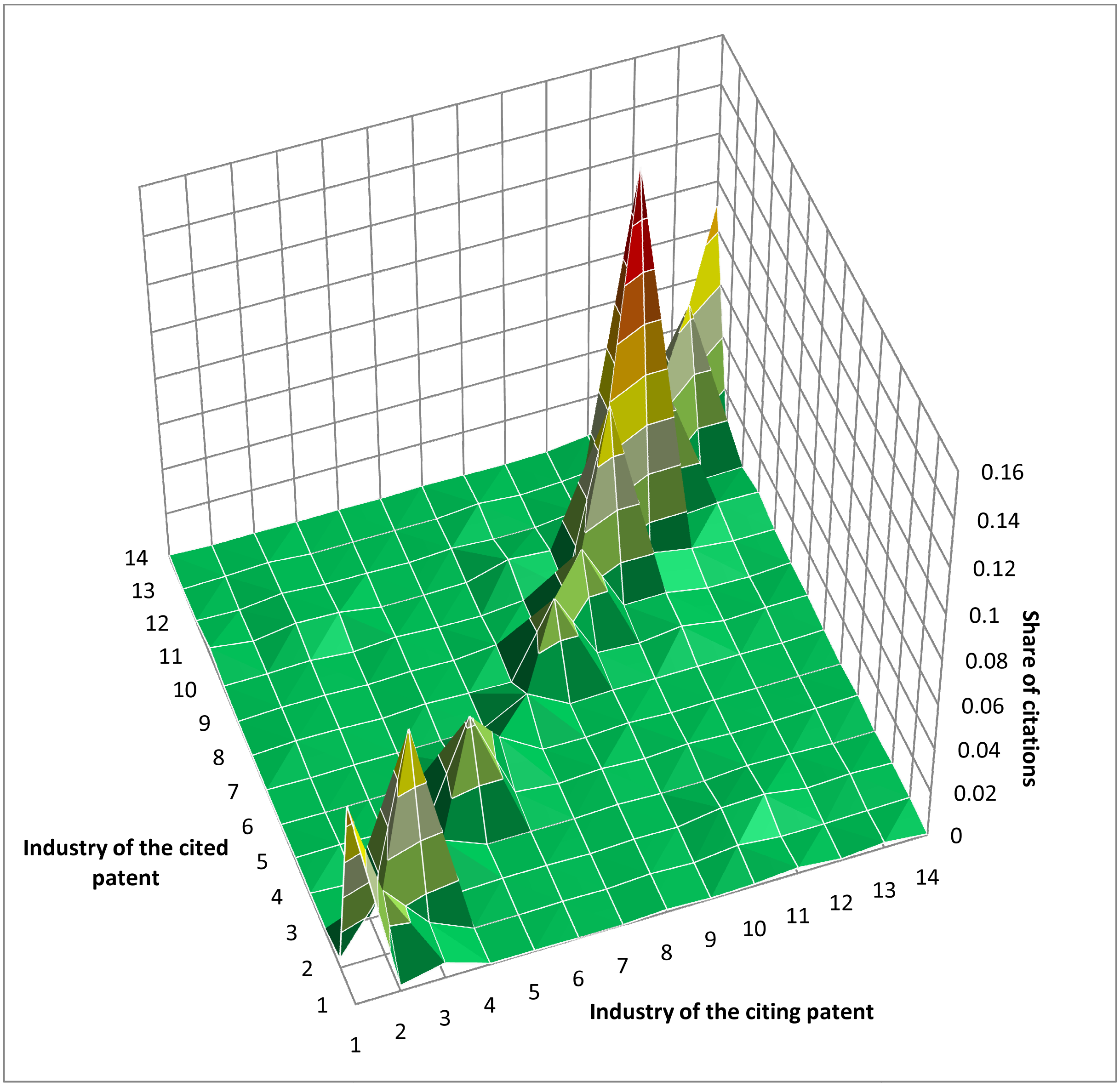

Figure 4.13 - Industry-citation map for dataset IN_PER2

Finally, the industry-citation map for dataset OUT_PER1 is depicted in Figure 4.14. The main shares are allotted to intra-industry citations of industry 13 (consumer goods), 11 and 14 (civil engineering). Although the percentage of combinations with no observations is much higher than for the previous datasets (74 of 196), the overall share of intra-industry citations is approximately the same (78.29%). The highest inter-industry citation level occurs for patents from industry 5 being cited by patents from industry 10 (1.00%). Other relatively high shares belong to citations from industry 2 to 3 (0.74%), industry 3 to 6 (0.66%) and industry 8 to 10 (0.66%). Concluding the descriptive part of the analysis, one can say that knowledge spillovers happen for the largest part within a certain industry. Nevertheless, there are certain citation connections especially between “proximate” industries, such as the industries concerned with machinery; the chemical and pharmaceutical industry; and the electrical, electronic and instruments industry.

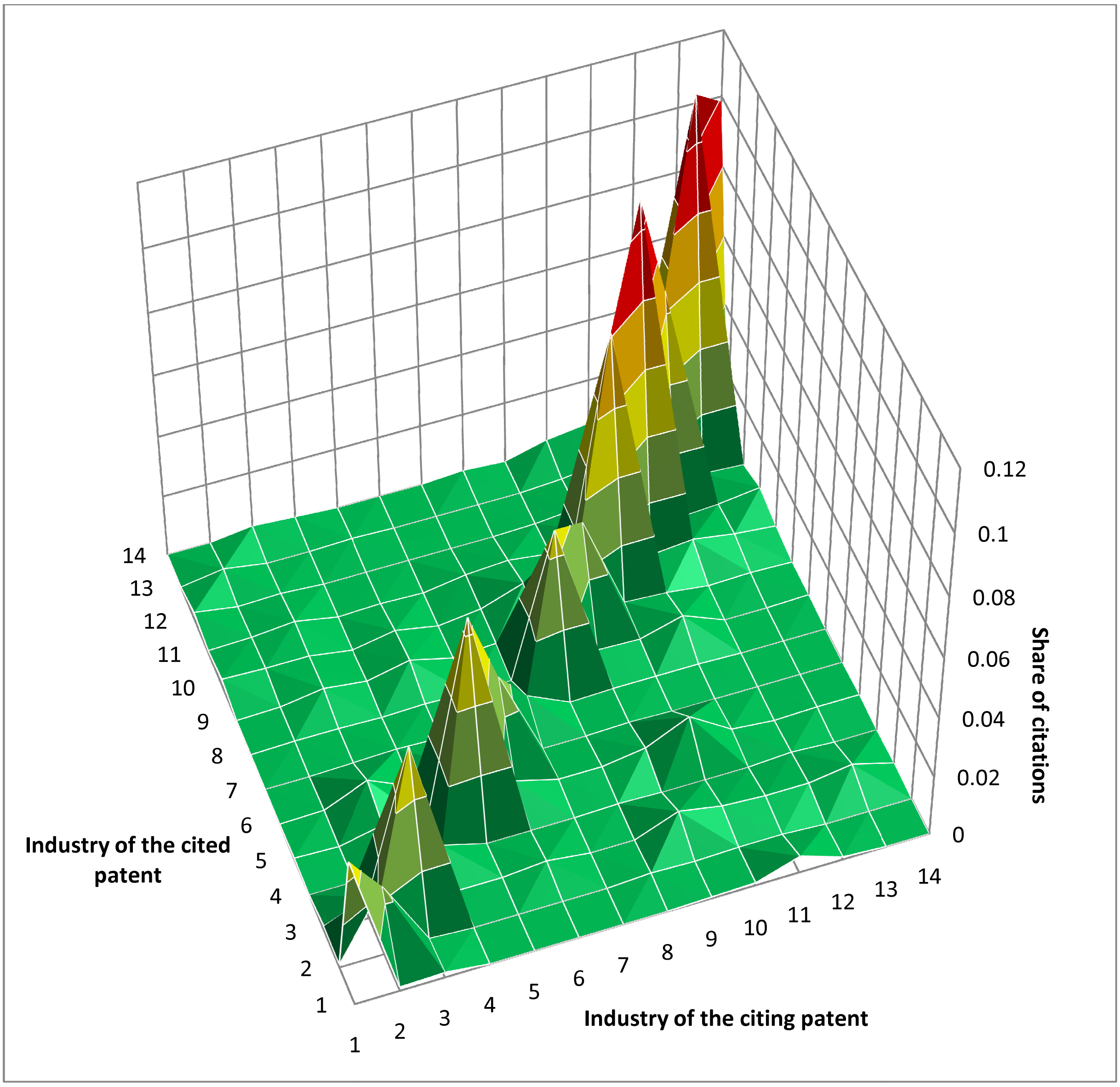

Figure 4.14 - Industry-citation map for dataset OUT_PER1

5 Econometric model and results

The previous chapter has provided the reader with a descriptive overview of the data in order to gain an understanding for the econometric analyses. This chapter will present the models used for the analyses as well as the results of the regressions. However, in order to conduct the chosen econometric analyses, the three datasets had to be altered slightly, as will be explained in detail in the following subsection.

5.1 The models

The focal points of the econometric analyses are patent citations since they are used as a proxy for knowledge spillovers. The goal of this thesis is to find out which parameters influence the probability of the occurrence of a patent citation (and thus a spillover of knowledge) and to which extent they do so. The intuition behind the used model is very simple: Either a given patent cites another given patent – or it does not. Hence, a binary response model is used because the dependent variable can either be equal to one (a citation occurs) or equal to zero (no citation occurs). As section 5.1.2 points out, the results for several binary response models are computed, including a probit, logit and complementary log-log model. At first, however, the datasets have to be adjusted in order to make them utilizable for the chosen models.

5.1.1 Adjustments of the datasets

The datasets described in section 4.3 consist of pairs of patents of a certain period between which a citation occurred. Thus, if one were to run a binary response model only with this data, the dependent variable would be equal to one for every observation. Lukach and Plasmans (2002) avoided this shortcoming by analyzing each industry separately and setting the dependent variable equal to one only for observations in which the citing patent was assigned to the selected industry. However, this procedure somehow lacks an appropriate interpretation. Thus, another path has been chosen in this thesis which requires an adjustment of the datasets.

The creation process is similar for all of the three datasets, IN_PER1, IN_PER2 and OUT_PER1. The starting point is the set of citation pairs in which each observation represents an actual citation consisting of a citing and a cited patent. In the left panel of Figure 5.1, the citing patents are denoted as P_i $(i = 1,..,k)$ and the cited patents are marked as Q_j $(j = 1,..,l)$. A pointer from a given patent P_i towards another patent Q_j indicates that patent P_i cites patent Q_j. A citing patent can issue multiple citations; similarly, a cited patent can receive multiple citations. In some cases, it is also possible that one and the same patent appears as a citing and as a cited patent in the same dataset.

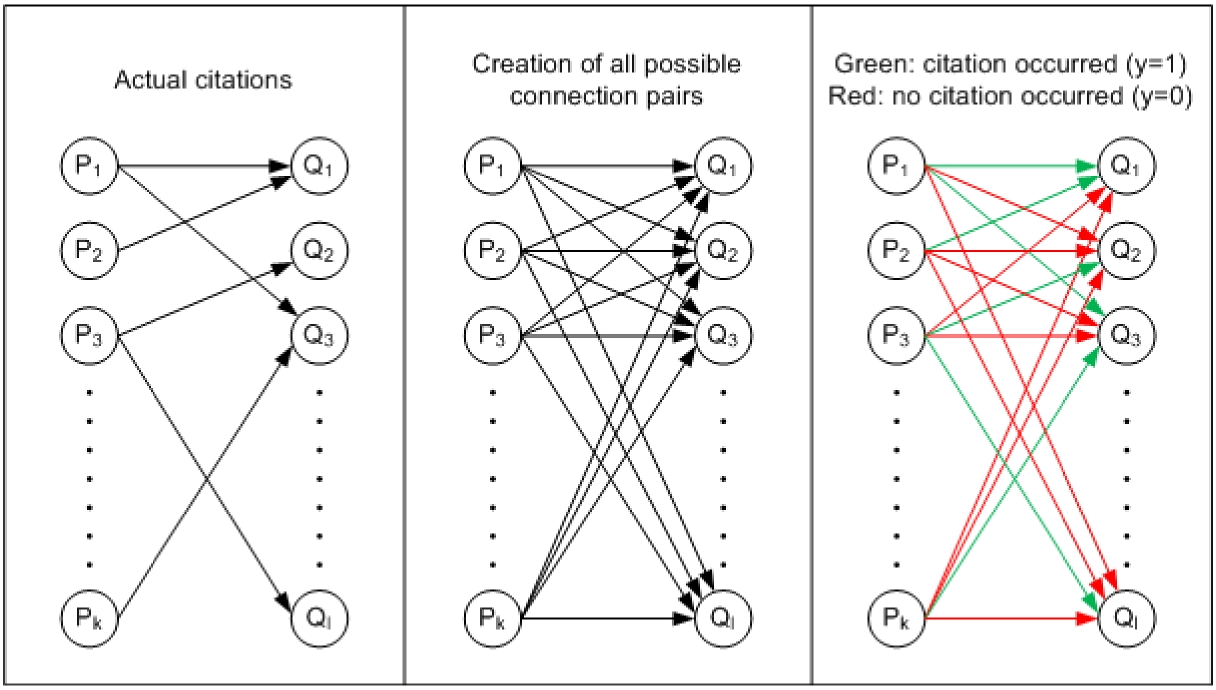

Figure 5.1 - Creation of the datasets for the regressions

The middle panel of Figure 5.1 depicts the next step in the creation of the datasets: All possible connection pairs between the group of citing patents and the group of cited patents are generated, thus producing the set of possible citations. Finally, the values of the dependent variable y (occurrence of a citation) are obtained by setting $y = 1$ whenever a connection pair is also part of the set of actual citations (indicated by the green pointers in the right panel of Figure 5.1, e.g. between P_3 and Q_2). Otherwise, y is set equal to zero, as signalized by the red pointers. For ease of reading, all patents that are elements of P_i will henceforth be referred to as “citing patents” regardless of whether a citation actually occurred in a given pair of patents. Similarly, all patents that are part of the set Q_j will be referred to as “cited patents”.

The motivation behind this procedure of dataset creation is straightforward: It is relatively simple to obtain a dataset consisting of patent citations (aside from the data cleaning process). However, it is not as easy to construct a comparable dataset of patent pairs between which no citation occurred out of the raw data from the PATSTAT database. It is simply not evident upon which criteria one should select a sample of “non-citation pairs” and how large this sample should be compared to the one of the citation pairs, since the probability that two random patents are citing each other is very low. Thus, we use the approach described above with the following interpretation, e.g. for the dataset constructed out of IN_PER1: Conditional on that an Austrian corporate patent has issued at least one citation during the first period, how is the probability that this patent cites another patent, which has been cited by at least one Austrian corporate patent, affected by several variables such as time and spatial proximity. Thus, we examine the determining factors of the probability of a patent citation for a sample of patents which have issued or received, respectively, at least one citation.

Due to the procedure of the creation of the datasets, it is also possible that combinations occur which are not feasible in reality. For instance, it can happen that a certain combination of patents has a negative time lag because the citing patent was applied for earlier than the cited patent. Although this is a valid possibility in reality when the search and examination phase of the application process of the citing patent requires a disproportional amount of time, it occurs only in a very small percentage of observations (e.g. of the actual citations in IN_PER1, only 5 of 10435 observations featured a negative time lag and were omitted from the dataset). However, in the construction of the datasets for the regression, many combinations with a negative time lag have occurred and were removed since they would have influenced the results significantly. Another set of unfeasible combinations appears when a patent is part of the group of citing patents as well as of the group of cited patents. Then, one of the created connections features one and the same patent on both ends of the citation pair. Since this is not a valid option, these connections were removed from the final datasets as well.

Table 5.1 - Basic statistics of the regression datasets

		REG_IN_PER1	REG_IN_PER2	REG_OUT_PER1
Number of citing patents		2638	1666	2025
Number of cited patents		8290	3369	1059
Number of observations		21083067	5590421	1939519
% of observations where an actual citation occurred		0.049%	0.064%	0.118%
Application year of the citing patent	mean	1995.620	2001.166	1998.146
	std. dev.	1.697	1.649	2.205
Time lag	mean	8.944	11.647	3.443
	std. dev.	5.802	6.866	2.411
% of intra-firm pairs		0.234%	0.111%	0.495%
% of intra-industry pairs		8.908%	9.996%	9.171%

Finally, in order to differentiate properly between the datasets of the actual citation pairs and the datasets created for the regressions, the following abbreviations will be used from this point forth:

- REG_IN_PER1: the regression dataset obtained from IN_PER1
- REG_IN_PER2: the regression dataset obtained from IN_PER2
- REG_OUT_PER1: the regression dataset obtained from OUT_PER1

Table 5.1 features a summary of the three new datasets. Due to the large number of citing and cited patents, the number of observations is very high for all three datasets, while the percentage of actual citations is rather low. This reflects the low probability that two random patents are part of a citation relationship. The large number of observations leads to some problems in the computation of the various regression models, as will be described in subsection 5.1.4.

5.1.2 Selection of the models

As mentioned before, the regression datasets are used to conduct several statistical analyses. Since the dependent variable is either equal to zero or equal to one, binary response models are utilized. More precisely, we use a probit, logit and complementary log-log model to examine the three datasets. The following explanations of these models are based on the books by Wooldridge (2005) and Long (1997).

5.1.2.1 Basic specification

Binary response models use the concept of a latent variable which is linked to the dependent variable via a function that results in values between zero and one. The purpose of binary response models is to calculate the probability that the dependent variable is equal to one (the "response probability"). The general form of binary response models is

$$\Pr(y = 1|x_1, \dots, x_k) = G(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k), \tag{5.1}$$

where y is the dependent variable, $x_1, \dots, x_k$ are the independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the regression coefficients and G is the linking function. Depending on the choice of the binary response model, the linking function G takes on different forms. In the probit model, the standard normal cumulative distribution function is used for G:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds, \tag{5.2}$$

where z can be any real number. In the logit model, on the other hand, G is the standard logistic cumulative distribution function:

$$G(z) = \Lambda(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{5.3}$$

Finally, in the complementary log-log model (sometimes also referred to as Weibull model), G is defined as:

$$G(z) = 1 - \exp[-\exp(z)] \tag{5.4}$$

The probit and logit model are symmetric, i.e. from the point on the probability curve where $\Pr(y = 1|x_1, \dots, x_k) = 0.5$, the magnitude of the impact on the probability caused by a change in x_i by a given amount δ is the same whether x_i is increased or decreased by δ. On the other hand, in the asymmetric complementary log-log model, the magnitude of the impact differs depending on whether the explanatory variable is increased or decreased. Due to the fact that the independent variables do not influence the probability directly but via a linking function, the regression coefficients $\beta_0, \beta_1, \dots, \beta_k$ do not represent the marginal impact of each variable on the probability of y being equal to one. Therefore, the influence of a change in one variable on the probability has to be determined otherwise. One possibility is to calculate the marginal effect of x_i by calculating the partial derivative of $G(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_k)$ with respect to x_i. For all models, the sign of the marginal effect of x_i is the same as the sign of the corresponding β_i. However, since only two of the explanatory variables are continuous and the other ones are dummies, this way of calculating the effect of a change in an independent variable on the probability is very limited in our case.

Thus, the second option is employed, which is to calculate the change in the probability for a discrete change of δ in x_i, holding all other variables at their mean. Possible values for δ include a unit change and a standard deviation change. For continuous independent variables, a centered standard deviation change is calculated around their respective means. For dummy variables, a change from zero to one is considered in the analyses. Furthermore, it is worth emphasizing that the objective of the regressions in this thesis is not to calculate an absolute value for the probability of the occurrence of a patent citation, but rather to find out more about the relative importance of each explanatory variable.

5.1.2.2 Goodness of fit

Another issue in the field of binary response models is the selection of an adequate measure of fit in order to evaluate the fit of the chosen model. As Long (1997) states, contrary to R^2 in the linear regression model, there is no generally accepted measure of fit for binary response models. Over the years, different approaches to calculate a measure of fit have been undertaken, which can be grouped in three categories:

- Pseudo-R^2's based on R^2 in the linear regression model
- Pseudo-R^2's using observed versus predicted values
- Information measures

The first category contains measures that transfer the various explanations of R^2 in the linear regression model, such as the percentage of explained variation and the likelihood ratio index, to general models. Measures of the second group count the number of correctly predicted values by the model. The third group contains measures of information, such as

Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). However, due to the diverse origin of the various measures of fit, it is not possible to make comparisons of the values of different measures of fit. Furthermore, it is important to remember that the information content of measures of fit is limited, as claimed by Long (1997, p. 102):

> *"[…] I am unaware of convincing evidence that selecting a model that maximizes the value of a given measure of fit results in a model that is optimal in any sense other than the model having a larger value of that measure."*

Despite these constraints, it might be useful to calculate the goodness of fit in order to find out if one of the three models is more appropriate for determining the probability of knowledge spillovers than the other ones. Thus, it was decided to calculate McFadden's R^2, which resembles the percentage of explained variation of the linear regression model using the log likelihood. R^2_{McF} is defined as

$$R^2_{McF} = 1 - \frac{\ln\left[\hat{L}\left(M_\beta\right)\right]}{\ln\left[\hat{L}\left(M_\alpha\right)\right]}, \tag{5.5}$$

where $\ln\left[\hat{L}(M_\alpha)\right]$ is the estimated log likelihood of model M_α, which includes only the intercept, and $\ln\left[\hat{L}\left(M_\beta\right)\right]$ is the estimated log likelihood of model M_β, which includes all of the variables of the respective regression. Thus, in order to calculate the chosen measure of fit, one has to calculate the log likelihood of each model. According to Long (1997), the likelihood of a binary response model is defined as

$$L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \prod_{y=1} G(\boldsymbol{x_n}\boldsymbol{\beta}) \prod_{y=0} [1 - G(\boldsymbol{x_n}\boldsymbol{\beta})], \tag{5.6}$$

where $L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})$ is the likelihood of the vector of coefficients, $\boldsymbol{\beta}$**,** dependent on the matrix of independent variables of the entire sample, $\boldsymbol{X}$, as well as on the vector of their respective dependent variables, $\boldsymbol{y}$, and $\boldsymbol{x_n}$ is a row vector of independent variables of the n-th observation. Furthermore, the index of the multiplication sign shall indicate that only those observations are used in each multiplication for which $y = 1$ and $y = 0$, respectively. Since the log likelihood is required for the calculation of R^2_{McF}, the logarithm of equation (5.6) is taken, yielding

$$\ln[L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})] = \sum_{y=1} \ln[G(\boldsymbol{x_n}\boldsymbol{\beta})] + \sum_{y=0} \ln[1 - G(\boldsymbol{x_n}\boldsymbol{\beta})]. \tag{5.7}$$

Equation (5.7) is used to estimate the log likelihood of both the complete model and the model without regressors for each of the binary response models and each dataset. Using the respective values, R^2_{McF} is calculated for each of the regressions.

5.1.3 Description of the variables

After shedding some light on the basic properties of the utilized models, this section focuses on the explanatory variables used in the regressions. Six factors have been chosen to explain the probability of the occurrence of a citation and thus of a knowledge spillover. In principle, the variables are the same for all three regression datasets.

5.1.3.1 Citing_appln_year

The variable *citing_appln_year* represents the year, in which the citing patent has been applied for. Although the exact date of the application was available, the year of the application was regarded as sufficient for the purpose of this thesis and thus facilitated the computation of the regressions. The range of this variable is restricted to six values for datasets REG_IN_PER1 (1993 – 1998) and REG_IN_PER2 (1999 – 2004), and to thirteen values for dataset REG_OUT_PER1 (1993 – 2005).

5.1.3.2 Time_lag

The variable *time_lag* measures the time lag of an observation, which is constructed as the difference between the year of application of the citing patent and the year of application of the cited patent. Observations in which the citing and the cited patent had been applied for within the same calendar year hold a time lag of zero. As mentioned before, even though there is the remote possibility of a negative time lag in reality, this option was not allowed for in the creation of the three regression datasets. Thus, *time_lag* is an integer greater or equal to zero. Furthermore, by containing the time lag, it was not possible to include the application year of the cited patent since this variable would have been a linear combination of the variables *citing_appln_year* and *time_lag*. The descriptive analysis suggests that the share of citations as a function of the time lag rises steeply at first until it reaches a peak and declines afterwards. In order to account for this behavior, *time_lag* enters the regressions linearly as well as quadratically.

5.1.3.3 Same_firm

The variable *same_firm* indicates whether an observation features a pair of patents that have been applied for by the same company. *Same_firm* is a dummy variable which is equal to one if both the citing and the cited patent have been applied for by the same firm; otherwise, it is equal to zero. This dummy shall indicate the influence of self citations on the probability of the occurrence of a citation. Due to the aforementioned restrictive definition of the boundaries of a firm (see section 4.2.2.2), a self citation can only take place if the applicant of the cited patent (for REG_IN_PER1 and REG_IN_PER2) or of the citing patent (for REG_OUT_PER1), respectively, is of Austrian origin.

5.1.3.4 Citing_industry

The dummy variable *citing_industry* represents the industry in which the citing patent is originated. The classification of the industries was carried out according to Table 4.2. Thus, there are fourteen different industries. *Citing_industry* shall capture the impact of the various industries on the probability of the occurrence of a citation. The base level of this variable is industry 1 (electrical engineering). Hence, the values for the other thirteen options will indicate the influence of the industry relative to the base industry.

5.1.3.5 Same_industry

The dummy variable *same_industry* indicates whether the citing and the cited patent of an observation belong to the same industry. The dummy is set equal to one if both the citing and the cited patent are attributed to the same industry; otherwise, it is set equal to zero. The purpose of this variable is to account for the influence of inter-industry citations.

5.1.3.6 Citing_firm_country / Cited_firm_country

Finally, the last variable is concerned with the geographical origin of the respective counterpart to the Austrian firm patent. For datasets REG_IN_PER1 and REG_IN_PER2, *cited_firm_country* indicates the country of origin of the applicant of the cited patent. For dataset REG_OUT_PER1, *citing_firm_country* specifies the country of origin of the applicant of the citing patent. Due to the large variety of countries in the datasets, they were aggregated on the six levels presented in section 4.3.4.2: Austria; Germany and Switzerland; other European countries; Japan; USA and Canada; and all other countries. This segmentation allows us to make some conclusions about the impact of geographical origin on the probability of a patent citation. The base level of this variable is Austria, thus only the other five options will appear in the results.

5.1.4 Computational issues

The regressions were computed with the statistical software *R* (R Development Core Team, 2010). *R* is a free, command line based program which covers a wide variety of statistical applications due to the extensibility of the available functions via a large number of additional packages. For instance, in order to connect *R* to the *MySQL* database for accessing the datasets, the package “RMySQL” was used.

As apparent from Table 5.1, the number of observations in each dataset is extremely large. Especially for REG_IN_PER1 and REG_IN_PER2 a direct computation of the regression was not feasible due to the limited availability of working memory. Hence, the computation of the regression had to be split up into several subsets. Therefore, a unique number was assigned to each observation in a dataset. Then, using the "sample" function in *R*, a subset of random observations was extracted from the dataset. Due to the fact that in the vast majority of observations the dependent variable is equal to zero, the size of the subsets had to be very large in order to contain an adequate number of actual citations. In fact, the size of the subsets was chosen as large as possible, resulting in subsets with more than 900000 observations for each of the three datasets. For a smaller size, it occurred that some of the industry dummies (especially those that appear only in few observations) featured an obviously incorrect value for the standard error (about two orders of magnitude larger than for other industry dummies). For each subset, a regression was computed and its results were stored in a separate matrix. This procedure was iterated until each observation from the original dataset had been used exactly once in a subset (except for a few remaining observations; see the explanation below).

Thus, the computation of a regression of a dataset was split up into a series of regressions of subsets. However, the results of these "subregressions" had to be put together again in order to provide a result for the entire dataset. So as to yield the estimates of the coefficients of the overall regression, the arithmetic mean of each of the estimates of the coefficients of the subregressions is calculated:

$$\hat{\beta}_{i,total} = \sum_{j=1}^{J} \left(\frac{1}{J} \ \hat{\beta}_{i,j} \right), \tag{5.8}$$

where $\hat{\beta}_{i,total}$ is the estimated coefficient of the i-th independent variable of the overall regression, J is the number of subregressions (iterations) and $\hat{\beta}_{i,j}$ is the estimated coefficient of the i-th independent variable in the j-th subregression. In order to calculate the standard error of the overall regression, we use the equation for the variance of sums of pairwise uncorrelated random variables (Wooldridge, 2005):

$$\mathrm{Var}\left(\sum_{j=1}^{J} a_j X_j \right) = \sum_{j=1}^{J} \left(a_j^2 \mathrm{Var}(X_j) \right) \tag{5.9}$$

The presumption that the results from each subregression are independent and thus uncorrelated is fulfilled due to the random selection of the subsamples. Therefore, the standard error for the coefficient of the i-th independent variable of the overall regression, $SE_{i,total}$, is calculated in the following way:

$$SE_{i,total} = \sqrt{\sum_{j=1}^{J}\left(\frac{1}{J^2}SE_{i,j}^2\right)} = \frac{1}{J}\sqrt{\sum_{j=1}^{J} SE_{i,j}^2}, \tag{5.10}$$

where $SE_{i,j}$ is the standard error of the i-th independent variable of the j-th subregression. In order to use equation (5.10) for the determination of the standard errors of the overall regression, the sample size of the subsets of data has to be constant. Since the ratio of total number of observations and number of iterations is in no case an integer, the next smaller integer is chosen as sample size. Thus, it occurs that a few observations are not used in the subregressions. However, since the number of unused observations is limited to $J - 1$, which is very small compared to the total number of observations as well as to the number of observations in a subset, there is only a very small error in this approach.

Using the values of the estimation of the coefficients as well as their standard errors, the t-values were calculated in order to gain an insight into the significance of the parameters. Furthermore, for the smallest of the three datasets, REG_OUT_PER1, it was also possible to compute the regression of the entire dataset at once (by using a more powerful hardware specification). The results of this computation can be found in Appendix A. By comparing the results from the regression of the entire dataset and the results from the iterative approach, one finds that equations (5.8) and (5.10) provide a very good approximation of the values of the overall regression.

5.2 Results

In this section, the results from the regressions of the three datasets REG_IN_PER1, REG_IN_PER2 and REG_OUT_PER1 will be presented. As mentioned in section 5.1.2, the discrete change for the variables *citing_appln_year* and *time_lag* is calculated as a centered standard deviation change. All other variables are dummies and thus the discrete change is calculated by setting the dummy from zero to one while holding all other variables at their means.

5.2.1 REG_IN_PER1

The results of the probit, logit and complementary log-log analysis of dataset REG_IN_PER1 can be found in Table 5.2, Table 5.3 and Table 5.4, respectively. At first sight, one notices that the variables have only a very small absolute influence on the overall probability of a patent citation and thus a knowledge spillover. This is due to the nature of the dataset which features a vast majority of observations in which the dependent variable is equal to zero. Furthermore, we see that the significance of the parameters is similar in all three models. Only for a few variables there are minor differences in the level of significance of the probit, logit and complementary log-log model. Furthermore, R^2_{McF} in the probit model is slightly higher compared to the values of the logit and complementary log-log model, suggesting that the probit model is somewhat more appropriate for the given set of data.

The coefficient of *citing_appln_year* is not significantly different from zero. Therefore, this parameter does not have an influence on the probability of a patent citation. Furthermore, while both *time_lag* terms are significant, the influence of a centered standard deviation change is approximately equal to zero. However, the signs of the two variables show the expected behavior: For small values of the time lag, the positive linear term prevails and increases the probability. For larger time lags, the negative influence of the quadratic term reduces the probability of a knowledge spillover.

The largest impact on the probability of the occurrence of a patent citation is obtained by the variables *same_firm* and *same_industry*. Both parameters are highly significant in all three regressions. While the impact of the dummy indicating whether two patents of a pair are part of the same industry is approximately constant for all three models, the *same_firm* dummy shows a quite different behavior for the probit model compared to both the logit and the complementary log-log model. For the latter two models, the influence of the patents being from the same firm on the probability of a knowledge spillover is almost twice as high as the influence of the *same_industry* dummy. For the probit model, on the other hand, the influence of the variable *same_firm* is more than thrice as high as the impact of the *same_industry* dummy and almost twice as high as the impact of the *same_firm* dummy in the other two models. In the probit model, the absolute value of the discrete change indicates an absolute change in the probability of more than 1%.

The dummies indicating the industry of the citing patent require special attention. Except for industries 3 (instruments) and 12 (transport), all of the coefficients are significantly different from zero. There are slight variations in the values of the discrete changes between the three models, especially between the probit model and the other two models. The interesting aspect of the industry variables are the signs of the coefficients: As mentioned in the description of the variables, the coefficients of the regression reflect the impact of the various industries relative to the base category, industry 1 (electrical engineering). Thus, relative to industry 1, the biggest positive influence on the probability originates from industries 2 (electronics), 4 (other machinery) and 7 (agricultural chemistry and machinery). While the probit model suggests that industry 2 has the largest positive impact on the probability, this spot is taken by industry 4 in the other two models. Interestingly, industries 4 and 7 are the ones with the lowest share of actual citations as indicated in the descriptive analysis in section 4.3.5.1.

Table 5.2 - Result of the probit analysis of dataset REG_IN_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	-3.2060	3.886	————	-0.83	
citing_appln_year	-0.0004	0.002	0.00000	-0.20	
time_lag	0.0040	0.002	0.00000	2.15	*
(time_lag)^2	-0.0003	0.000		-3.35	***
same_industry	1.0244	0.007	0.00354	147.55	***
same_firm	1.3716	0.021	0.01214	65.50	***
citing_industry					
2 (electronics)	0.1497	0.022	0.00013	6.80	***
3 (instruments)	0.0021	0.018	0.00000	0.12	
4 (other machinery)	0.1458	0.040	0.00013	3.68	***
5 (chemistry)	-0.0432	0.018	-0.00003	-2.41	*
6 (pharmaceuticals and biotechnology)	0.1154	0.020	0.00010	5.91	***
7 (agricultural chemistry and machinery)	0.1315	0.038	0.00011	3.47	***
8 (material engineering)	-0.0454	0.019	-0.00003	-2.40	*
9 (chemical and environmental engineering)	0.0587	0.020	0.00004	2.98	**
10 (materials processing and handling)	-0.0554	0.017	-0.00003	-3.26	**
11 (mechanical engineering)	-0.1696	0.017	-0.00009	-9.80	***
12 (transport)	-0.0223	0.020	-0.00001	-1.10	
13 (consumer goods)	-0.0613	0.019	-0.00004	-3.31	***
14 (civil engineering)	-0.1909	0.019	-0.00009	-10.07	***
cited_firm_country					
DE & CH	0.3477	0.018	0.00013	19.27	***
Europe	0.3590	0.019	0.00014	19.36	***
JP	0.3344	0.019	0.00012	17.84	***
Other	0.3776	0.026	0.00015	14.32	***
US & CA	0.3557	0.017	0.00013	20.57	***
Size of dataset	21083067				
Size of subset per iteration	958321				
Number of iterations	22				
Number of unused observations	5				
R^2_{McF}	0.1903				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
cited_firm_country	AT				
	***	0.001			
Significance levels	**	0.01			
	*	0.05			

Table 5.3 - Result of the logit analysis of dataset REG_IN_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	-10.0636	12.185	———	-0.83	
citing_appln_year	0.0000	0.006	0.00000	0.01	
time_lag	0.0137	0.006	0.00000	2.32	*
(time_lag)^2	-0.0010	0.000		-3.65	***
same_industry	3.4803	0.024	0.00348	142.63	***
same_firm	3.8092	0.065	0.00657	58.41	***
citing_industry					
2 (electronics)	0.4867	0.070	0.00011	6.95	***
3 (instruments)	-0.0605	0.057	-0.00001	-1.06	
4 (other machinery)	0.6135	0.131	0.00015	4.69	***
5 (chemistry)	-0.1825	0.057	-0.00003	-3.20	**
6 (pharmaceuticals and biotechnology)	0.2976	0.062	0.00006	4.84	***
7 (agricultural chemistry and machinery)	0.5211	0.122	0.00012	4.26	***
8 (material engineering)	-0.1945	0.060	-0.00003	-3.23	**
9 (chemical and environmental engineering)	0.1610	0.062	0.00003	2.58	**
10 (materials processing and handling)	-0.2321	0.054	-0.00004	-4.29	***
11 (mechanical engineering)	-0.5941	0.055	-0.00008	-10.75	***
12 (transport)	-0.0694	0.065	-0.00001	-1.07	
13 (consumer goods)	-0.2649	0.058	-0.00004	-4.53	***
14 (civil engineering)	-0.6659	0.059	-0.00009	-11.29	***
cited_firm_country					
DE & CH	1.1578	0.062	0.00012	18.67	***
Europe	1.1865	0.064	0.00012	18.68	***
JP	1.0914	0.064	0.00011	16.98	***
Other	1.2614	0.087	0.00014	14.43	***
US & CA	1.1574	0.060	0.00012	19.36	***
Size of dataset	21083067				
Size of subset per iteration	958321				
Number of iterations	22				
Number of unused observations	5				
R^2_{McF}	0.1891				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
cited_firm_country	AT				
Significance levels	***	0.001			
	**	0.01			
	*	0.05			

Table 5.4 - Result of the complementary log-log analysis of dataset REG_IN_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	-10.0437	12.133	———	-0.83	
citing_appln_year	0.0000	0.006	0.00000	0.00	
time_lag	0.0138	0.006	0.00000	2.35	*
(time_lag)^2	-0.0010	0.000		-3.69	***
same_industry	3.4773	0.024	0.00348	142.64	***
same_firm	3.7805	0.065	0.00641	58.29	***
citing_industry					
2 (electronics)	0.4860	0.070	0.00011	6.98	***
3 (instruments)	-0.0617	0.057	-0.00001	-1.09	
4 (other machinery)	0.6129	0.130	0.00015	4.70	***
5 (chemistry)	-0.1836	0.057	-0.00003	-3.23	**
6 (pharmaceuticals and biotechnology)	0.2870	0.061	0.00006	4.69	***
7 (agricultural chemistry and machinery)	0.5186	0.121	0.00012	4.27	***
8 (material engineering)	-0.1965	0.060	-0.00003	-3.27	**
9 (chemical and environmental engineering)	0.1593	0.062	0.00003	2.56	*
10 (materials processing and handling)	-0.2319	0.054	-0.00004	-4.31	***
11 (mechanical engineering)	-0.5941	0.055	-0.00008	-10.79	***
12 (transport)	-0.0694	0.064	-0.00001	-1.08	
13 (consumer goods)	-0.2678	0.058	-0.00004	-4.61	***
14 (civil engineering)	-0.6622	0.059	-0.00009	-11.30	***
cited_firm_country					
DE & CH	1.1569	0.062	0.00012	18.67	***
Europe	1.1857	0.063	0.00012	18.68	***
JP	1.0904	0.064	0.00011	16.98	***
Other	1.2603	0.087	0.00014	14.43	***
US & CA	1.1565	0.060	0.00012	19.35	***
Size of dataset	21083067				
Size of subset per iteration	958321				
Number of iterations	22				
Number of unused observations	5				
R^2_{McF}	0.1891				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
cited_firm_country	AT				
	***	0.001			
Significance levels	**	0.01			
	*	0.05			

In contrast, industries 14 (civil engineering) and 11 (mechanical engineering) constitute the major negative impact on the dependent variable, despite each having one of the highest shares of patent citations according to the descriptive analysis. Industry 10 (materials processing and handling), which has the highest share of actual citations, also has a negative influence on the probability of a patent citation. However, for this industry, the change in the probability is less than half as large as for the aforementioned industries 11 and 14. Furthermore, one has also to consider that some information regarding the dependency of a citation on the industry might be contained in the *same_industry* dummy. Computing the regressions without the *same_industry* dummy results in a change of the sign of some of the *citing_industry* coefficients as well as in more coefficients not being significantly different from zero (see Appendix B.1 for the exact results). Regardless of the industry, there is always a positive influence of the probability of a citation if both patents of a pair belong to the same industry due to the magnitude of the *same_industry* variable.

Finally, the dummies indicating the country of origin of the applicant of the cited patent are significant for all three analyses. However, the impact of these dummies in the probit model is slightly higher than in the logit and complementary log-log model. One can see that all countries have a positive influence on the probability of a knowledge spillover compared to the base category, Austria. However, for the interpretation of this result one also has to take the variable *same_firm* into account. Since the *same_firm* dummy can only be equal to one if the cited patent originates from Austria (which is, in fact, the case in the majority of cited Austrian patents), some information regarding the geographical influence may be contained in the *same_firm* dummy. Thus, if the cited patent is held by the same firm as the citing one (and is, therefore, an Austrian patent), the probability of a patent citation is higher than if the cited patent is held by any non-Austrian entity. Austrian inter-firm citations, on the other hand, have a lower probability compared to international citations. If one omits the *same_firm* dummy from the regression, the coefficients of the international country dummies obtain a negative sign (see Appendix B.1).

Similar to the above mentioned influence of the affiliation of the citing patent to a certain industry, we can observe that the category “other countries” seems to have the largest influence on the probability in all three models, despite having the smallest share of citations in the descriptive analysis. The probability of a knowledge spillover between a random Austrian and a random Japanese patent is the smallest of all international country combinations. Moreover, since the coefficient for Germany and Switzerland is not significantly higher than the values of the other country dummies, the results do not indicate a special influence of spatial or linguistic proximity among the applicants of the citing and the cited patent.

5.2.2 REG_IN_PER2

The regression results for the second dataset, REG_IN_PER2, are summarized in Table 5.5, Table 5.6 and Table 5.7. For this dataset, all three regressions feature almost the same significance levels for all coefficients. Contrary to the first dataset, the logit and the complementary log-log model show a slightly higher R^2_{McF} compared to the probit model. The magnitude of R^2_{McF} of all three models is just below the one of the regressions of the first dataset.

The coefficients of the time variables show the exact opposite in terms of significance compared to the results of the first period: While the coefficients of both time lag terms are not significantly different from zero, the coefficient for *citing_appln_year* has a high significance level. The influence of the application year is negative. According to the regressions, a later application date leads to a decreased probability of a patent citation. However, the impact of a centered standard deviation change in this variable is rather small compared to the impact of discrete changes in other factors.

Just like in the regressions for dataset REG_IN_PER1, the variables with the largest impact on the probability for REG_IN_PER2 are the dummies *same_firm* and *same_industry*. Both coefficients suggest a positive influence on the probability if the citing and the cited patent are owned by the same firm or part of the same industry, respectively. While setting the *same_industry* dummy from zero to one has approximately the same impact in the probit, logit and complementary log-log model, this is not the case for the *same_firm* dummy. In the probit model, setting the *same_firm* dummy from zero to one increases the probability by approximately 1.29%. In the logit and complementary log-log model, the same discrete change results in an increased probability of about 0.6%, thus suggesting an impact only half as large.

The dummies indicating the industry affiliation of the citing patent are significant for all but one industry (industry 3 – instruments). The values of a discrete change in one variable are on roughly the same level in all three models, although there are again some differences between the probit regression and the other two models. The industries with the largest positive impact on the probability relative to the base category are industry 4 (other machinery), 7 (agricultural chemistry and machinery) as well as 6 (pharmaceuticals and biotechnology). Once again, it is apparent that these are the industries with the lowest share of both patents and citations in the descriptive analysis. On the other hand, there are only few industries with a negative coefficient. The ones resulting in the largest decrease in the probability are industry 11 (mechanical engineering) and 14 (civil engineering). However, their influence with respect to the base category is rather small.

Table 5.5 - Result of the probit analysis of dataset REG_IN_PER2

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	54.3926	6.988	————	7.78	***
citing_appln_year	-0.0291	0.003	-0.00003	-8.34	***
time_lag	-0.0013	0.003	0.00000	-0.40	
(time_lag)^2	0.0000	0.000		0.29	
same_industry	1.0888	0.012	0.00480	87.45	***
same_firm	1.3384	0.043	0.01287	30.98	***
citing_industry					
2 (electronics)	0.1106	0.032	0.00010	3.44	***
3 (instruments)	-0.0135	0.024	-0.00001	-0.57	
4 (other machinery)	0.3618	0.054	0.00050	6.64	***
5 (chemistry)	0.1089	0.030	0.00009	3.69	***
6 (pharmaceuticals and biotechnology)	0.2845	0.048	0.00034	5.91	***
7 (agricultural chemistry and machinery)	0.2746	0.049	0.00032	5.57	***
8 (material engineering)	0.1205	0.030	0.00011	4.01	***
9 (chemical and environmental engineering)	0.1316	0.031	0.00012	4.28	***
10 (materials processing and handling)	-0.0514	0.024	-0.00003	-2.13	*
11 (mechanical engineering)	-0.1396	0.021	-0.00008	-6.65	***
12 (transport)	0.0685	0.027	0.00005	2.50	*
13 (consumer goods)	0.0596	0.026	0.00005	2.31	*
14 (civil engineering)	-0.1224	0.026	-0.00007	-4.76	***
cited_firm_country					
DE & CH	0.2956	0.032	0.00014	9.30	***
Europe	0.2931	0.033	0.00014	8.91	***
JP	0.2998	0.033	0.00014	9.15	***
Other	0.2892	0.048	0.00013	6.08	***
US & CA	0.3004	0.030	0.00014	9.91	***
Size of dataset	5590421				
Size of subset per iteration	931736				
Number of iterations	6				
Number of unused observations	5				
R^2_{McF}	0.1826				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
cited_firm_country	AT				
	***	0.001			
Significance levels	**	0.01			
	*	0.05			

Table 5.6 - Result of the logit analysis of dataset REG_IN_PER2

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	175.2978	21.629	———	8.10	***
citing_appln_year	-0.0925	0.011	-0.00003	-8.55	***
time_lag	-0.0060	0.010	0.00000	-0.63	
(time_lag)^2	0.0001	0.000		0.41	
same_industry	3.6674	0.043	0.00475	84.67	***
same_firm	3.5574	0.120	0.00608	29.55	***
citing_industry					
2 (electronics)	0.3587	0.100	0.00008	3.59	***
3 (instruments)	-0.0649	0.072	-0.00001	-0.90	
4 (other machinery)	1.4016	0.161	0.00055	8.72	***
5 (chemistry)	0.3705	0.090	0.00008	4.12	***
6 (pharmaceuticals and biotechnology)	0.9982	0.166	0.00031	6.01	***
7 (agricultural chemistry and machinery)	1.0814	0.154	0.00035	7.01	***
8 (material engineering)	0.3916	0.093	0.00009	4.20	***
9 (chemical and environmental engineering)	0.4210	0.095	0.00009	4.42	***
10 (materials processing and handling)	-0.1589	0.075	-0.00003	-2.13	*
11 (mechanical engineering)	-0.4535	0.065	-0.00007	-7.02	***
12 (transport)	0.2441	0.084	0.00005	2.92	**
13 (consumer goods)	0.1765	0.079	0.00003	2.23	*
14 (civil engineering)	-0.3978	0.078	-0.00006	-5.08	***
cited_firm_country					
DE & CH	0.9251	0.104	0.00012	8.89	***
Europe	0.9168	0.107	0.00012	8.54	***
JP	0.9182	0.107	0.00012	8.57	***
Other	0.8575	0.157	0.00011	5.47	***
US & CA	0.9355	0.100	0.00012	9.39	***
Size of dataset	5590421				
Size of subset per iteration	931736				
Number of iterations	6				
Number of unused observations	5				
R^2_{McF}	0.1835				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
cited_firm_country	AT				
	***	0.001			
Significance levels	**	0.01			
	*	0.05			

Table 5.7 - Result of the complementary log-log analysis of dataset REG_IN_PER2

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	174.9862	21.538	———	8.12	***
citing_appln_year	-0.0923	0.011	-0.00003	-8.58	***
time_lag	-0.0061	0.010	0.00000	-0.64	
(time_lag)^2	0.0001	0.000		0.42	
same_industry	3.6643	0.043	0.00474	84.74	***
same_firm	3.5272	0.119	0.00592	29.67	***
citing_industry					
2 (electronics)	0.3576	0.100	0.00008	3.59	***
3 (instruments)	-0.0645	0.072	-0.00001	-0.89	
4 (other machinery)	1.3989	0.160	0.00055	8.75	***
5 (chemistry)	0.3676	0.089	0.00008	4.11	***
6 (pharmaceuticals and biotechnology)	1.0008	0.165	0.00031	6.06	***
7 (agricultural chemistry and machinery)	1.0862	0.153	0.00035	7.09	***
8 (material engineering)	0.3899	0.093	0.00009	4.20	***
9 (chemical and environmental engineering)	0.4198	0.095	0.00009	4.43	***
10 (materials processing and handling)	-0.1573	0.074	-0.00003	-2.12	*
11 (mechanical engineering)	-0.4519	0.064	-0.00007	-7.01	***
12 (transport)	0.2446	0.083	0.00005	2.94	**
13 (consumer goods)	0.1735	0.079	0.00003	2.21	*
14 (civil engineering)	-0.3979	0.078	-0.00006	-5.11	***
cited_firm_country					
DE & CH	0.9235	0.104	0.00012	8.88	***
Europe	0.9152	0.107	0.00012	8.54	***
JP	0.9163	0.107	0.00012	8.56	***
Other	0.8557	0.157	0.00010	5.47	***
US & CA	0.9337	0.100	0.00012	9.38	***
Size of dataset	5590421				
Size of subset per iteration	931736				
Number of iterations	6				
Number of unused observations	5				
R^2_{McF}	0.1835				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
cited_firm_country	AT				
	***	0.001			
Significance levels	**	0.01			
	*	0.05			

While the coefficients of the country dummies are highly significant in all regressions, one can observe that the values for a discrete change in the variables in the probit model are consistently higher than the values in the logit and the complementary log-log model. Furthermore, one again has to take the *same_firm* dummy into account when interpreting the results for cited Austrian firms (the results from alternative model specifications can be found in Appendix B.2). It turns out that Austrian intra-firm citations have the highest probability, while Austrian inter-firm citations have the lowest one. Moreover, we find that the impact of a discrete change is approximately the same for all of the five groups of countries. The group of other countries has a slightly smaller positive impact on the probability compared to the other categories. It seems, however, that the only big geographical influence on the probability is the differentiation between Austrian and non-Austrian applicants.

5.2.3 REG_OUT_PER1

As mentioned before, for dataset REG_OUT_PER1, which is the smallest of the three sets in terms of the number of observations, it was possible to compute the regression directly. Nevertheless, it was decided to compute the regression iteratively as well in order to find out more about the accuracy of the calculation of the coefficients and their standard errors. Furthermore, the consistent use of this approach allows a comparison with the results from the other datasets. Table 5.8, Table 5.9 and Table 5.10 present the iterative regression results for dataset REG_OUT_PER1 for the probit, logit and complementary log-log model, respectively. The results of the regressions computed non-iteratively can be found in Appendix A.

Once again, we find that the significance levels of the regression coefficients are similar in the three models, although some parameters are more significant in the logit and complementary log-log model compared to the probit model. Furthermore, we find that R^2_{McF} in the probit model is higher than in the other two models. Moreover, the level of R^2_{McF} in the regressions of this dataset is the highest of the three datasets, suggesting that the chosen models provide a better fit for dataset REG_OUT_PER1. Contrary to the other two datasets, the coefficients of all time variables are significant. In the logit and complementary log-log model, the probability changes for a centered standard deviation change of the time variables are below the values in the probit model. The signs of the *time_lag* coefficients reflect the same situation as in the results of the first dataset: For small values, the larger positive coefficient of the linear term prevails, while for higher values a decrease in the probability is suggested due to the negative sign of the quadratic term. Holding all other variables at their mean, a centered standard deviation change of the time lag has, in contrast to *citing_appln_year*, a positive effect on the probability of a knowledge spillover.

Table 5.8 - Result of the probit analysis of dataset REG_OUT_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	29.5532	12.266	————	2.41	*
citing_appln_year	-0.0168	0.006	-0.00005	-2.73	**
time_lag	0.1249	0.011	0.00007	11.19	***
(time_lag)^2	-0.0148	0.001		-10.81	***
same_industry	1.1101	0.017	0.00900	66.86	***
same_firm	1.4276	0.038	0.02566	37.96	***
citing_industry					
2 (electronics)	0.1624	0.047	0.00035	3.47	***
3 (instruments)	-0.0044	0.040	-0.00001	-0.11	
4 (other machinery)	0.2801	0.091	0.00075	3.08	**
5 (chemistry)	-0.0225	0.040	-0.00004	-0.56	
6 (pharmaceuticals and biotechnology)	0.0983	0.051	0.00019	1.92	
7 (agricultural chemistry and machinery)	0.2548	0.072	0.00065	3.54	***
8 (material engineering)	-0.0171	0.044	-0.00003	-0.39	
9 (chemical and environmental engineering)	0.0850	0.044	0.00016	1.91	
10 (materials processing and handling)	-0.0870	0.038	-0.00012	-2.31	*
11 (mechanical engineering)	-0.1491	0.036	-0.00019	-4.15	***
12 (transport)	-0.0222	0.040	-0.00004	-0.56	
13 (consumer goods)	-0.0719	0.038	-0.00011	-1.92	
14 (civil engineering)	-0.2487	0.040	-0.00028	-6.16	***
citing_firm_country					
DE & CH	0.4264	0.033	0.00044	12.86	***
Europe	0.4194	0.035	0.00042	11.94	***
JP	0.4018	0.037	0.00039	11.00	***
Other	0.4655	0.050	0.00052	9.30	***
US & CA	0.4185	0.032	0.00042	13.06	***
Size of dataset	1939519				
Size of subset per iteration	969759				
Number of iterations	2				
Number of unused observations	1				
R^2_{McF}	0.2345				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
citing_firm_country	AT				
	***	0.001			
Significance levels	**	0.01			
	*	0.05			

Table 5.9 - Result of the logit analysis of dataset REG_OUT_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	65.1306	34.774	————	1.87	
citing_appln_year	-0.0374	0.017	-0.00003	-2.15	*
time_lag	0.3554	0.032	0.00005	10.98	***
(time_lag)^2	-0.0438	0.004		-10.86	***
same_industry	3.4685	0.054	0.00871	64.45	***
same_firm	3.6740	0.108	0.01444	34.14	***
citing_industry					
2 (electronics)	0.5008	0.138	0.00031	3.63	***
3 (instruments)	-0.0598	0.116	-0.00003	-0.52	
4 (other machinery)	0.9793	0.276	0.00078	3.55	***
5 (chemistry)	-0.1165	0.117	-0.00005	-1.00	
6 (pharmaceuticals and biotechnology)	0.1754	0.144	0.00009	1.22	
7 (agricultural chemistry and machinery)	0.8817	0.209	0.00067	4.23	***
8 (material engineering)	-0.1719	0.125	-0.00007	-1.37	
9 (chemical and environmental engineering)	0.2013	0.128	0.00011	1.57	
10 (materials processing and handling)	-0.3306	0.108	-0.00013	-3.07	**
11 (mechanical engineering)	-0.4900	0.103	-0.00018	-4.75	***
12 (transport)	-0.0794	0.114	-0.00004	-0.70	
13 (consumer goods)	-0.3027	0.107	-0.00012	-2.83	**
14 (civil engineering)	-0.7654	0.113	-0.00025	-6.80	***
citing_firm_country					
DE & CH	1.3087	0.105	0.00039	12.45	***
Europe	1.2957	0.111	0.00038	11.72	***
JP	1.2145	0.115	0.00034	10.57	***
Other	1.4210	0.153	0.00045	9.31	***
US & CA	1.2671	0.102	0.00036	12.40	***
Size of dataset	1939519				
Size of subset per iteration	969759				
Number of iterations	2				
Number of unused observations	1				
R^2_{McF}	0.2321				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
citing_firm_country	AT				
Significance levels	***	0.001			
	**	0.01			
	*	0.05			

Table 5.10 - Result of the complementary log-log analysis of dataset REG_OUT_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	61.8067	34.353	———	1.80	
citing_appln_year	-0.0358	0.017	-0.00003	-2.08	*
time_lag	0.3521	0.032	0.00004	10.96	***
(time_lag)^2	-0.0438	0.004		-10.88	***
same_industry	3.4622	0.054	0.00873	64.41	***
same_firm	3.6242	0.107	0.01389	33.97	***
citing_industry					
2 (electronics)	0.5131	0.136	0.00032	3.76	***
3 (instruments)	-0.0564	0.115	-0.00003	-0.49	
4 (other machinery)	1.0139	0.267	0.00083	3.80	***
5 (chemistry)	-0.0866	0.115	-0.00004	-0.75	
6 (pharmaceuticals and biotechnology)	0.1608	0.142	0.00008	1.13	
7 (agricultural chemistry and machinery)	0.8887	0.203	0.00067	4.38	***
8 (material engineering)	-0.1762	0.124	-0.00008	-1.43	
9 (chemical and environmental engineering)	0.2121	0.126	0.00011	1.68	
10 (materials processing and handling)	-0.3337	0.107	-0.00013	-3.11	**
11 (mechanical engineering)	-0.4803	0.103	-0.00018	-4.67	***
12 (transport)	-0.0735	0.113	-0.00003	-0.65	
13 (consumer goods)	-0.2911	0.106	-0.00012	-2.75	**
14 (civil engineering)	-0.7517	0.112	-0.00025	-6.74	***
citing_firm_country					
DE & CH	1.3089	0.105	0.00039	12.46	***
Europe	1.2936	0.110	0.00038	11.71	***
JP	1.2037	0.115	0.00033	10.46	***
Other	1.4235	0.152	0.00045	9.37	***
US & CA	1.2657	0.102	0.00037	12.38	***
Size of dataset	1939519				
Size of subset per iteration	969759				
Number of iterations	2				
Number of unused observations	1				
R^2_{McF}	0.2320				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
citing_firm_country	AT				
Significance levels	***	0.001			
	**	0.01			
	*	0.05			

The coefficient of the *same_industry* dummy predicts a relatively constant change in the probability for all three models. In fact, this parameter again constitutes the second largest influence on the dependent variable, increasing the probability by more than 0.9% when set from zero to one, holding all other variables at their means. The largest impact is obtained by the dummy *same_firm*. However, similar to the other two datasets, there is a severe difference in the magnitude of its influence. In the probit model, setting *same_firm* from zero to one leads to an increase in the citation probability of more than 2.5%. In the other two models, the same change amounts only to a shift in the probability of approximately 1.4%.

Many of the industry dummies are – in contrast to the other two datasets – not significantly different from zero. The coefficient of industry 13 (consumer goods) is not significant in the probit regression, but significant at the 1% level in the logit and complementary log-log regression. Of all the significant industry coefficients, we once again find the dummies of industry 4 (other machinery) and 7 (agricultural chemistry and machinery) the ones with the largest positive influence on the dependent variable. The largest negative impact on the probability of a knowledge spillover is caused by industry 14 (civil engineering) and 11 (mechanical engineering). The absolute value of the maximum negative discrete change is less than half as large as the maximum positive discrete change. Once again, all of these values have to be seen with respect to the influence of the base category, industry 1 (electrical engineering).

Finally, the coefficients of the country dummies are all highly significant. Just like for the previous two datasets, the probit model consistently predicts a larger impact on the probability due to a discrete change in one of the five country dummies than the logit and the complementary log-log model. Furthermore, all international country combinations obtain a higher probability of knowledge spillovers compared to the intra-Austrian case when the *same_firm* parameter is equal to zero. Germany and Switzerland, the USA and Canada, and other European countries have roughly the same influence on the dependent variable. Thus, there is again no evidence of increased knowledge spillovers due to spatial proximity. If the citing patent originates from the group of other countries, there is an even higher chance of knowledge spillovers than for the three aforementioned groups of countries. For Japanese patents, however, the probability of the occurrence of a patent citation is the lowest of the five international groups. The results of regressions with alternative model specifications can be found in Appendix B.3.

5.3 Comparisons

Aside from finding out which factors influence knowledge spillover and to which extent they do so, one objective of this thesis is to reveal if these aspects change over time. Furthermore, it might be interesting to scrutinize if there are any substantial differences between in- and outgoing knowledge spillovers. In order to assess these problems, three different datasets were constructed. In this subchapter, the results of the regressions of these datasets will be compared.

5.3.1 Comparison of the two periods of incoming knowledge spillovers

At first, the results of the regressions of the datasets concerned with incoming knowledge spillovers, REG_IN_PER1 and REG_IN_PER2, are compared. All of the six regressions show a similar goodness of fit, ranging between 0.18 and 0.19. The time indicators, *citing_appln_year* and *time_lag* (respectively *time_lag^2*), show a rather opposing behavior. While the application year of the citing patent has a negative influence on the citation probability in the second period, there is no significant impact of this variable in the first period at all. On the other hand, neither the linear nor the quadratic *time_lag* term is significantly different from zero in the second period, while both terms are significant in the first period and show the expected behavior. As mentioned before, this behavior consists of an increasing probability for small values of the time lag until a peak is reached and a monotonous decline of the probability afterwards.

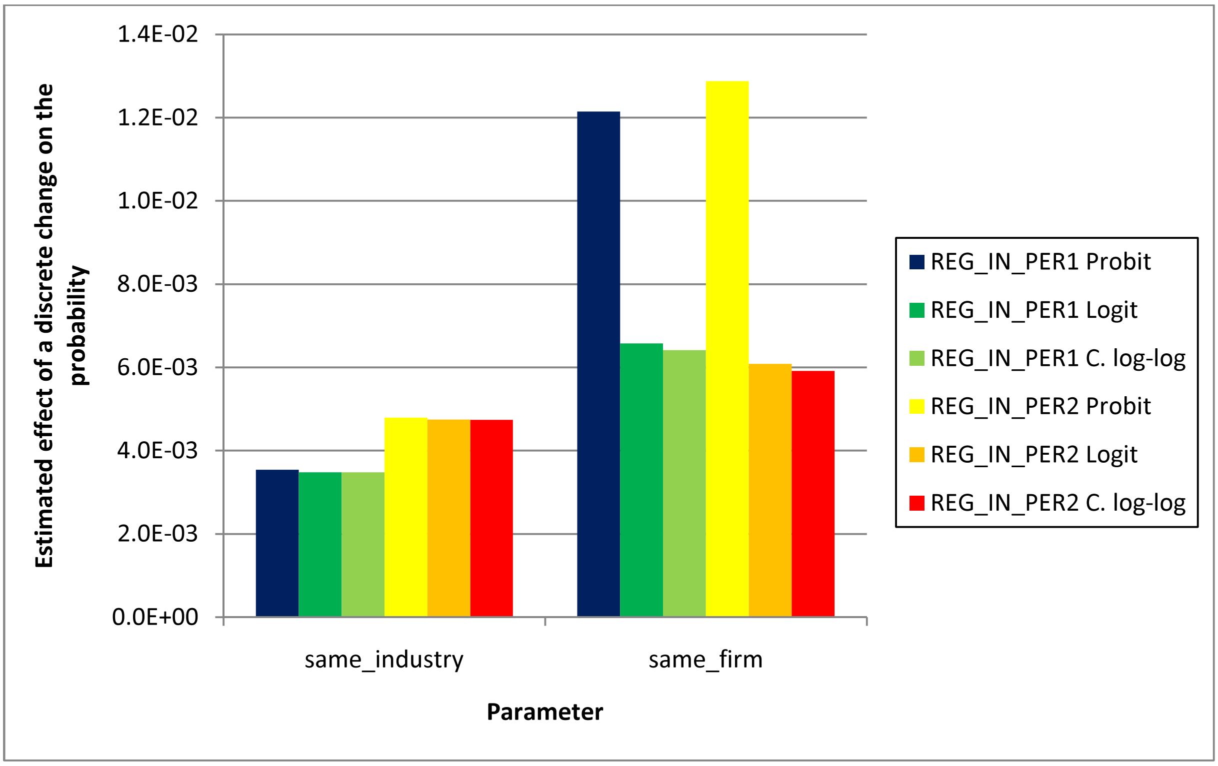

Figure 5.2 - Effects of *same_industry* and *same_firm* on incoming knowledge spillovers

A comparison of the effects of the variables *same_firm* and *same_industry* in the various regressions can be found in Figure 5.2. According to this graph, a discrete change in the *same_industry* dummy from zero to one (holding all other variables at their mean) has a larger effect in the second period compared to the first one. Thus, the importance of intra-industry knowledge spillovers appears to have risen over time. For the *same_firm* dummy, the interpretation is not as clear. On the one hand, the probit models suggest an increase of the positive impact of intra-firm citations in the second period. On the other hand, the logit and complementary log-log model estimate that the importance of intra-firm citations has slightly decreased in the latter period of time. Therefore, it is not possible to make an unambiguous statement about the development of intra-firm knowledge spillovers.

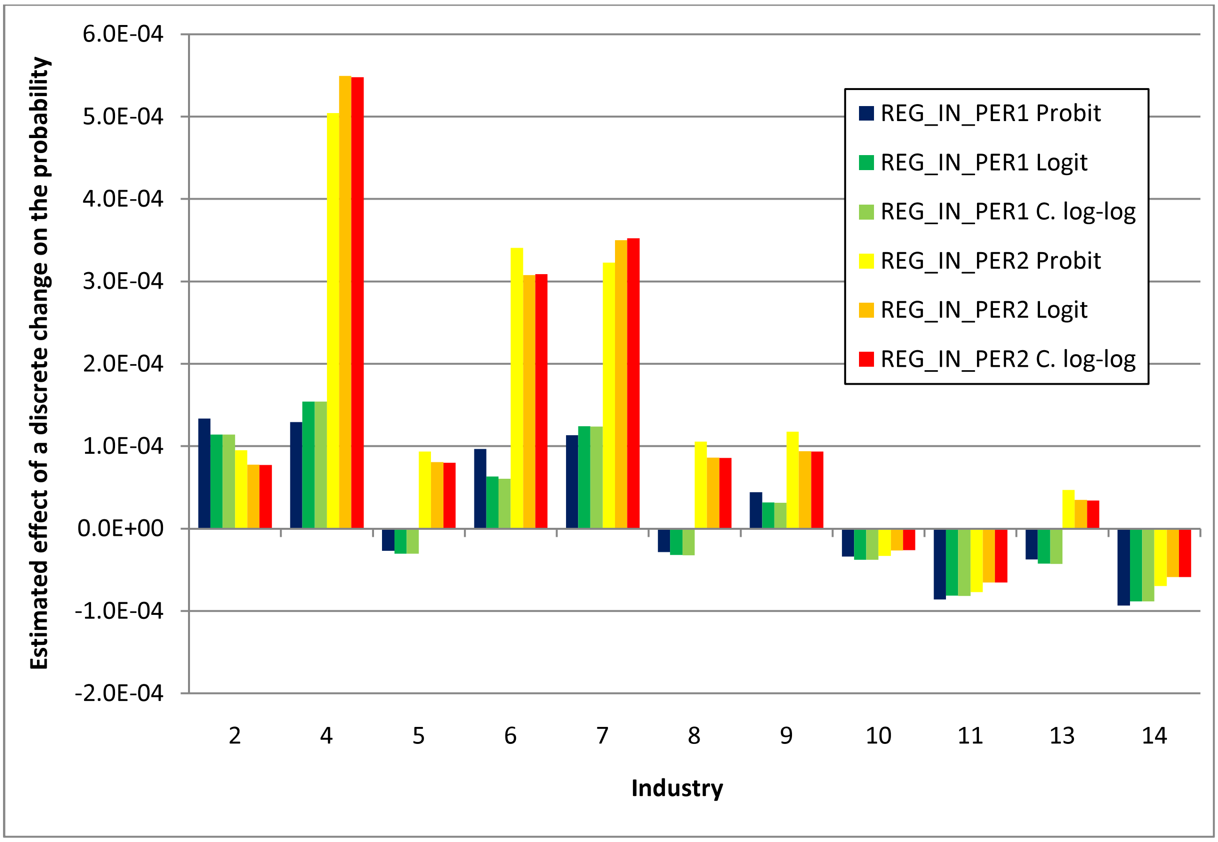

Figure 5.3 - Effects of *citing_industry* on incoming knowledge spillovers

The impact of the affiliation of the citing patent to a certain industry is depicted in Figure 5.3. Only the effects of industries with coefficients significantly different from zero were considered in this chart. It is important to remember that the effects of a discrete change in these dummies have to be viewed as relative to the base industry 1 (electrical engineering). For some of the industries, the impact on the probability of a knowledge spillover has remained almost constant over the two periods, e.g. for industries 10 (materials processing and handling) and 11 (mechanical engineering). Both industries have a negative influence on the probability. The negative influence of industry 14 (civil engineering) has decreased in the second period. For other industries, such as 4 (other machinery), 6 (pharmaceuticals and biotechnology), 7 (agricultural chemistry and machinery) and 9 (chemical and environmental engineering), the positive influence on the probability has increased (for some of them even multiplied) in the latter period. The influence of industry 2 (electronics) has slightly declined in the latter period.

Lastly, there are even industries for which the sign of the influence has changed from minus to plus. This group includes chemistry (5), material engineering (8) and consumer goods (13). These observations show that the industry dependency of the probability of a knowledge spillover has developed very differently for the various groups of industries. The positive impact of industries associated with chemistry and pharmaceuticals increased over time, while the influence of mechanical industries remained at a constant, but negative level.

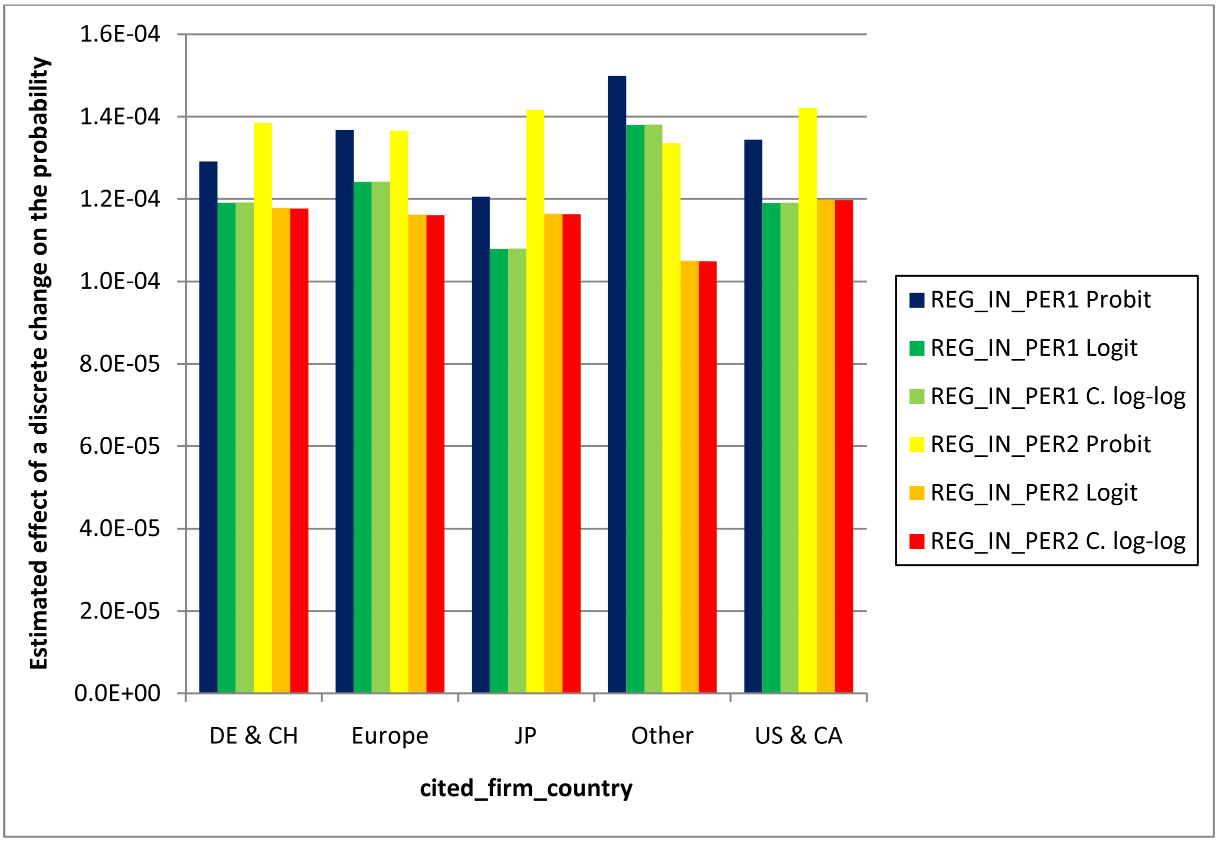

Figure 5.4 - Effects of *cited_firm_country* on incoming knowledge spillovers

The development of the impact of the country of origin of the applicant of the cited patent is displayed in Figure 5.4. The results of the graph have to be interpreted as relative to the base category, Austria. At first sight, the influence of the country of origin does not vary quite as much as the influence of the industry dummies. On closer inspection, however, one can notice a few interesting details. The positive influence of Japanese and North American patents on the knowledge spillover probability has increased from period one to period two in all types of regressions. The impact of the group of other countries has declined in the second period, making it the least influential category.

The development of the dummies for German and Swiss as well as for other European countries is not as unambiguous in the three types of regressions. For the German speaking countries, the probit model suggests an increasing influence, while the logit and complementary log-log models predict a slight decrease of the probability. In neither period, there is an exceptional influence of the dummy for German and Swiss patents recognizable.

For other European countries, the probit model does not indicate any change in the influence of the parameter, while the other two regression models suggest a decline of the impact. The fact that all of the international groups of countries have a positive impact on the probability compared to Austria is, as stated before, influenced by the choice to include the *same_firm* dummy. For both periods, regressions without *same_firm* suggest that other countries have a negative influence or no influence at all compared to Austrian patents.

Table 5.11 - Estimated maximum probabilities of incoming knowledge spillovers (incl. self citations)

		REG_IN_PER1	REG_IN_PER2
Intercept		1	1
citing_appln_year		1993	1999
time_lag		7	0
same_industry		1	1
same_firm		1	1
citing_industry		2 / 4[3]	4
cited_firm_country		AT	AT
$\widehat{\Pr}(y=1)$	Probit	0.0765	0.1398
	Logit	0.1146	0.2762
	C. log-log	0.1182	0.3086

Table 5.12 - Estimated maximum probabilities of incoming knowledge spillovers (excl. self citations)

		REG_IN_PER1	REG_IN_PER2
Intercept		1	1
citing_appln_year		1993	1999
time_lag		7	0
same_industry		1	1
same_firm		0	0
citing_industry		2 / 4	4
cited_firm_country		Other	US & CA
$\widehat{\Pr}(y=1)$	Probit	0.0077	0.0170
	Logit	0.0100	0.0270
	C. log-log	0.0101	0.0272

Aside from a comparison of the influences of the parameter relative to each other, it might also be interesting to calculate the actual probability of a knowledge spillover for selected values. Table 5.11 and Table 5.12 present the maximum values of the probability for

[3] In the probit regression of REG_IN_PER1, the industry with the largest positive influence on the probability is industry 2. However, in the logit and complementary log-log analysis of the same dataset, this spot is obtained by industry 4. Thus, the predicted probabilities of REG_IN_PER1 were calculated with industry 2 for the probit model and industry 4 for the other two models.

REG_IN_PER1 and REG_IN_PER2 with and without self citations, respectively. In Table 5.11 the variable *same_firm* is set equal to one, which implies that the country of origin of the cited patent has to be Austria. One finds that the predictions of the three different models vary quite strongly, especially between the probit and the other two regressions. Furthermore, the estimated maximum probabilities of the second period are more than twice as high compared to the first period (for the logit and complementary log-log model). The maximum probabilities without self citations draw a similar picture, however at a smaller scale. The values are about one tenth of the values including self citations. This illustrates the impact of the *same_firm* dummy. It is, however, not possible to draw any conclusions from a direct comparison of the maximum values of the two datasets due to the large difference in the number of observations as well as in the percentage of actual citations between both datasets.

5.3.2 Comparison of incoming and outgoing knowledge spillovers

Additionally to the comparison of the development of incoming knowledge spillovers over time, it is also useful to examine if there are any differences between incoming and outgoing knowledge spillovers. Since the covered period of Austrian patents in datasets REG_IN_PER1 and REG_OUT_PER1 is the same, it is possible to compare the results of the regressions of these datasets. According to the R^2_{McF}, the chosen models provide a better fit for the dataset concerned with outgoing knowledge spillovers (R^2_{McF} greater than 0.23).

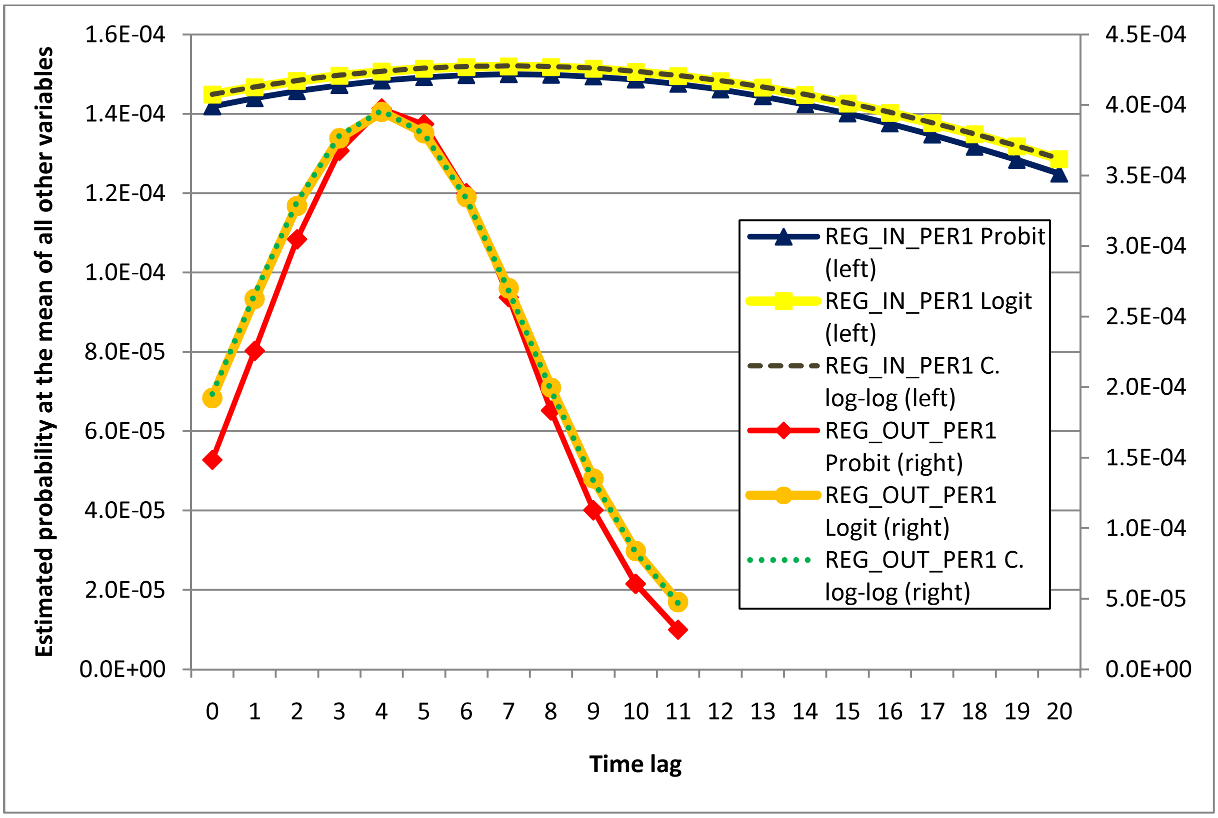

Figure 5.5 - Estimated probability dependent on *time_lag* for incoming and outgoing knowledge spillovers

While *citing_appln_year* does not have a significant influence on the probability of a spillover for incoming knowledge flows, this is not the case for outgoing spillovers. Each additional year decreases the probability that a given patent will cite an Austrian corporate patent. Figure 5.5 displays the estimated probability dependent on the time lag, holding all other variables at their means. For outgoing knowledge spillovers, this ceteris paribus graph has a shape similar to the one in the descriptive analysis, predicting the maximum probability of a citation for an average patent pair at a time lag of four years. For incoming knowledge spillovers, the graph is less pronounced. The curve is very flat, reaching its maximum value at a time lag of seven years. For both datasets, the logit and the complementary log-log regression predict virtually the same graph. The results of the probit analysis, however, differ slightly and provide lower probability values.

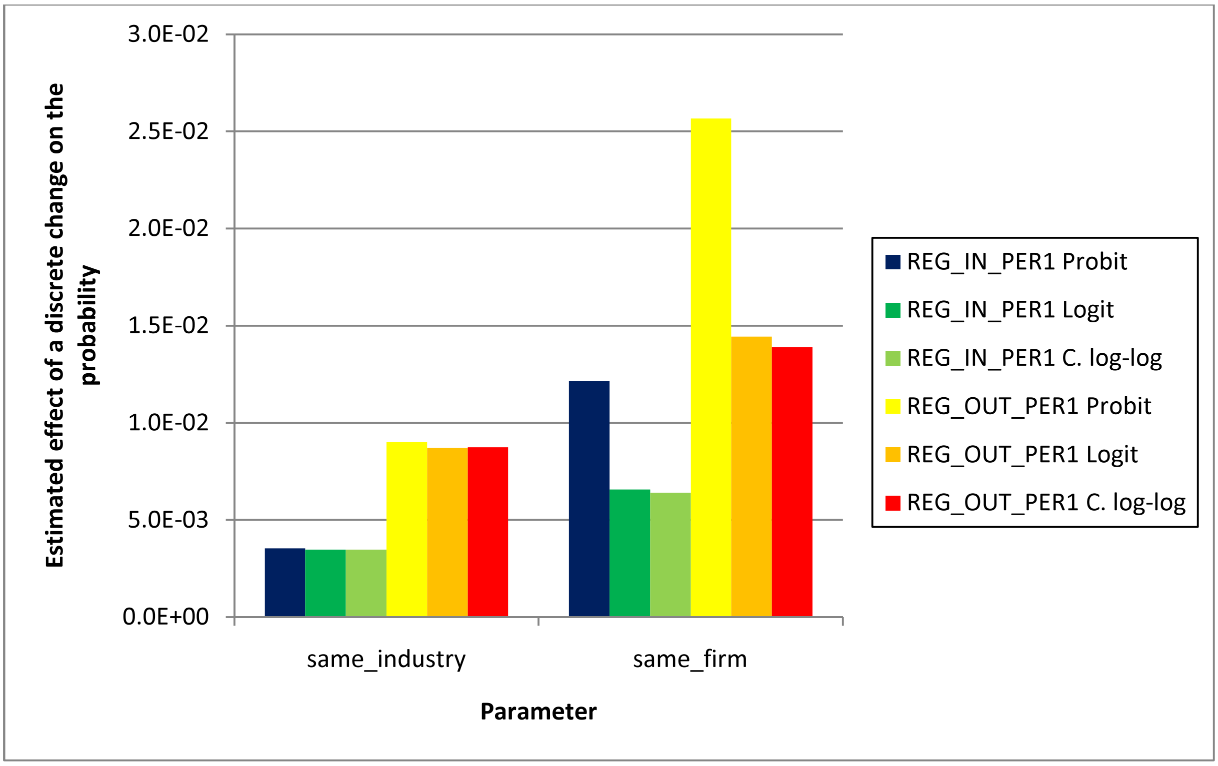

Figure 5.6 - Effects of *same_industry* and *same_firm* on incoming and outgoing knowledge spillovers

The effects of the parameters *same_industry* and *same_firm* are summarized in Figure 5.6. It is apparent from this graph that both variables have a larger influence on the probability for outgoing rather than for incoming knowledge spillovers. For all regression models, the positive impact of *same_industry* is more than twice as high for outgoing spillovers compared to incoming ones. The same statement holds for the parameter *same_firm*, but there is a major difference between the magnitudes of the values of the various regression models. The effect of a discrete change estimated with the probit model is almost twice as large compared to the estimation with the logit and complementary log-log model. Nevertheless, self citations have a far greater influence on outgoing knowledge spillovers than on incoming knowledge flows.

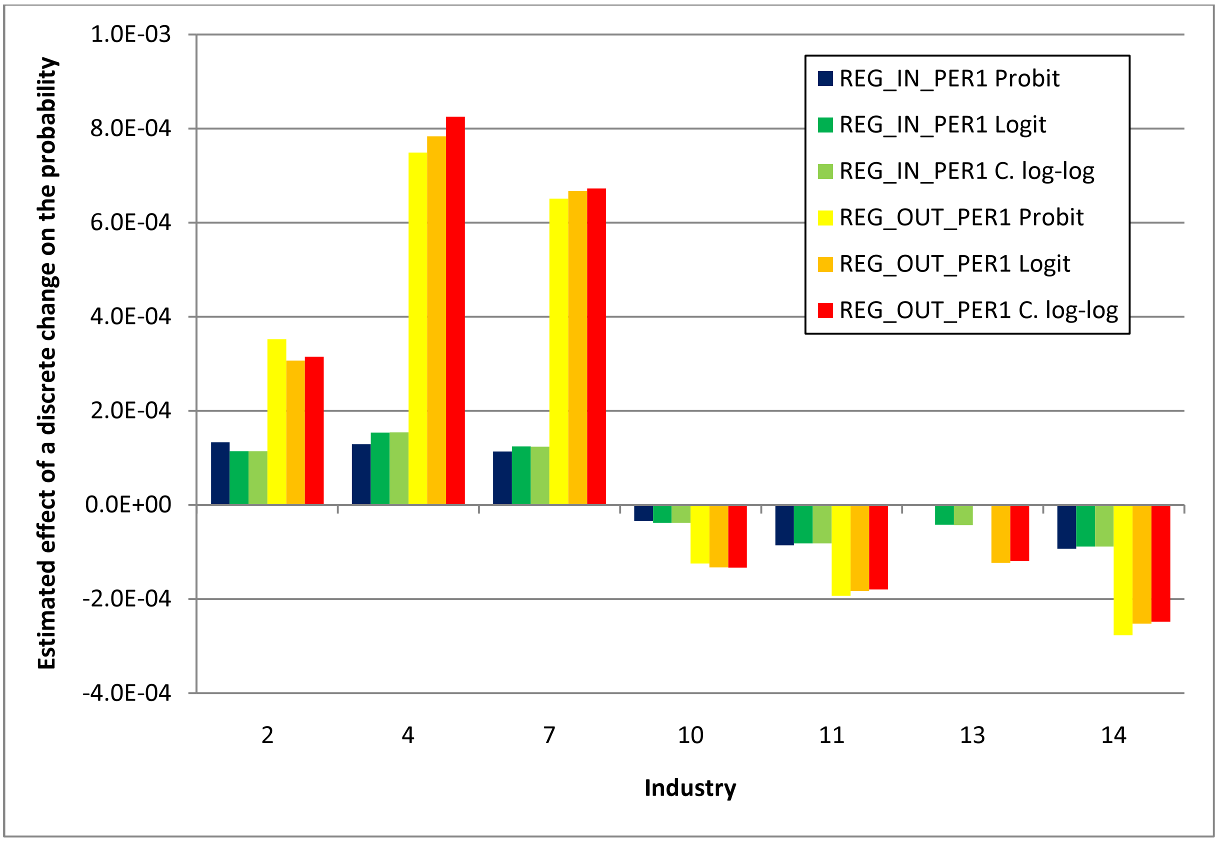

Figure 5.7 - Effects of *citing_industry* on incoming and outgoing knowledge spillovers[4]

Figure 5.7 depicts a comparison of the effects of the industry dummies in the regressions of REG_IN_PER1 and REG_OUT_PER1. The chart consists of fewer industries than for the previous comparison since more dummies are not significantly different from zero in the regressions of dataset REG_OUT_PER1. For the significant variables, there is a trend clearly evident from Figure 5.7: The direction of the impact on the probability is the same for incoming and outgoing knowledge spillovers, but the magnitudes can differ quite strongly. Industries 2 (electronics), 4 (other machinery) and 7 (agricultural chemistry and machinery) have a positive influence on the probability of a knowledge spillover. However, the magnitude of the impact is two to four times higher for outgoing flows of knowledge. On the other hand, industries 10 (materials processing and handling), 11 (mechanical engineering), 13 (consumer goods) and 14 (civil engineering) constitute a negative influence compared to the base category on the probability of both incoming and outgoing knowledge spillovers, again with the heavier impact in the regressions concerned with outgoing knowledge spillovers. Summarizing this aspect, the major difference between incoming and outgoing spillovers

[4] The coefficient of industry 13 is not significantly different from zero in the probit regression. Thus, the probit results of this variable were not considered in Figure 5.7. However, in the logit and complementary log-log regressions, the coefficient of industry 13 is significant at the 1% level.

concerning the influence of the industry affiliation is the magnitude of its impact. Furthermore, many industries do not have an impact at all on the probability of outgoing knowledge spillovers.

A comparison of the effects of the country of origin of the counterpart of the Austrian patent can be found in Figure 5.8. The graph reveals that the geographical influence is much larger for outgoing knowledge spillovers than for incoming flows of knowledge. In fact, the geographical influence on outgoing knowledge spillovers is approximately thrice as large as on incoming knowledge spillovers. Furthermore, one finds that the relative importance of the five groups of countries is very similar for incoming and outgoing knowledge spillovers. In both directions, the group of other countries has the largest impact on the probability compared to the reference category, Austria. On the other hand, Japan has the smallest impact on the dependent variable in both cases. Spatial and linguistic proximity does not have an outstanding impact on the probability in either dataset. The impact of German and Swiss patents does not differ significantly from the one of other European and North American patents in both directions. Thus, similar to the conclusion of the industry analysis, the major difference in the geographical influence between incoming and outgoing knowledge spillovers is the magnitude of the impact, regardless of which binary response model is used in the regression.

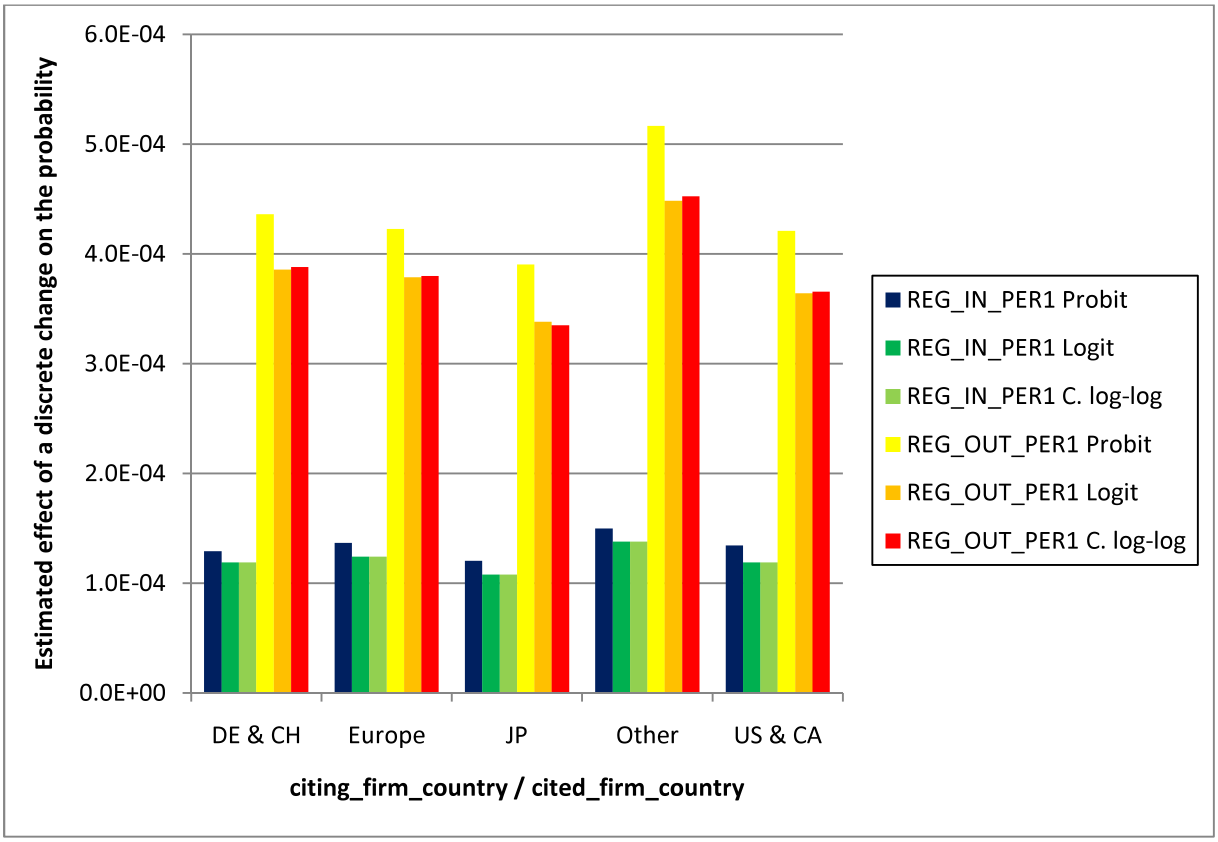

Figure 5.8 - Effects of the country dummies on incoming and outgoing knowledge spillovers

Table 5.13 - Estimated maximum probabilities for incoming and outgoing knowledge spillovers (incl. self citations)

		REG_IN_PER1	REG_OUT_PER1
Intercept		1	1
citing_appln_year		1993	1997
time_lag		7	4
same_industry		1	1
same_firm		1	1
citing_industry		2 / 4	4
cited_firm_country		AT	AT
	Probit	0.0765	0.1827
$\widehat{\Pr}(y=1)$	Logit	0.1146	0.3068
	C. log-log	0.1182	0.3502

Table 5.14 - Estimated maximum probabilities for incoming and outgoing knowledge spillovers (excl. self citations)

		REG_IN_PER1	REG_OUT_PER1
Intercept		1	1
citing_appln_year		1993	1997
time_lag		7	4
same_industry		1	1
same_firm		0	0
citing_industry		2 / 4	4
cited_firm_country		Other	Other
	Probit	0.0077	0.0309
$\widehat{\Pr}(y=1)$	Logit	0.0100	0.0444
	C. log-log	0.0101	0.0466

Finally, Table 5.13 and Table 5.14 present a comparison of the estimated maximum probabilities including and excluding self citations, respectively. For outgoing knowledge flows, the maximum value is attained with an application year of the citing patent greater than the minimum value, although the coefficient of this variable has a negative sign. This is due to the fact that the impact of the time lag holds its maximum at four years and its magnitude exceeds the one of the application year of the citing patent. Since the earliest application year of the cited patent in the dataset is 1993, the application year of the citing patent has to be at least 1997 when considering a time lag of four years. We find that the maximum values for outgoing knowledge spillovers exceed the ones for incoming flows by far. Although there are severe differences in the predictions among the three models, the maximum probability for outgoing knowledge spillovers is at least three times higher than for incoming knowledge flows. However, as mentioned before, a direct comparison of the maximum probabilities of the two datasets is not viable. Once again, there is a large difference between including and excluding self citations. Furthermore, one finds that the specifications to obtain the maximum probability are very similar, apart from the difference in the time lag. However, as noted in the descriptive analysis, due to the truncation bias of “cutting off” possible citations of patents with a later application date in the outgoing spillovers dataset, the time lag providing the maximum value of the probability might actually be higher than four years.

6 Conclusions

This thesis was intended to shed some light on the influences on knowledge spillovers originated from or received by Austrian firms. In order to achieve this goal, the patent citation approach of Jaffe and Trajtenberg (1996; 1999) was adapted for Austrian patents applied for at the European Patent Office (EPO) and at the United States Patent and Trademark Office (USPTO). Although the existence and the overall concept of knowledge spillovers sound very intuitive and reasonable at first, there is a certain vagueness inherent in this field of research. First of all, there have been developed several notions of what is regarded as a knowledge spillover and what is attributable to other phenomena. Furthermore, there is no unique way of measuring and quantifying knowledge spillovers hitherto. Thus, it is very difficult to integrate the effects of knowledge flows into quantitative economic models.

The econometric approach used in this thesis utilizes patent citations as a proxy for knowledge spillovers in order to assess the influence of various parameters on the probability of a flow of knowledge. However, this method clearly has its limitations: Firstly, patents represent only a small fraction of the set of inventions and innovations developed by firms. Secondly, a patent citation does not necessarily indicate a knowledge transfer among the applicants of the involved patents, since citations can also be added by the patent examiner at the patent office without awareness of either applicant. Nevertheless, this approach constitutes a valuable tool for gaining more insight into which factors influence the occurrence of knowledge spillovers.

6.1 Summarizing the results

Patent data from the PATSTAT database of the EPO was used for the regressions in this thesis and arranged in three different datasets in order to allow for comparisons over time and between incoming and outgoing knowledge spillovers. The dataset concerned with incoming knowledge spillovers during the second period has to be interpreted with caution due to the possibility of missing data. The focus of the analyses was laid on the influence of application date, geographical origin and industry affiliation of the involved patents. Already the descriptive analysis of the datasets revealed some interesting insights. As expected, the majority of the citations occur between patents of the same industry. Furthermore, the enhanced engagement of Austrian firms in certain industry sectors, such as the mechanical and chemical industry, is apparent from the number of patents as well as citations.

The analysis of the time lag between citing and cited patent discloses a typical trend: The number of citations increases for small values of the time lag until a peak is reached at a time lag of three to six years. After the peak, the number of citations decreases at a smaller rate.

This behavior is a strong argument for the existence of two processes working in opposite directions: the diffusion process and the obsolescence process. The examination of the geographical origin of the citation counterpart of the Austrian patent yields a strong connection between Austria and the USA, Germany as well as Japan. Moreover, there was a strong tendency of Austrian firms to cite patents applied for by the same firm.

Furthermore, regressions with binary response models were computed in order to assess the influence of the variables on the probability of a knowledge spillover. The results suggest that the impact of the time variables is rather limited. The year of application of the citing patent appears to have only a small influence on the probability compared to other variables. The time lag, on the other hand, shows the expected behavior in the two datasets of the first period: a positive linear term and a smaller, negative quadratic term. This resembles the results gained in the preliminary analysis. Nevertheless, the absolute as well as the relative impact (in comparison to the other variables) of these parameters is rather small.

The influence of the various industries of the patents was accounted for in two ways: firstly, by taking the industry of the citing patent into account; secondly, by introducing a variable indicating if the two patents of a pair are part of the same industry. The latter variable constitutes a major influence on the probability, exceeding the influence of the particular industry variables by far. Thus, relative to the base category, there is always an increase in the probability if both patents are part of the same industry, regardless of which industry they are affiliated to. The values of the industry dummies suggest that those industries with the smallest share of actual citations in the descriptive analysis have the largest probability of a patent citation. Thus, there may be, for instance, few agricultural patents issued to Austrian companies, but for each issued patent the probability of a knowledge spillover is much higher than for a patent pertaining to, e.g., a mechanical industry.

Concerning the influence of the country of origin of the applicant of the cited, respectively the citing patent, there is no sign of increased knowledge spillovers due to spatial proximity. In general, there is very little difference in the influences of the five groups of countries compared to the reference category, Austria. The values of their coefficients are in approximately the same range for each respective dataset, with slight variations in the exact order of the dummies. As mentioned before, the choice to include the variable *same_firm* as well as its properties influence also the results of the country dummies. If *same_firm* is not considered in the analysis, intra-Austrian knowledge spillovers are the most probable of all possible country combinations.

The comparisons between the results of the regressions of the various datasets have yielded some interesting outcomes. An analysis of the development of incoming knowledge spillovers over time provides evidence that the impact of most of the dummy variables has increased or at least stayed at the same level. For some industry dummies, even the sign of the coefficient changed from minus to plus relative to the base category. Only a few country dummies showed a decline in the second period. The comparison between incoming and outgoing knowledge spillovers of the first period suggests that the influence of the time variables is similar in both cases. However, the curve of the probability as a function of the time lag is not as flat for outgoing knowledge spillovers as for incoming ones and the maximum value appears at a smaller time lag. In general, one can say that the tendency of the impact of the variables is the same in both cases, but the magnitude of the impact on the probability is much larger for outgoing spillovers. The only exceptions are the coefficients of several industry dummies that are not significantly different from zero for outgoing knowledge spillovers.

6.2 Further research

Since the method of using patent citations to track down knowledge spillovers is a rather recent development, there is still much room for improvement as well as the examination of other research questions. A possible advancement concerns certainly the use of more complete data. While there already exists a database, which is issued by the National Bureau of Economic Research, that is focused especially on patent citations (Hall, Jaffe and Trajtenberg, 2002), this database includes only patents applied for at the USPTO. The PATSTAT database issued by the EPO, on the other hand, suffers from incomplete records and double entries, which require a lot of data preprocessing. Newer versions of the database certainly provide an improvement concerning the quality of the data. Thus, conducting an analysis similar to the one of this thesis with newer and more complete data will enhance its explanatory power and might yield new insights.

Another possible route for future analyses is to use different models or a different approach in constructing the dataset. This thesis explored which factors influence the citation probability between patents that have issued or received, respectively, at least one citation in the relevant periods of time. The availability of more computing power and a more complete patent database might allow an analysis of the probability on the basis of all patents, regardless of whether they were already involved in a citation relationship. This may, for instance, be done by drawing random samples of patent pairs from the database and indicating if there exists a citation relationship between such a pair. However, the sample has to be sufficiently large in order to account for the small share of actual citations among the set of possible patent combinations. Furthermore, it is necessary to address the truncation bias that appears when an analysis is focused on patents that received citations. A first approach can be found in Trajtenberg (1990), but it is necessary to find a more general solution.

Finally, it appears inherently useful to combine data from a patent database with other data in order assess how non-patent related factors influence the probability of knowledge spillovers. For instance, it might be interesting to examine how the size of a firm, its R&D expenditures or competition aspects influence the occurrence of patent citations. However, it is also possible to improve the expressiveness of the variables that were already used in this thesis. For example, by including a finer classification of the origin of the applicant or inventor of a patent (other than countries or groups of countries), one is more likely to find signs of a localization of knowledge spillovers. A first approach in this direction was conducted by Maurseth and Verspagen (2002) for European patent data.

List of references

ALCÁCER, J. AND GITTELMAN, M. (2006) Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. *The Review of Economics and Statistics*, Vol. 88 (4), pp. 774-779.

AMIR, R. (2000) Modelling Imperfectly Appropriable R&D via Spillovers. *International Journal of Industrial Organization*, Vol. 18 (7), pp. 1013-1032.

BACCHIOCCHI, E. AND MONTOBBIO, F. (2010) International Knowledge Diffusion and Home-bias effect: Do USPTO and EPO Patent Citations Tell the Same Story? *The Scandinavian Journal of Economics*, Vol. 112 (3), pp. 441-470.

BOUND, J., CUMMINS, C., GRILICHES, Z., HALL, B. H. AND JAFFE, A. (1984) Who Does R & D and Who Patents? In: Griliches, Z. (ed.): *R & D, Patents, and Productivity*. The University of Chicago Press, Chicago, IL, pp. 21-54.

BRESCHI, S. AND LISSONI, F. (2001a) Knowledge Spillovers and Local Innovation Systems: A Critical Survey. *Industrial and Corporate Change*, Vol. 10 (4), pp. 975-1005.

BRESCHI, S. AND LISSONI, F. (2001b) Localised Knowledge Spillovers vs. Innovative Milieux: Knowledge "Tacitness" Reconsidered. *Papers in Regional Science*, Vol. 80 (3), pp. 255-273.

EUROSTAT (2011) Introduction to NUTS. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction, (accessed 2011-04-18).

CHESBROUGH, H. W. (2003) *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business School Press, Boston, MA.

CHRISTENSEN, C. M. (1997) *The Innovator's Dilemma. The Revolutionary Book that Will Change the Way You Do Business*. Harvard Business School Press, Boston, MA.

COHEN, W. M. AND LEVINTHAL, D. A. (1990) Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, Vol. 35 (1), pp. 128-152.

CORNES, R. AND SANDLER, T. (1996) *The Theory of Externalities, Public Goods and Club Goods*. 2nd Ed., Cambridge University Press, Cambridge, United Kingdom.

CRISCUOLO, P. AND VERSPAGEN, B. (2008) Does It Matter Where Patent Citations Come From? Inventor vs. Examiner Citations in European Patents. *Research Policy*, Vol. 37 (10), pp. 1892-1908.

D'ASPREMONT, C. AND JACQUEMIN, A. (1988) Cooperative and Noncooperative R&D in Duopoly with Spillovers. *The American Economic Review*, Vol. 78 (5), pp. 1133-1137.

D'ASPREMONT, C. AND JACQUEMIN, A. (1990) Cooperative and Noncooperative R&D in Duopoly with Spillovers: Erratum. *The American Economic Review*, Vol. 80 (3), pp. 641-642.

DE BONDT, R. (1997) Spillovers and Innovative Activities. *International Journal of Industrial Organization*, Vol. 15 (1), pp. 1-28.

DUGUET, E. AND MACGARVIE, M. (2005) How Well Do Patent Citations Measure Flows of Technology? Evidence from French Innovation Surveys. *Economics of Innovation and New Technology*, Vol. 14 (5), pp. 375-393.

EPO (2006) Data Catalog for the EPO Worldwide Patent Statistical Database, Version 3.22. Appended to the April 2006 Edition of the PATSTAT Database. Updated Version (4.10) available at: http://www.epo.org/patents/patent-information/raw-data/test/product-14-24_de.html (accessed 2011-04-18).

GLAESER, E. L., KALLAL, H. D., SCHEINKMAN, J. A. AND SHLEIFER, A. (1992) Growth in Cities. *Journal of Political Economy*, Vol. 100 (6), pp. 1126-1152.

GRILICHES, Z. (1979) Issues in Assessing the Contribution of Research and Development to Productivity Growth. *The Bell Journal of Economics*, Vol. 10 (1), pp. 92-116.

GRILICHES, Z. (1990) Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature*, Vol. 28 (4), pp. 1661-1707.

GRILICHES, Z. (1992) The Search for R&D Spillovers. *The Scandinavian Journal of Economics*, Vol. 94, Supplement. Proceedings of a Symposium on Productivity Concepts and Measurement Problems: Welfare, Quality and Productivity in the Service Industries, pp. 29-47.

GROSSMAN, G. M. AND HELPMAN, E. (1991) *Innovation and Growth in the Global Economy*. MIT Press, Cambridge, MA.

HALL, B. H. (2000) Innovation and Market Value. In: Barell, R., Mason, G. and O'Mahony, M. (eds.): *Productivity, Innovation and Economic Performance*. Cambridge University Press, Cambridge, United Kingdom, pp. 177-198.

HALL, B. H., JAFFE, A. B. AND TRAJTENBERG, M. (2002) The NBER Patent-Citations Data File: Lessons, Insights and Methodological Tools. In: Jaffe, A. B. and Trajtenberg, M. (eds.): *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge, MA, pp. 403-460.

HALL, B. H., JAFFE, A. AND TRAJTENBERG, M. (2005) Market Value and Patent Citations. *The RAND Journal of Economics*, Vol. 36 (1), pp. 16-38.

HENDERSON, R., JAFFE, A. B. AND TRAJTENBERG, M. (1998) Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-1988. *The Review of Economics and Statistics*, Vol. 80 (1), pp. 119-127.

HENRIQUES, I. (1990) Cooperative and Noncooperative R&D in Duopoly with Spillovers. *The American Economic Review*, Vol. 80 (3), pp. 638-640.

HOWELLS, J. R. L. (2002) Tacit Knowledge, Innovation and Economic Geography. *Urban Studies*, Vol. 39 (5-6), pp. 871-884.

HU, A. G. Z. AND JAFFE, A. B. (2003) Patent Citations and International Knowledge Flow: The Cases of Korea and Taiwan. *International Journal of Industrial Organization*, Vol. 21 (6), pp. 849-880.

HUBER, F. (2007) *Social Networks and Knowledge Spillovers*. Peter Lang/Europäischer Verlag der Wissenschaften, Frankfurt/Main, Germany.

JAFFE, A. B. AND LERNER, J. (2001) Reinventing Public R&D: Patent Policy and the Commercialization of National Laboratory Technologies. *The RAND Journal of Economics*, Vol. 32 (1), pp. 167-198.

JAFFE, A. B. AND TRAJTENBERG, M. (1996) Flows of Knowledge from Universities and Federal Laboratories: Modeling the Flow of Patent Citations over Time and across Institutional and Geographic Boundaries. *Proceedings of the National Academy of Sciences*, Vol. 93 (23), pp. 12671-12677.

JAFFE, A. B. AND TRAJTENBERG, M. (1999) International Knowledge Flows: Evidence from Patent Citations. *Economics of Innovation and New Technology*, Vol. 8 (1-2), pp. 105-136.

JAFFE, A. B., TRAJTENBERG, M. AND FOGARTY, M. S. (2000) Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors. *The American Economic Review*, Vol. 90 (2), pp. 215-218.

JAFFE, A. B., TRAJTENBERG, M. AND FOGARTY, M. S. (2002) The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees. In: Jaffe, A. B. and Trajtenberg, M. (eds.): *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge, MA, pp. 379-402.

JAFFE, A. B., TRAJTENBERG, M. AND HENDERSON, R. (1993) Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, Vol. 108 (3), pp. 577-598.

KAMIEN, M. I., MULLER, E. AND ZANG, I. (1992) Research Joint Ventures and R&D Cartels. *The American Economic Review*, Vol. 82 (5), pp. 1293-1306.

KRUGMAN, P. (1991) *Geography and Trade*. MIT Press, Cambridge, MA.

LANGINIER, C. (2005) Using Patents to Mislead Rivals. *Canadian Journal of Economics*, Vol. 38 (2), pp. 520-545.

LANJOUW, J. O., PAKES, A. AND PUTNAM, J. (1998) How to Count Patents and Value Intellectual Property: The Uses of Patent Renewal and Application Data. *The Journal of Industrial Economics*, Vol. 46 (4), pp. 405-432.

LONG, J. S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications, London, United Kingdom.

LUKACH, R. AND PLASMANS, J. (2002) Measuring Knowledge Spillovers Using Patent Citations: Evidence from the Belgian Firm's Data. *CESifo Working Paper*, No. 754, July, Munich, DE.

MALMBERG, A. AND MASKELL, P. (2002) The Elusive Concept of Localization Economies: Towards a Knowledge-based Theory of Spatial Clustering. *Environment and Planning A*, Vol. 34 (3), pp. 429-449.

MANSFIELD, E. (1986) Patents and Innovation: An Empirical Study. *Management Science*, Vol. 32 (2), pp. 173-181.

MARTIN, S. (2002) *Advanced Industrial Economics*. 2^{nd} Ed., Blackwell Publishing, Oxford, United Kingdom.

MAURSETH, P. B. AND VERSPAGEN, B. (2002) Knowledge Spillovers in Europe: A Patent Citation Analysis. *The Scandinavian Journal of Economics*, Vol. 104 (4), pp. 531-545.

PAKES, A. (1985) On Patents, R & D, and the Stock Market Rate of Return. *Journal of Political Economy*, Vol. 93 (2), pp. 390-409.

PAKES, A. AND GRILICHES, Z. (1984) Patents and R & D at the Firm Level: A First Look. In: Griliches, Z. (ed.): *R & D, Patents, and Productivity*. The University of Chicago Press, Chicago, IL, pp. 55-72.

PATSTAT (2006) Worldwide Patent Statistical Database issued by the European Patent Office, Edition April 2006, Vienna, Austria.

PAVITT, K. (1985) Patent Statistics as Indicators of Innovative Activities: Possibilities and Problems. *Scientometrics*, Vol. 7 (1-2), pp. 77-99.

R DEVELOPMENT CORE TEAM (2010) R: A Language and Environment for Statistical Computing, Version 2.11.1. R Foundation for Statistical Computing, Vienna, Austria. Downloadable at: http://www.R-project.org (accessed 2011-04-18).

ROMER, P. M. (1990) Endogenous Technological Change. *Journal of Political Economy*, Vol. 98 (5), pt. 2, pp. 71-102.

SALANT, S. W. AND SHAFFER, G. (1998) Optimal Asymmetric Strategies in Research Joint Ventures. *International Journal of Industrial Organization*, Vol. 16 (2), pp. 195-208.

SCHMOCH, U. (2008) Concept of a Technology Classification for Country Comparisons. Final Report to the World Intellectual Property Organisation (WIPO). Available at: http://www.wipo.int/ipstats/en/statistics/patents/ (accessed 2011-04-18).

SUZUMURA, K. (1992) Cooperative and Noncooperative R&D in an Oligopoly with Spillover. *The American Economic Review*, Vol. 82 (5), pp. 1307-1320.

TRAJTENBERG, M. (1990) A Penny for Your Quotes: Patent Citations and the Value of Innovation. *The RAND Journal of Economics*, Vol. 21 (1), pp. 172-187.

UTTERBACK, J. M. (1994) *Mastering the Dynamics of Innovation*. Harvard Business School Press, Boston, MA.

VERSPAGEN, B., VAN MOERGASTEL, T. AND SLABBERS, M. (1994) MERIT Concordance Table: IPC – ISIC (rev. 2). *MERIT Research Memorandum*, No. 2/94-004.

WIPO (2004) *WIPO Intellectual Property Handbook: Policy, Law and Use*. 2nd Ed., WIPO Publication No. 489, Geneva, Switzerland.

WIPO (2011) Guide to the IPC, Version 2011. Available at: http://www.wipo.int/classifications/ipc/en/general/ (accessed 2011-04-18).

WOOLDRIDGE, J. M. (2005) *Introductory Econometrics: A Modern Approach*. 3rd Ed., South Western Publishing, Mason, OH.

Appendix

A. Results of non-iterative computations

Table A.1 - Result of the non-iterative computation of the probit analysis of REG_OUT_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	29.1110	12.255	———	2.38	*
citing_appln_year	-0.0166	0.006	-0.00005	-2.70	**
time_lag	0.1248	0.011	0.00007	11.19	***
(time_lag)^2	-0.0148	0.001		-10.82	***
same_industry	1.1091	0.017	0.00903	66.90	***
same_firm	1.4271	0.038	0.02574	37.99	***
citing_industry					
2 (electronics)	0.1620	0.047	0.00035	3.47	***
3 (instruments)	-0.0043	0.040	-0.00001	-0.11	
4 (other machinery)	0.2817	0.091	0.00076	3.11	**
5 (chemistry)	-0.0222	0.040	-0.00004	-0.55	
6 (pharmaceuticals and biotechnology)	0.0992	0.051	0.00019	1.94	
7 (agricultural chemistry and machinery)	0.2676	0.070	0.00070	3.83	***
8 (material engineering)	-0.0154	0.044	-0.00002	-0.35	
9 (chemical and environmental engineering)	0.0851	0.044	0.00016	1.92	
10 (materials processing and handling)	-0.0871	0.038	-0.00012	-2.31	*
11 (mechanical engineering)	-0.1482	0.036	-0.00019	-4.13	***
12 (transport)	-0.0213	0.040	-0.00003	-0.54	
13 (consumer goods)	-0.0715	0.037	-0.00011	-1.91	
14 (civil engineering)	-0.2481	0.040	-0.00028	-6.15	***
citing_firm_country					
DE & CH	0.4275	0.033	0.00044	12.92	***
Europe	0.4197	0.035	0.00043	11.96	***
JP	0.4018	0.036	0.00039	11.02	***
Other	0.4655	0.050	0.00052	9.31	***
US & CA	0.4181	0.032	0.00042	13.07	***
Size of dataset	1939519				
R^2_{McF}	0.2345				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
citing_firm_country	AT				
Significance levels	***	0.001			
	**	0.01			
	*	0.05			

Table A.2 - Result of the non-iterative computation of the logit analysis of REG_OUT_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	64.2095	34.743	———	1.85	
citing_appln_year	-0.0370	0.017	-0.00003	-2.13	*
time_lag	0.3550	0.032	0.00005	10.97	***
(time_lag)^2	-0.0439	0.004		-10.87	***
same_industry	3.4686	0.054	0.00878	64.45	***
same_firm	3.6718	0.108	0.01452	34.15	***
citing_industry					
2 (electronics)	0.5109	0.137	0.00032	3.73	***
3 (instruments)	-0.0581	0.116	-0.00003	-0.50	
4 (other machinery)	1.0120	0.268	0.00083	3.77	***
5 (chemistry)	-0.0913	0.115	-0.00004	-0.79	
6 (pharmaceuticals and biotechnology)	0.1766	0.144	0.00009	1.23	
7 (agricultural chemistry and machinery)	0.8903	0.207	0.00068	4.30	***
8 (material engineering)	-0.1684	0.125	-0.00007	-1.35	
9 (chemical and environmental engineering)	0.2123	0.127	0.00011	1.67	
10 (materials processing and handling)	-0.3308	0.108	-0.00013	-3.07	**
11 (mechanical engineering)	-0.4886	0.103	-0.00018	-4.74	***
12 (transport)	-0.0768	0.113	-0.00003	-0.68	
13 (consumer goods)	-0.2912	0.107	-0.00012	-2.73	**
14 (civil engineering)	-0.7643	0.112	-0.00025	-6.80	***
citing_firm_country					
DE & CH	1.3098	0.105	0.00039	12.47	***
Europe	1.2944	0.110	0.00038	11.72	***
JP	1.2132	0.115	0.00034	10.57	***
Other	1.4293	0.152	0.00046	9.42	***
US & CA	1.2663	0.102	0.00037	12.40	***
Size of dataset	1939519				
R^2_{McF}	0.2321				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
citing_firm_country	AT				
Significance levels	***	0.001			
	**	0.01			
	*	0.05			

Table A.3 - Result of the non-iterative computation of the complementary log-log analysis of REG_OUT_PER1

	Estimate	Std. Error	Discrete Change	t-Value	
Intercept	61.1709	34.336	———	1.78	
citing_appln_year	-0.0355	0.017	-0.00003	-2.06	*
time_lag	0.3497	0.032	0.00005	10.92	***
(time_lag)^2	-0.0434	0.004		-10.84	***
same_industry	3.4609	0.054	0.00876	64.48	***
same_firm	3.6212	0.106	0.01390	34.03	***
citing_industry					
2 (electronics)	0.5055	0.136	0.00031	3.73	***
3 (instruments)	-0.0616	0.114	-0.00003	-0.54	
4 (other machinery)	1.0105	0.264	0.00083	3.82	***
5 (chemistry)	-0.0945	0.114	-0.00004	-0.83	
6 (pharmaceuticals and biotechnology)	0.1544	0.142	0.00008	1.09	
7 (agricultural chemistry and machinery)	0.8766	0.203	0.00067	4.32	***
8 (material engineering)	-0.1830	0.123	-0.00008	-1.49	
9 (chemical and environmental engineering)	0.2035	0.126	0.00011	1.62	
10 (materials processing and handling)	-0.3349	0.107	-0.00014	-3.14	**
11 (mechanical engineering)	-0.4888	0.102	-0.00018	-4.79	***
12 (transport)	-0.0795	0.112	-0.00004	-0.71	
13 (consumer goods)	-0.2985	0.105	-0.00012	-2.83	**
14 (civil engineering)	-0.7596	0.111	-0.00025	-6.85	***
citing_firm_country					
DE & CH	1.3059	0.105	0.00039	12.46	***
Europe	1.2914	0.110	0.00038	11.71	***
JP	1.2094	0.115	0.00034	10.56	***
Other	1.4271	0.151	0.00046	9.44	***
US & CA	1.2641	0.102	0.00037	12.40	***
Size of dataset	1939519				
R^2_{McF}	0.2320				
Base categories for dummy variables					
citing_industry	1 (electrical engineering)				
citing_firm_country	AT				
	***	0.001			
Significance levels	**	0.01			
	*	0.05			

B. Results of regressions with alternative specifications

B.1 REG_IN_PER1

Table B.1 - Results of alternative probit analyses of REG_IN_PER1

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	1.7753 (3.790)	-3.2488 (3.518)	0.5787 (3.392)
citing_appln_year	-0.0027 (0.002)	-0.0002 (0.002)	-0.0019 (0.002)
time_lag	0.0015 (0.002)	0.0027 (0.002)	0.0008 (0.002)
(time_lag)^2	-0.0002* (0.000)	-0.0002** (0.000)	-0.0001 (0.000)
same_industry	1.0697*** (0.007)	————	————
same_firm	————	1.7879*** (0.019)	————
citing_industry			
2 (electronics)	0.1594*** (0.022)	0.0793*** (0.020)	0.0845*** (0.019)
3 (instruments)	0.0014 (0.018)	0.1151*** (0.016)	0.1227*** (0.016)
4 (other machinery)	0.1559*** (0.039)	-0.0182 (0.035)	-0.0204 (0.035)
5 (chemistry)	-0.0433* (0.018)	0.0445** (0.016)	0.0486** (0.016)
6 (pharmaceuticals and biotechnology)	0.1489*** (0.019)	0.0999*** (0.017)	0.1225*** (0.017)
7 (agricultural chemistry and machinery)	0.1586*** (0.036)	-0.0196 (0.034)	-0.0060 (0.033)
8 (material engineering)	-0.0380* (0.019)	0.0153 (0.017)	0.0285 (0.017)
9 (chemical and environmental engineering)	0.0605** (0.019)	0.0734*** (0.018)	0.0752*** (0.017)
10 (materials processing and handling)	-0.0549** (0.017)	0.0781*** (0.015)	0.0921*** (0.015)
11 (mechanical engineering)	-0.1727*** (0.017)	0.0335* (0.015)	0.0436** (0.015)
12 (transport)	-0.0110 (0.020)	-0.0565** (0.018)	-0.0180 (0.018)
13 (consumer goods)	-0.0322 (0.018)	0.0327* (0.016)	0.0704*** (0.016)
14 (civil engineering)	-0.0351* (0.018)	-0.0766*** (0.017)	0.0678*** (0.016)

Table B.1 - Results of alternative probit analyses of REG_IN_PER1 (continued)

cited_firm_country			
DE & CH	-0.1184***	0.3425***	-0.1127***
	(0.011)	(0.016)	(0.010)
Europe	-0.1071***	0.3490***	-0.1062***
	(0.012)	(0.017)	(0.010)
JP	-0.1242***	0.3299***	-0.1247***
	(0.012)	(0.017)	(0.011)
Other	-0.0896***	0.3652***	-0.0906***
	(0.022)	(0.024)	(0.020)
US & CA	-0.1085***	0.3399***	-0.1153***
	(0.010)	(0.015)	(0.008)
Size of dataset	21083067	21083067	21083067
Size of subset per iteration	958321	958321	958321
Number of iterations	22	22	22
Number of unused observations	5	5	5
R^2_{McF}	0.1615	0.0552	0.0024
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
cited_firm_country	AT		
Significance levels	***	0.001	
	**	0.01	
	*	0.05	
Standard errors in parentheses.			

Table B.2 - Results of alternative logit analyses of REG_IN_PER1

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	4.3795	-4.7001	6.2491
	(12.070)	(12.154)	(12.046)
citing_appln_year	-0.0065	-0.0022	-0.0069
	(0.006)	(0.006)	(0.006)
time_lag	0.0042	0.0094	0.0033
	(0.006)	(0.006)	(0.006)
(time_lag)^2	-0.0006*	-0.0008**	-0.0005
	(0.000)	(0.000)	(0.000)
same_industry	3.6633***	———	———
	(0.024)		
same_firm	———	5.5463***	———
		(0.064)	
citing_industry			
2 (electronics)	0.5142***	0.2771***	0.3024***
	(0.070)	(0.070)	(0.070)
3 (instruments)	-0.0676	0.3722***	0.4356***
	(0.057)	(0.057)	(0.057)

Table B.2 - Results of alternative logit analyses of REG_IN_PER1 (continued)

4 (other machinery)	0.6819***	-0.1027	-0.0898
	(0.131)	(0.130)	(0.130)
5 (chemistry)	-0.1837**	0.1467**	0.1743**
	(0.057)	(0.057)	(0.057)
6 (pharmaceuticals and biotechnology)	0.4873***	0.2986***	0.4357***
	(0.061)	(0.061)	(0.061)
7 (agricultural chemistry and machinery)	0.6349***	-0.0401	-0.0322
	(0.122)	(0.121)	(0.121)
8 (material engineering)	-0.1463*	0.0237	0.1036
	(0.060)	(0.060)	(0.060)
9 (chemical and environmental engineering)	0.1699**	0.2371***	0.2682***
	(0.062)	(0.062)	(0.062)
10 (materials processing and handling)	-0.2477***	0.2671***	0.3294***
	(0.054)	(0.054)	(0.054)
11 (mechanical engineering)	-0.6065***	0.0880	0.1564**
	(0.055)	(0.055)	(0.055)
12 (transport)	-0.0552	-0.2308***	-0.0653
	(0.064)	(0.065)	(0.064)
13 (consumer goods)	-0.1197*	0.0636	0.2565***
	(0.058)	(0.058)	(0.058)
14 (civil engineering)	-0.1380*	-0.2742***	0.2524***
	(0.057)	(0.058)	(0.056)
cited_firm_country			
DE & CH	-0.3822***	1.2781***	-0.3966***
	(0.034)	(0.062)	(0.034)
Europe	-0.3571***	1.3030***	-0.3719***
	(0.037)	(0.063)	(0.036)
JP	-0.4194***	1.2344***	-0.4401***
	(0.038)	(0.064)	(0.038)
Other	-0.2844***	1.3551***	-0.3215***
	(0.070)	(0.087)	(0.070)
US & CA	-0.3784***	1.2732***	-0.4032***
	(0.030)	(0.060)	(0.029)
Size of dataset	21083067	21083067	21083067
Size of subset per iteration	958321	958321	958321
Number of iterations	22	22	22
Number of unused observations	5	5	5
R^2_{McF}	0.1625	0.0553	0.0023
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
cited_firm_country	AT		
	***	0.001	
Significance levels	**	0.01	
	*	0.05	
Standard errors in parentheses.			

Table B.3 - Results of alternative complementary log-log analyses of REG_IN_PER1

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	4.3003 (12.046)	-4.5946 (12.128)	6.2457 (12.043)
citing_appln_year	-0.0065 (0.006)	-0.0022 (0.006)	-0.0069 (0.006)
time_lag	0.0042 (0.006)	0.0094 (0.006)	0.0033 (0.006)
(time_lag)^2	-0.0006* (0.000)	-0.0008** (0.000)	-0.0005 (0.000)
same_industry	3.6609*** (0.024)	————	————
same_firm	————	5.5305*** (0.064)	————
citing_industry			
2 (electronics)	0.5131*** (0.070)	0.2768*** (0.070)	0.3023*** (0.070)
3 (instruments)	-0.0680 (0.057)	0.3699*** (0.057)	0.4355*** (0.057)
4 (other machinery)	0.6835*** (0.130)	-0.1039 (0.130)	-0.0899 (0.130)
5 (chemistry)	-0.1837** (0.057)	0.1461* (0.057)	0.1743** (0.057)
6 (pharmaceuticals and biotechnology)	0.4863*** (0.061)	0.2955*** (0.061)	0.4356*** (0.061)
7 (agricultural chemistry and machinery)	0.6356*** (0.122)	-0.0387 (0.121)	-0.0322 (0.121)
8 (material engineering)	-0.1462* (0.060)	0.0224 (0.060)	0.1036 (0.060)
9 (chemical and environmental engineering)	0.1693** (0.062)	0.2357*** (0.062)	0.2681*** (0.062)
10 (materials processing and handling)	-0.2477*** (0.054)	0.2664*** (0.054)	0.3294*** (0.054)
11 (mechanical engineering)	-0.6059*** (0.055)	0.0864 (0.055)	0.1564** (0.055)
12 (transport)	-0.0553 (0.064)	-0.2315*** (0.064)	-0.0653 (0.064)
13 (consumer goods)	-0.1196* (0.058)	0.0615 (0.058)	0.2565*** (0.058)
14 (civil engineering)	-0.1379* (0.057)	-0.2735*** (0.057)	0.2524*** (0.056)
cited_firm_country			
DE & CH	-0.3815*** (0.034)	1.2779*** (0.062)	-0.3965*** (0.034)
Europe	-0.3565*** (0.037)	1.3028*** (0.063)	-0.3718*** (0.036)

Table B.3 - Results of alternative complementary log-log analyses of REG_IN_PER1 (continued)

JP	-0.4188***	1.2342***	-0.4400***
	(0.038)	(0.064)	(0.038)
Other	-0.2839***	1.3548***	-0.3214***
	(0.070)	(0.087)	(0.070)
US & CA	-0.3779***	1.2731***	-0.4031***
	(0.030)	(0.060)	(0.029)
Size of dataset	21083067	21083067	21083067
Size of subset per iteration	958321	958321	958321
Number of iterations	22	22	22
Number of unused observations	5	5	5
R^2_{McF}	0.1625	0.0553	0.0023
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
cited_firm_country	AT		
Significance levels	***	0.001	
	**	0.01	
	*	0.05	
Standard errors in parentheses.			

B.2 REG_IN_PER2

Table B.4 - Results of alternative probit analyses of REG_IN_PER2

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	52.3269***	47.7434***	46.8481***
	(6.901)	(6.264)	(6.150)
citing_appln_year	-0.0280***	-0.0256***	-0.0250***
	(0.003)	(0.003)	(0.003)
time_lag	-0.0007	-0.0011	-0.0010
	(0.003)	(0.003)	(0.003)
(time_lag)^2	0.0000	0.0000	0.0000
	(0.000)	(0.000)	(0.000)
same_industry	1.1113***	————	————
	(0.012)		
same_firm	————	1.7831***	————
		(0.040)	
citing_industry			
2 (electronics)	0.1081***	-0.0649*	-0.0687*
	(0.032)	(0.029)	(0.028)
3 (instruments)	-0.0105	-0.0543**	-0.0528*
	(0.023)	(0.021)	(0.021)
4 (other machinery)	0.3813***	0.0287	0.0233
	(0.052)	(0.047)	(0.046)

Table B.4 - Results of alternative probit analyses of REG_IN_PER2 (continued)

5 (chemistry)	0.1143***	-0.0553*	-0.0484
	(0.029)	(0.026)	(0.026)
6 (pharmaceuticals and biotechnology)	0.2563***	-0.0610	-0.0624
	(0.052)	(0.046)	(0.046)
7 (agricultural chemistry and machinery)	0.2823***	-0.0596	-0.0589
	(0.049)	(0.044)	(0.043)
8 (material engineering)	0.1222***	-0.0732**	-0.0715**
	(0.030)	(0.027)	(0.026)
9 (chemical and environmental engineering)	0.1334***	-0.0463	-0.0424
	(0.031)	(0.027)	(0.027)
10 (materials processing and handling)	-0.0540*	-0.0745***	-0.0704***
	(0.024)	(0.021)	(0.021)
11 (mechanical engineering)	-0.1478***	-0.0093	-0.0044
	(0.021)	(0.019)	(0.019)
12 (transport)	0.0774**	-0.0732**	-0.0561*
	(0.027)	(0.024)	(0.024)
13 (consumer goods)	0.0676**	-0.0440	-0.0319
	(0.026)	(0.023)	(0.022)
14 (civil engineering)	-0.0324	-0.1585***	-0.0751***
	(0.024)	(0.022)	(0.021)
cited_firm_country			
DE & CH	-0.0352	0.2760***	-0.0429*
	(0.022)	(0.028)	(0.019)
Europe	-0.0359	0.2728***	-0.0461*
	(0.024)	(0.029)	(0.021)
JP	-0.0208	0.2783***	-0.0398
	(0.023)	(0.029)	(0.020)
Other	-0.0474	0.2591***	-0.0593
	(0.043)	(0.043)	(0.038)
US & CA	-0.0244	0.2784***	-0.0401*
	(0.020)	(0.027)	(0.017)
Size of dataset	5590421	5590421	5590421
Size of subset per iteration	931736	931736	931736
Number of iterations	6	6	6
Number of unused observations	5	5	5
R^2_{McF}	0.1671	0.0306	0.0019
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
cited_firm_country	AT		
	***	0.001	
Significance levels	**	0.01	
	*	0.05	
Standard errors in parentheses.			

Table B.5 - Results of alternative logit analyses of REG_IN_PER2

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	165.2457***	172.1654***	168.4719***
	(21.542)	(21.564)	(21.490)
citing_appln_year	-0.0870***	-0.0901***	-0.0877***
	(0.011)	(0.011)	(0.011)
time_lag	-0.0046	-0.0062	-0.0031
	(0.010)	(0.010)	(0.010)
(time_lag)^2	0.0001	0.0002	0.0001
	(0.000)	(0.000)	(0.000)
same_industry	3.7513***	———	———
	(0.043)		
same_firm	———	5.3334***	———
		(0.119)	
citing_industry			
2 (electronics)	0.3628***	-0.2239*	-0.2418*
	(0.100)	(0.100)	(0.099)
3 (instruments)	-0.0674	-0.1908**	-0.1873**
	(0.072)	(0.072)	(0.072)
4 (other machinery)	1.4127***	0.0926	0.0748
	(0.161)	(0.159)	(0.158)
5 (chemistry)	0.3890***	-0.1887*	-0.1702
	(0.090)	(0.089)	(0.089)
6 (pharmaceuticals and biotechnology)	0.9655***	-0.2165	-0.2297
	(0.166)	(0.164)	(0.164)
7 (agricultural chemistry and machinery)	1.0865***	-0.1940	-0.2126
	(0.154)	(0.152)	(0.152)
8 (material engineering)	0.3848***	-0.2598**	-0.2525**
	(0.093)	(0.092)	(0.092)
9 (chemical and environmental engineering)	0.4327***	-0.1697	-0.1509
	(0.095)	(0.095)	(0.094)
10 (materials processing and handling)	-0.1891*	-0.2494***	-0.2482***
	(0.074)	(0.074)	(0.074)
11 (mechanical engineering)	-0.4849***	-0.0390	-0.0178
	(0.065)	(0.064)	(0.064)
12 (transport)	0.2445**	-0.2764***	-0.1981*
	(0.084)	(0.083)	(0.083)
13 (consumer goods)	0.2087**	-0.1614*	-0.1123
	(0.079)	(0.078)	(0.078)
14 (civil engineering)	-0.1093	-0.5667***	-0.2633***
	(0.075)	(0.075)	(0.073)
cited_firm_country			
DE & CH	-0.1404*	0.9992***	-0.1484*
	(0.067)	(0.104)	(0.067)
Europe	-0.1451*	0.9866***	-0.1624*
	(0.072)	(0.107)	(0.072)

Table B.5 - Results of alternative logit analyses of REG_IN_PER2 (continued)

JP	-0.1170	1.0101***	-0.1384
	(0.072)	(0.106)	(0.071)
Other	-0.1889	0.9363***	-0.2115
	(0.135)	(0.156)	(0.134)
US & CA	-0.1143	1.0083***	-0.1399*
	(0.061)	(0.099)	(0.059)
Size of dataset	5590421	5590421	5590421
Size of subset per iteration	931736	931736	931736
Number of iterations	6	6	6
Number of unused observations	5	5	5
R^2_{McF}	0.1684	0.0308	0.0019
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
cited_firm_country	AT		
Significance levels	***	0.001	
	**	0.01	
	*	0.05	
Standard errors in parentheses.			

Table B.6 - Results of alternative complementary log-log analyses of REG_IN_PER2

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	164.8711***	172.0717***	168.4184***
	(21.490)	(21.527)	(21.483)
citing_appln_year	-0.0868***	-0.0901***	-0.0877***
	(0.011)	(0.011)	(0.011)
time_lag	-0.0046	-0.0064	-0.0031
	(0.010)	(0.010)	(0.010)
(time_lag)^2	0.0001	0.0002	0.0001
	(0.000)	(0.000)	(0.000)
same_industry	3.7484***	————	————
	(0.043)		
same_firm	————	5.3138***	————
		(0.118)	
citing_industry			
2 (electronics)	0.3620***	-0.2237*	-0.2417*
	(0.100)	(0.099)	(0.099)
3 (instruments)	-0.0675	-0.1908**	-0.1872**
	(0.072)	(0.072)	(0.072)
4 (other machinery)	1.4127***	0.0923	0.0747
	(0.160)	(0.158)	(0.158)
5 (chemistry)	0.3882***	-0.1880*	-0.1702
	(0.089)	(0.089)	(0.089)

Table B.6 - Results of alternative complementary log-log analyses of REG_IN_PER2 (continued)

6 (pharmaceuticals and biotechnology)	0.9655***	-0.2141	-0.2296
	(0.165)	(0.164)	(0.164)
7 (agricultural chemistry and machinery)	1.0870***	-0.1922	-0.2125
	(0.153)	(0.152)	(0.152)
8 (material engineering)	0.3836***	-0.2596**	-0.2525**
	(0.093)	(0.092)	(0.092)
9 (chemical and environmental engineering)	0.4316***	-0.1699	-0.1509
	(0.095)	(0.094)	(0.094)
10 (materials processing and handling)	-0.1890*	-0.2484***	-0.2481***
	(0.074)	(0.074)	(0.074)
11 (mechanical engineering)	-0.4842***	-0.0395	-0.0178
	(0.064)	(0.064)	(0.064)
12 (transport)	0.2438**	-0.2775***	-0.1981*
	(0.083)	(0.083)	(0.083)
13 (consumer goods)	0.2081**	-0.1617*	-0.1123
	(0.078)	(0.078)	(0.078)
14 (civil engineering)	-0.1093	-0.5663***	-0.2632***
	(0.074)	(0.075)	(0.073)
cited_firm_country			
DE & CH	-0.1405*	0.9991***	-0.1483*
	(0.067)	(0.104)	(0.067)
Europe	-0.1452*	0.9865***	-0.1623*
	(0.072)	(0.107)	(0.072)
JP	-0.1174	1.0101***	-0.1384
	(0.072)	(0.106)	(0.071)
Other	-0.1893	0.9362***	-0.2115
	(0.135)	(0.156)	(0.134)
US & CA	-0.1145	1.0082***	-0.1398*
	(0.060)	(0.099)	(0.059)
Size of dataset	5590421	5590421	5590421
Size of subset per iteration	931736	931736	931736
Number of iterations	6	6	6
Number of unused observations	5	5	5
R^2_{McF}	0.1684	0.0308	0.0019
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
cited_firm_country	AT		
	***	0.001	
Significance levels	**	0.01	
	*	0.05	
Standard errors in parentheses.			

B.3 REG_OUT_PER1

Table B.7 - Results of alternative probit analyses of REG_OUT_PER1

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	38.5849**	16.7315	25.7210*
	(11.789)	(10.956)	(10.303)
citing_appln_year	-0.0211***	-0.0102	-0.0144**
	(0.006)	(0.005)	(0.005)
time_lag	0.1106***	0.1042***	0.0911***
	(0.011)	(0.010)	(0.009)
(time_lag)^2	-0.0135***	-0.0130***	-0.0117***
	(0.001)	(0.001)	(0.001)
same_industry	1.1760***	————	————
	(0.016)		
same_firm	————	1.8248***	————
		(0.034)	
citing_industry			
2 (electronics)	0.1807***	-0.0012	0.0102
	(0.046)	(0.042)	(0.041)
3 (instruments)	0.0067	-0.0085	0.0162
	(0.040)	(0.035)	(0.034)
4 (other machinery)	0.3080***	0.0387	0.0434
	(0.089)	(0.083)	(0.080)
5 (chemistry)	-0.0287	0.0159	0.0156
	(0.040)	(0.036)	(0.034)
6 (pharmaceuticals and biotechnology)	0.1236*	0.0571	0.0501
	(0.049)	(0.045)	(0.042)
7 (agricultural chemistry and machinery)	0.2597***	0.0051	-0.0028
	(0.071)	(0.068)	(0.064)
8 (material engineering)	0.0665	-0.0285	0.0230
	(0.043)	(0.039)	(0.037)
9 (chemical and environmental engineering)	0.0906*	0.0223	0.0313
	(0.043)	(0.040)	(0.038)
10 (materials processing and handling)	-0.0840*	0.0134	0.0304
	(0.037)	(0.033)	(0.032)
11 (mechanical engineering)	-0.1490***	-0.0092	0.0134
	(0.035)	(0.032)	(0.031)
12 (transport)	-0.0061	-0.0176	0.0191
	(0.039)	(0.035)	(0.034)
13 (consumer goods)	-0.0400	0.0488	0.0775*
	(0.037)	(0.033)	(0.032)
14 (civil engineering)	-0.1017**	-0.0687	0.0527
	(0.038)	(0.035)	(0.033)

Table B.7 - Results of alternative probit analyses of REG_OUT_PER1 (continued)

citing_firm_country			
DE & CH	-0.0598**	0.4032***	-0.0584**
	(0.022)	(0.029)	(0.019)
Europe	-0.0664**	0.3956***	-0.0642**
	(0.025)	(0.031)	(0.022)
JP	-0.0674*	0.3802***	-0.0696**
	(0.027)	(0.032)	(0.024)
Other	-0.0277	0.4281***	-0.0333
	(0.044)	(0.044)	(0.039)
US & CA	-0.0660**	0.3875***	-0.0715***
	(0.020)	(0.028)	(0.018)
Size of dataset	1939519	1939519	1939519
Size of subset per iteration	969759	969759	969759
Number of iterations	2	2	2
Number of unused observations	1	1	1
R^2_{McF}	0.1873	0.0921	0.0057
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
citing_firm_country	AT		
Significance levels	***	0.001	
	**	0.01	
	*	0.05	
Standard errors in parentheses.			

Table B.8 - Results of alternative logit analyses of REG_OUT_PER1

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	97.2298**	43.4477	88.2751**
	(34.454)	(34.534)	(34.200)
citing_appln_year	-0.0529**	-0.0260	-0.0477**
	(0.017)	(0.017)	(0.017)
time_lag	0.3246***	0.3308***	0.3113***
	(0.032)	(0.032)	(0.032)
(time_lag)^2	-0.0416***	-0.0430***	-0.0408***
	(0.004)	(0.004)	(0.004)
same_industry	3.7311***	———	———
	(0.052)		
same_firm	———	5.2979***	———
		(0.106)	
citing_industry			
2 (electronics)	0.5648***	-0.0133	0.0325
	(0.137)	(0.136)	(0.136)
3 (instruments)	-0.0286	-0.0648	0.0560
	(0.116)	(0.115)	(0.115)

Table B.8 - Results of alternative logit analyses of REG_OUT_PER1 (continued)

4 (other machinery)	1.1023***	0.1550	0.1468
	(0.269)	(0.267)	(0.267)
5 (chemistry)	-0.0985	0.0201	0.0513
	(0.115)	(0.115)	(0.115)
6 (pharmaceuticals and biotechnology)	0.3790**	0.1350	0.1627
	(0.141)	(0.142)	(0.141)
7 (agricultural chemistry and machinery)	0.9015***	0.0486	-0.0157
	(0.220)	(0.218)	(0.217)
8 (material engineering)	0.1904	-0.1470	0.0760
	(0.123)	(0.124)	(0.123)
9 (chemical and environmental engineering)	0.2187	0.0671	0.1048
	(0.127)	(0.127)	(0.126)
10 (materials processing and handling)	-0.3268**	0.0091	0.1001
	(0.107)	(0.107)	(0.107)
11 (mechanical engineering)	-0.5055***	-0.0646	0.0436
	(0.103)	(0.103)	(0.102)
12 (transport)	-0.0419	-0.1091	0.0646
	(0.113)	(0.113)	(0.113)
13 (consumer goods)	-0.1308	0.0695	0.2584*
	(0.106)	(0.106)	(0.105)
14 (civil engineering)	-0.3104**	-0.3050**	0.1774
	(0.109)	(0.111)	(0.108)
citing_firm_country			
DE & CH	-0.1925**	1.4092***	-0.1938**
	(0.064)	(0.105)	(0.063)
Europe	-0.2103**	1.3919***	-0.2113**
	(0.072)	(0.111)	(0.072)
JP	-0.2264**	1.3261***	-0.2298**
	(0.079)	(0.115)	(0.079)
Other	-0.1130	1.4915***	-0.1110
	(0.128)	(0.152)	(0.127)
US & CA	-0.2367***	1.3593***	-0.2361***
	(0.059)	(0.102)	(0.058)
Size of dataset	1939519	1939519	1939519
Size of subset per iteration	969759	969759	969759
Number of iterations	2	2	2
Number of unused observations	1	1	1
R^2_{McF}	0.1882	0.0924	0.0058
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
citing_firm_country	AT		
	***	0.001	
Significance levels	**	0.01	
	*	0.05	
Standard errors in parentheses.			

Table B.9 - Results of alternative complementary log-log analyses of REG_OUT_PER1

	Alternative specification 1	Alternative specification 2	Alternative specification 3
Intercept	96.4235**	42.1899	88.2187**
	(34.296)	(34.324)	(34.179)
citing_appln_year	-0.0525**	-0.0254	-0.0477**
	(0.017)	(0.017)	(0.017)
time_lag	0.3232***	0.3284***	0.3111***
	(0.032)	(0.032)	(0.032)
(time_lag)^2	-0.0415***	-0.0428***	-0.0408***
	(0.004)	(0.004)	(0.004)
same_industry	3.7249***	———	———
	(0.052)		
same_firm	———	5.2713***	———
		(0.105)	
citing_industry			
2 (electronics)	0.5619***	-0.0143	0.0324
	(0.136)	(0.136)	(0.136)
3 (instruments)	-0.0294	-0.0681	0.0560
	(0.115)	(0.115)	(0.115)
4 (other machinery)	1.1039***	0.1578	0.1467
	(0.267)	(0.266)	(0.266)
5 (chemistry)	-0.0983	0.0171	0.0513
	(0.114)	(0.114)	(0.114)
6 (pharmaceuticals and biotechnology)	0.3773**	0.1306	0.1626
	(0.141)	(0.141)	(0.140)
7 (agricultural chemistry and machinery)	0.9003***	0.0512	-0.0158
	(0.218)	(0.217)	(0.217)
8 (material engineering)	0.1894	-0.1514	0.0759
	(0.123)	(0.123)	(0.123)
9 (chemical and environmental engineering)	0.2165	0.0657	0.1047
	(0.126)	(0.126)	(0.126)
10 (materials processing and handling)	-0.3266**	0.0060	0.1000
	(0.107)	(0.107)	(0.107)
11 (mechanical engineering)	-0.5044***	-0.0681	0.0436
	(0.102)	(0.102)	(0.102)
12 (transport)	-0.0421	-0.1131	0.0645
	(0.113)	(0.113)	(0.113)
13 (consumer goods)	-0.1303	0.0621	0.2583*
	(0.105)	(0.105)	(0.105)
14 (civil engineering)	-0.3092**	-0.3089**	0.1773
	(0.108)	(0.111)	(0.108)
citing_firm_country			
DE & CH	-0.1919**	1.4085***	-0.1937**
	(0.063)	(0.105)	(0.063)
Europe	-0.2096**	1.3917***	-0.2112**
	(0.072)	(0.111)	(0.072)

Table B.9 - Results of alternative complementary log-log analyses of REG_OUT_PER1 (continued)

JP	-0.2258**	1.3250***	-0.2297**
	(0.079)	(0.115)	(0.079)
Other	-0.1132	1.4910***	-0.1109
	(0.127)	(0.152)	(0.127)
US & CA	-0.2363***	1.3589***	-0.2360***
	(0.058)	(0.102)	(0.058)
Size of dataset	1939519	1939519	1939519
Size of subset per iteration	969759	969759	969759
Number of iterations	2	2	2
Number of unused observations	1	1	1
R^2_{McF}	0.1882	0.0924	0.0058
Base categories for dummy variables			
citing_industry	1 (electrical engineering)		
citing_firm_country	AT		
	***	0.001	
Significance levels	**	0.01	
	*	0.05	
Standard errors in parentheses.			