MASTER

The Spatial Stickiness of Knowledge Diffusion Within -and Around the Ecosystems of Brainport Eindhoven and the High Tech Campus Eindhoven

van Griensven, Jesper

Award date:
2023

The Spatial Stickiness of Knowledge Diffusion Within -and Around the Ecosystems of Brainport Eindhoven and the High Tech Campus Eindhoven

Master Thesis

Author: Jesper Hendricus Petrus van Griensven
E-mail: j.h.p.v.griensven@student.tue.nl
Student number: 1266586

Supervisors:

First academic supervisor:	Dr. Emilio Raiteri, IE&IS
Second academic supervisor:	Dr. Elena Mas Tur, IE&IS
Third academic supervisor:	Dr. Bert Sadowski, IE&IS

Faculty:	Industrial Engineering & Innovation Sciences (IE&IS)
Group:	Technology, Innovation & Society (TIS)
MSc program:	Innovation Sciences (IS)

Version 1.0

Eindhoven, June 26th, 2023

Abstract

Purpose – In this research, it is studied whether the diffusion of knowledge is spatially stickier within -and around ecosystems. Since ecosystems do have proximity-related advantages, which contribute to the increased diffusion and creation of knowledge, it is hypothesized that the diffusion of knowledge is spatially stickier within -and around these regions.

Methodology – With the use of the dataset by de Rassenfosse et al. (2019), patent data was collected about two ecosystems; Brainport Eindhoven (i) and the High Tech Campus Eindhoven (HTCE) (ii), for the years 2003 – 2014. A set of control patents was created by matching patents elsewhere from the Netherlands and Belgium to the ecosystem patents on a similarity in International Patent Classification (IPC) and year of filing. With the use of Ordinary Least Squares (OLS) regression, the distances of citation links originating from cited ecosystem -and control patents were compared.

Results – The results do show that the diffusion of knowledge is spatially stickier within -and around ecosystems, however, the effects of localization are not explained by the diffusion of knowledge within the ecosystems, but rather driven by firms outside the ecosystems.

Contribution – This research complements the current academic literature on knowledge diffusion and ecosystems by conducting empirical research on how knowledge is diffused within -and around ecosystems.

Keywords – Ecosystems – Knowledge diffusion – Proximity – High Tech Campus Eindhoven – Brainport Eindhoven – Localization – Knowledge flows – Patent citations – Geography

Acknowledgments

First and foremost, I would like to thank my first supervisor, Emilio Raiteri, who was of great value to me throughout the process. Emilio helped me shape the scope of this research and came up with the idea to focus this research on ecosystems, instead of knowledge diffusion in general, which enhanced the relevance and novelty of this research. Furthermore, Emilio helped me to shape my ideas into scientific research and taught me how to do academic research.

Secondly, I want to thank Mr. Benjamin Büttner, who taught me how to use Microsoft SQL and was always willing to help me. Thirdly, I want to thank Mrs. Elena Mas Tur and Mr. Bert Sadowski for their willingness to be my second and third supervisor. Fourthly, I want to thank Mr. Rudi Bekkers who gave me access to the university's remote computer which enabled me to do the empirical part of this research. I want to thank Mr. Paul van Son of the High Tech Campus Site Management, and Mr. Bart Jan Niestadt of V.O. Patents, for helping me orient this research. Lastly, I would like to thank my family and friends who were there for me all the time and who kept me motivated and going throughout the thesis.

Summary

Introduction

For years, both the city -and metropole of Eindhoven, better known as Brainport Eindhoven, are considered the booming heart of innovation and the high-tech industry in the Netherlands. Moreover, Brainport Eindhoven is the Netherlands' most valuable region, representing 11% of the total added value of the Dutch industry (Brainport Eindhoven, 2022). The success of Brainport Eindhoven is narrated by open innovation, ecosystems, and collaboration among multiple sectors and disciplines. The collaborations and interdependencies led to the creation of a non-hierarchical system, which scholars later started to define as "ecosystems" (Jacobides et al., 2018). Within Brainport Eindhoven, multiple ecosystems can be found, e.g., the High Tech Campus Eindhoven (HTCE). The HTCE is often referred to as "Europe's smartest square kilometer" since the research labs of patenting-intensive companies like Philips, NXP, and Signify are all located at the campus. With the HTCE and other ecosystems and institutions, Brainport Eindhoven is one big ecosystem, ranking No. 7 on the list of the world's leading tech ecosystems, and being the highest-ranked ecosystem in Europe after Oxford and Cambridge (Dealroom, 2022).

According to Robertson et al. (2023), ecosystems do have proximity-related advantages which contribute to the increased diffusion and creation of knowledge. Furthermore, Robertson et al. posit that knowledge is "sticky" within the ecosystem and diffuses locally within the ecosystem. Therefore, it is likely to assume that knowledge diffusion is somewhat peculiar concerning ecosystems. This led to the main research question of this research: Is the diffusion of knowledge within -and around ecosystems spatially stickier compared to non-ecosystems?

For more than three decades, patent citations are used by scholars as a proxy to measure knowledge diffusion and multiple studies have found that patent citations are bounded in space (Abramo et al., 2020; Almeida & Kogut, 1999; Belenzon & Schankerman, 2013; Breschi & Lissoni, 2009; Jaffe et al., 1993; Thompson & Fox-Kean, 2005; Zucker et al., 1998). With the use of patenting data from the HTCE and Brainport Eindhoven, the RQ described above will be anticipated and is formulated into a second sub-research question: Are patent citations, originating from cited High Tech Campus Eindhoven (HTCE) -and Brainport Eindhoven patents, more localized compared to patent citations originating from cited 'non-ecosystem' patents?

Literature Review

The literature review in this research is divided into two sub-chapters. The first part focuses on ecosystems whereas the second part focuses on patent citations as a proxy for knowledge diffusion. In the history of innovation, the emergence of ecosystems was facilitated by a paradigm shift in innovation in which the 'post World War II' closed-loop paradigm, was substituted by the open innovation paradigm (Chesbrough, 2006). In the current literature, four different and commonly studied ecosystems can be distinguished: business, innovation, entrepreneurial, and knowledge-based ecosystems (Cobben et al., 2022). In its essence,

Brainport Eindhoven can be categorized as an entrepreneurial ecosystem, since the mobilization of talent, start-ups, and networks are the key characteristics of the ecosystem. The HTCE can be considered a knowledge-based ecosystem (Borgh et al., 2012).

The diffusion of knowledge is facilitated by different forms of proximity: Cognitive proximity, organizational proximity, social proximity, institutional proximity, and geographical proximity (Boschma, 2005). According to Boschma (2005), geographical proximity is neither a necessary nor a sufficient condition for interactive social learning. Besides geographical proximity, the other forms of proximity are strongly represented within the ecosystems of the HTCE and Brainport Eindhoven. These findings are in line with the results found in the paper by Robertson et al. (2023), in which it is argued that ecosystems do have proximity-related advantages, and thereby the idea is strengthened that knowledge is stickier within -and around the HTCE and Brainport Eindhoven.

The second part of the literature review focuses on patent citations as a proxy for knowledge diffusion. Over the years, the use of this proxy has been utilized extensively. One of the major limitations of this proxy is that most of the patent citations are added by examiners at the patent office, as found by Alcácer & Gittelman (2006) and Criscuolo & Verspagen (2008). Besides the influence and biases of examiner citations, Corsino et al. (2019) found that measurement errors of patent citations, originating from firms, are rooted in firms' incentives to cite prior art. Although the limitations of patent citations as a proxy for knowledge diffusion, in the seminal paper by Jaffe et al. (1993), it was found that patent citations are bounded in space. In the paper by Breschi & Lissoni (2009), it was found that the localization of patent citations is mediated by the mobility of inventors.

Methodology

To study the effect of ecosystems, the aim was to compare the distance of citation links that originate from cited ecosystem patents to the distance of the controls. With the use of the dataset by de Rassenfosse et al. (2019), a dataset that included geographical data of 18.8 million first filings from the years 1980-2014, patent data about the HTCE and Brainport was collected. The dataset by de Rassenfosse et al. (2019) included the the latitude and longitude of each patent. Since the HTCE was founded in 2003, it was chosen to collect the ecosystems from the years 2003 – 2010. Consequently, since the dataset by de Rassenfosse et al. (2019) included data up to the year 2014, the citation lag was set at a minimum of four years. Concerning the control patents, a group of patents similar to the ecosystem patents was matched on a similarity in International Patent Classification (IPC) and year of application. Besides the dataset by de Rassenfosse et al. (2019), PATSTAT, the patent database of the European Patent Office (EPO), was used to collect additional information of the patents. The set of control patents was collected from the Netherlands and Belgium, excluding the region of Brainport Eindhoven. From the pool of possible control patents, for each ecosystem patent, a control patent was matched at random. Microsoft SQL was used to treat the data. Eventually, using the formula of Haversine, the distances of all citation links were calculated.

To answer the research question, three different hypotheses were formulated, and tested by three different distance metrics. In short, first, it was tested whether patent citations that originate from cited ecosystem patents are in general more localized. To test this, for each patent, the average of all citation links and the shortest citation link, was calculated. Secondly, concerning the second hypothesis, it was tested whether the diffusion of knowledge within the ecosystems is spatially stickier. To test this, for each patent, it was registered whether the patent had a citation link within the range of 30 -and 50 kilometers. These ranges represent more or less the borders of Brainport Eindhoven. Thirdly, it was tested whether patent citations that originate from cited ecosystem patents are more localized within -and around the ecosystems. To test this, for each patent, it was registered whether the patent had a citation link within the ranges of 100, 200, and 500 kilometers. The range of 200 kilometers represents the area of the Benelux. The hypotheses were tested using multiple descriptive statistics and OLS regression.

Results

The descriptive statistics show that the distribution of applicants is skewed in the ecosystem groups. Concerning the HTCE patents, approximately 95% of all the patents are filed by Philips and NXP. Concerning the Brainport group, approximately 80% of the patents are filed by Philips, NXP, and ASML. With the use of a t-test, for each distance metric, the means of the ecosystem and control groups were compared. The results of the t-test are rather counterintuitive as the effects of localization are supported for the control groups. By plotting the density of patents over the distance of the shortest citation link, it is found that the distributions are highly skewed when citation links with no spatial distance (zero kilometers) are included. Next to that, when these citation links are excluded, the graphs do show that within a range of 50 kilometers, the share of ecosystem patents is more frequent compared to their peer control patents.

To control for the effect of other predictor variables, the hypotheses were tested using OLS regression. The results do show that, for both ecosystem groups, patent citations that do originate from cited ecosystem patents are in general more localized. Secondly, within a range of 50 kilometers, the results do not show that patent citations originating from cited ecosystem patents are localized. Thirdly, on the other hand, the results do show that patent citations originating from cited ecosystems patents are more localized between the ranges of 100 – 500 kilometers. Although the results do favor localization, it is found that the models were very sensitive to the following three circumstances: the fixed firm effect (i), the in -or exclusion of citation links with a distance of zero kilometers (ii), the in -or exclusion of patents with only one citation link (iii). The fixed firm effect anticipates the influence of the individual firm on the relationship that was studied in this research.

Conclusion and Discussion

Based on the results, it can be concluded that H1 is partially accepted since for both ecosystem groups the distance of the shortest citation link of ecosystem patents was found to be shorter compared to the control patents, whereas for the average distance of citation links, a significant result was only found in the HTCE group. Although the density plots indicated a higher density of ecosystem systems within a range of 50 kilometers, no statistical support was found in favor

of hypothesis 2. Therefore, H2 is rejected. With respect to the third hypothesis, it was found that, for both ecosystem groups, patent citations originating from cited ecosystem patents were estimated to have more citation links within the ranges of 100 – 500 kilometers. Thereby, H3 is accepted. Concerning the main research question, it can be concluded that the diffusion of knowledge is spatially stickier within -and around ecosystems, however, the effects of localization are not explained by the diffusion of knowledge within the ecosystems, but rather driven by firms outside the ecosystems.

The results and conclusions described above must be interpreted considering the following factors. First of all, the localization effects that were found in this research are only significant when it is controlled for the fixed firm effect (i), when citation links with no spatial distance are excluded (ii), and when patents with only one citation link are included in the regression models (iii). Secondly, patent citations are not a perfect and rather 'noisy' proxy to measure knowledge diffusion (Corsino et al., 2019). All in all, although the imperfections and sensitivities of the method and the regression model, the findings in this research indicate that the diffusion of knowledge is spatially stickier within -and around ecosystems.

Table of Contents

1. Introduction

For years, both the city -and metropole of Eindhoven, better known as Brainport Eindhoven, are considered the booming heart of innovation and the high-tech industry in the Netherlands. Besides Philips, the founding father of high-tech Eindhoven, companies like ASML, NXP, Signify, DAF, and VDL Group are among the highest in the country's list of companies' private R&D expenditures. Together, at a national level, they represent 25,3% of the total private R&D expenditures (Brainport Eindhoven, 2022). Additionally, back in 2021, Philips, Signify, ASML, and NXP were accountable for 40% of all patent applications in the Netherlands (Brainport Eindhoven, 2022). For more than a decade, the economic growth of Brainport Eindhoven is superior to the national rates of economic growth (Brainport Eindhoven, 2022). All in all, Brainport Eindhoven represents about 5,2% of the national economy (Brainport Eindhoven, 2022). With respect to the national industry, Brainport Eindhoven represents a total added value of 11%. Consequently, Brainport Eindhoven is the most valuable region in the Netherlands concerning industry (Brainport Eindhoven, 2022).

The outstanding numbers of Brainport Eindhoven are not just a 'coincidence'. These numbers are not the result of a handful of successful high-tech firms, that operate in isolation, and appear to be located within Brainport Eindhoven. No, there is more than that, the story of Eindhoven and Brainport Eindhoven is a rather interesting one, narrated by open innovation, ecosystems, and collaboration among multiple disciplines and sectors.

More than one century ago, back in 1892, Philips built its first small factory to produce lightning bulbs (Romme, 2022). With a lack of adequate infrastructure in Eindhoven and the surrounding villages, Philips started investing in new neighborhoods, schools, and other social services. With an increasing demand for technologically skilled personnel, Philips helped initiate the arrival of a new university which later became known as 'Eindhoven University of Technology'. Over the years, in the shadow of Philips, a fruitful soil was created for entrepreneurship. In 2006, Brainport Eindhoven was founded, an institution that bridges entrepreneurship, knowledge institutions, and local municipalities within -and around the Brainport region.

Over the years, the collaborations and interdependencies led to the creation of a non-hierarchical system, which scholars later started to define as "ecosystems" (Jacobides et al., 2018). Within Brainport Eindhoven, multiple ecosystems can be found, e.g., the High Tech Campus Eindhoven (HTCE), the Brainport Industries Campus (BIC), and the Automotive Campus in Helmond. These three ecosystems all operate in close proximity. With the presence of multiple ecosystems, the Technical University of Eindhoven, and other intuitions, altogether, Brainport Eindhoven is one big ecosystem, ranking No. 7 on the list of the world's leading tech ecosystems, and being the highest-ranked ecosystem in Europe after Oxford and Cambridge (Dealroom, 2022).

A good example of the impact of the ecosystem on the region is the high success rate of startups within Brainport Eindhoven. Every year, in -and around Eindhoven, the 10 most promising start-ups of the region are awarded the Gerard & Anton award. In the Netherlands, the average success rate of start-ups transforming from a start-up into a scale-up is 16 percent (Van Leest et al., 2022). Over the years 2014-2020, among the 70 most promising Brainport start-ups, 52 of the start-ups managed to sustain and develop a successful scale-up (Van Leest et al., 2022). In an interview with the award-winning start-ups, the start-ups addressed the importance of the ecosystem by mentioning platforms like HTCE, TU Eindhoven, The Gate, and Innovation Space, acting as facilitators for start-ups to mobilize their operations (Van Leest et al., 2022). Both The Gate and Innovation Space are platforms, located at the TU Eindhoven, to support high-tech start-ups in the early stage of their development. All in all, even though the abundant numbers of the success rate of start-ups in the Brainport region are somewhat biased since only the most promising start-ups were selected, these numbers do display the strength of the region and its ecosystem.

As mentioned above, within the Brainport region, multiple sub-ecosystems can be found. Among these systems, the HTCE can be considered the most outstanding one. The HTCE is an innovative cluster of around 220 high-tech companies located in the center of the Brainport Region. The campus was initially a Philips research lab, but in 2002, Philips opened the campus to other companies to create an environment of open innovation. The campus can be best described as a knowledge-based ecosystem (Borgh et al., 2012). In such a system, knowledge-intensive firms, operating in close proximity, depend on one another for the efficiency and effectiveness of their operations. The HTCE is often referred to as “Europe’s smartest square kilometer” since the research labs of patenting-intensive companies like Philips, NXP, and Signify are all located at the campus, whereas ASML is located just around the corner in Veldhoven. The importance of the HTCE does not remain unnoticed by the national top: the Dutch King visited the campus two times in the last two years.

1.1 Ecosystems and Knowledge Diffusion

With the phenomena of ecosystems being embedded in recent literature on innovation, the acknowledgment of such regions hints that these regions are somewhat peculiar. With the example of the HTCE and Brainport Eindhoven, in which actors operate and collaborate intensively in close proximity, it intuitively makes sense to assume that ecosystems are peculiar with respect to knowledge diffusion.

In general, the diffusion of knowledge is considered to be one of the fundamental facilitators of innovation. The diffusion of knowledge is facilitated by geographical proximity. However, simply “being there” is not enough for knowledge to diffuse (Boschma, 2005; Capello & Varga, 2013; Paci et al., 2014). Boschma (2005) found that geographical proximity is neither a necessary nor a sufficient condition for knowledge to diffuse, it rather complements proximities such as cognitive, organizational, social -and institutional proximity.

According to Robertson et al. (2023), ecosystems do have proximity-related advantages which contribute to the increased diffusion and creation of knowledge. Moreover, Robertson et al. found a positive relationship between knowledge diffusion and innovation performance within ecosystems, positing that actors should create, diffuse and acquire knowledge within the ecosystem. Here, Robertson et al. introduce the idea that knowledge is "sticky" within the ecosystem and that knowledge diffuses locally within the ecosystem.

1.1.1 Knowledge Diffusion and Patent Citations

As can be derived from the previous paragraphs, the ecosystem of the Brainport region, and its sub-systems, are characterized by cooperation, the exchange of resources, and the abundant innovative output in the format of patent applications. The innovativeness of a region or firm is often measured by its R&D expenditures (Hall et al., 2005; Trajtenberg M, 1990). In addition to R&D expenditures, patents are commonly considered and used as a proxy to measure innovation (Choi et al., 2021; Hall et al., 2005; Trajtenberg M, 1990). A patent is a legal document that gives an inventor the exclusive right to use the patented invention. Like scientific literature, a patent can be cited by another patent. Patent citations are relevant since they indicate the novelty of an invention over the prior existing knowledge of content related to the invention, better known as prior art. Data about patent citations is often used by scholars for a handful of reasons. For example, patent citations are used to indicate the relevance of a patent publication. If a particular patent has been cited frequently, it indicates that the patent has had a major contribution to the corresponding industrial field. Besides this area of research, for more than three decades, patent citations are used by scholars as a proxy to measure knowledge diffusion (Abramo et al., 2020; Breschi & Lissoni, 2009; Castaldi et al., 2015; Jaffe et al., 1993; Nelson, 2009; Thompson & Fox-Kean, 2005).

As is often the case with diffusion, the closer you get to the source, the stronger the rate of diffusion. Think about radionuclides, a radiating heat source, or the diffusion of gas in a room. Concerning the diffusion of knowledge, and knowledge spillovers, scholars claim that the same holds for knowledge. In other words, the idea is that the diffusion of knowledge is bounded in space (Kijek & Kijek, 2019). More than a century ago, Marshall (1920) asked himself why industries are concentrated in cities. One of his theories suggested that the geographic concentration of industries is favored because of knowledge spillovers. Here, Marshall (1920) introduces the assumption that knowledge spillovers are facilitated by spatial proximity. Approximately a century later, the relationship between knowledge diffusion and spatial proximity is still a trending topic in research. Over the years, multiple studies have shown that the diffusion of knowledge, measured by patent citations, is indeed, bounded in space (Abramo et al., 2020; Almeida & Kogut, 1999; Belenzon & Schankerman, 2013; Breschi & Lissoni, 2009; Jaffe et al., 1993; Thompson & Fox-Kean, 2005; Zucker et al., 1998).

1.2 Problem Statement and Research Questions

As can be derived from the previous section on knowledge diffusion and patent citations, the diffusion of knowledge is sticky. Since ecosystems do have proximal conditions that favor knowledge diffusion (Robertson et al., 2023), this research assumes that the diffusion of knowledge is even more sticky within -and around ecosystems. Although it is captured in the literature that ecosystems do have proximal conditions that favor knowledge diffusion, and thereby make it more likely that knowledge is stickier within these systems, empirical findings supporting this relationship are limited in the current scientific literature. Secondly, in the three decades that the localization of knowledge diffusion, indicated by patent citations, has been studied, the localization of patent citations originating from cited ecosystem patents has not been studied yet. Therefore, with the lack of empirical research on ecosystems and knowledge diffusion, this study aims to find out whether patent citations, originating from cited ecosystem patents, are localized.

It does so by analyzing patent data from two ecosystems over the years 2003 – 2014: Brainport Eindhoven (i) and the High Tech Campus Eindhoven (ii). These two ecosystems are chosen as a case study for a handful of reasons. First of all, Brainport Eindhoven is considered one of Europe's leading high-tech ecosystems. Secondly, since Brainport Eindhoven is accountable for approximately 40% of Dutch patent applications, the availability of patent data is abundant. Thirdly, being a student of Eindhoven University of Technology, I'm, grateful and excited to study the ecosystem I do live in. Down below, the aim of this research has been formulated into an overarching -and main research question. Besides the main research question, a sub-question is formulated which serves the main research question. The sub-question will be answered by focusing on knowledge diffusion in general (i), the diffusion of knowledge within the ecosystems (ii), and the diffusion of knowledge within -and outside the ecosystems (iii). Therefore, three hypotheses have been formulated and will be tested throughout the research.

Main research question*:* Is the diffusion of knowledge within -and around ecosystems spatially stickier compared to non-ecosystems?

Sub-question: Are patent citations, originating from cited High Tech Campus Eindhoven (HTCE) -and Brainport Eindhoven patents, more localized compared to patent citations originating from cited 'non-ecosystem' patents?

Hypothesis 1: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized compared to patent citations originating from cited 'non-ecosystem' patents.

Hypothesis 2: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized within the ecosystem compared to patent citations originating from cited 'non-ecosystem' patents.

Hypothesis 3: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized within -and around the ecosystem compared to patent citations originating from cited 'non-ecosystem' patents.

1.3 Contribution

By anticipating the problem statement described above, the contribution of this report to the Innovation Sciences can be considered twofold. First of all, this master thesis complements the academic literature on both ecosystems and knowledge diffusion by conducting empirical research on how knowledge is diffused within -and around ecosystems. Next to that, this master thesis contributes to the field of research which focuses on the geographical feature of knowledge diffusion.

1.4 Outline of the Report

Next, in the Literature review (Chapter 2), the relevant literature related to ecosystems, knowledge diffusion, and the use of patent citations, as a proxy for knowledge diffusion, will be elaborated. With the use of this literature, the aim is to create a solid research design, which will later be presented in the chapter on the Methodology (Chapter 3). In this chapter, it is explained how, with the use of a quantitative research approach, the research questions will be answered. Next, in the chapter on Results (Chapter 4), the outcomes of the quantitative analysis will be presented. Finally, in the Conclusion (Chapter 5), it will be elaborated on what the results tell and what conclusions can be drawn from it. Finally, in the Discussion (Chapter 6), it will be discussed what the results mean and how they should be interpreted. Next to that, in this chapter, the limitations, contributions, and recommendations will be addressed.

2. Literature Review

The literature that this research draws upon is divided into two sub-chapters. The first chapter is centered around the concept of ecosystems. Firstly, it will be illustrated how the paradigm shift towards open innovation shaped ecosystems. After, the concept of ecosystems will be elaborated and applied to the ecosystems of this case study: Brainport Eindhoven and the HTCE. The chapter will wrap up by elaborating on how various forms of proximity affect knowledge diffusion with respect to innovation, ecosystems, Brainport Eindhoven, and the HTCE.

The second chapter will be centered around the concept of patent citations as a proxy for knowledge diffusion. First, an introduction will be made to patents and patent citations. After that, a reflection will be made on the area of literature that assessed the use of patent citations as a proxy for knowledge diffusion. Finally, it will be elaborated on how scholars used and applied patent citations to measure (localized) knowledge flows.

2.1 Ecosystems

2.1.1 Open Innovation and Ecosystems

2.1.1.1 Open Innovation

In the history of innovation, open innovation is a relatively new concept. For a long time, 'closed innovation' used to be the conventional paradigm for innovation (Chesbrough, 2006). During World War II, the US government recognized how science and R&D created products that were decisive for the outcome of the war, e.g., the development of the atomic bomb. Before this period, entrepreneurs were not that interested in the commercialization of scientific findings. After the war, the US government significantly increased its R&D investments (Chesbrough, 2006). Around this time and in the years to come, the golden age of R&D was characterized by deep vertical integration (Chesbrough, 2006). For any product on the market, there were very few capable alternatives. In addition, due to dominant market positions in product markets, it was easy to capture value from one's R&D when controlling the entire value chain of business activities.

However, eventually, the paradigm of closed innovation started to erode, due to a changing landscape of knowledge. Due to globalization, knowledge became more accessible. Because of that, the significance of R&D, being the most important asset for revenue, started to decrease. Innovation started to open up through the diffusion of knowledge instead of the isolated creation of knowledge. Business was no longer solely focused on inventing and commercializing new knowledge. Instead, as a firm, you would win by making the best use of internal and external knowledge in a timely way, creatively combining that knowledge in new and different forms to create new products and services (Chesbrough, 2006). In that way, knowledge diffusion became more and more relevant to innovation processes. As explained by Wu et al. (2021), concerning

open innovation and knowledge diffusion, firms purposefully interconnect and exchange knowledge with external entities to acquire external knowledge. The aim to acquire external valuable knowledge, by "opening up" internally and exchanging internal knowledge, is the fundamental principle of open innovation.

2.1.1.2 Ecosystems

Since actors and organizations started to exchange knowledge, collaborations, and interdependencies among these actors started to grow. As a result, sometimes, these collaborations and interdependencies led to the creation of non-hierarchical systems (Jacobides et al., 2018). Scholars defined such systems as 'ecosystems', referring to biological ecosystems.

In their research, Cobben et al. (2022) reviewed the conceptual boundaries of four commonly studied ecosystems: business, innovation, entrepreneurial, and knowledge-based ecosystems. The business ecosystem, as introduced by Moore (1993), can be described as an ecosystem of companies, active in multiple industries, that focuses on one focal firm (Moore, 1993, as cited in Cobben et al., 2022). As an example, Moore introduces the Apple ecosystem. Apple is the leader of a business ecosystem that covers an extended web of suppliers from different industries. The innovation ecosystem can be described as a system of collaborative arrangements between firms in which firms' individual offerings and contributions are combined to anticipate a customer-facing solution (Adner, 2006, as cited in Cobben et al., 2022). Granstrand & Holgersson (2020) define the innovation ecosystem as the evolving set of actors, artifacts, and activities, bounded by relations and institutions, which are important for the innovative performance of a population of actors.

The knowledge-based ecosystem can be defined as a group of knowledge-intensive firms, located in close proximity, that depend on one another for the effectiveness and efficiency of their operations (Borgh et al., 2012, as cited in Cobben et al., 2022). The entrepreneurial ecosystem focuses on people, and entrepreneurs, whose entrepreneurial activities constitute the ecosystem (Stam, 2015). In such a system, the value creation of these entrepreneurs is organized by a variety of governance modes, all confined within a particular institutional context (Stam, 2015 as cited in Cobben et al., 2022). Stam (2015) illustrates that the degree of entrepreneurial activity within an ecosystem is the result of the presence of entrepreneurial elements like networks, leadership, talent, finance, knowledge, support, culture, and demand. Yang et al. (2022) stress the importance of entrepreneurial ecosystems by showing that entrepreneurial ecosystems promote municipal economic growth, according to a study on 32 cities in China between 2008 – 2018.

2.1.1.3 Brainport Eindhoven and the High-Tech Campus Eindhoven

In its essence, Brainport Eindhoven can be categorized an entrepreneurial ecosystem, since the mobilization of talent, start-ups, and networks are the key characteristics of the ecosystem. A good example of the entrepreneurial ecosystem at Brainport Eindhoven is the HighTech XL. HighTech XL is an open innovation-oriented platform created by the HTCE and other partners within Brainport Eindhoven (High Tech Campus Eindhoven, 2023). The goal of HighTech XL is to connect the technological skills of the Brainport region with entrepreneurship. These teams

of entrepreneurs and technologists are matured by the program of HighTech XL and connected to investors and mentors. To make this happen, the HighTech XL created the so-called 'Eindhoven Start-up Alliance'. In this alliance, facilitated by Philips, ASML, HTCE, and BOM, collaborations are arranged between start-ups, multinational corporations, SMEs, and research institutes. Although, in its totality, Brainport Eindhoven can be considered an entrepreneurial ecosystem, within this ecosystem, the other conceptual ecosystems can be found as well. For example, the HTCE, which is located in Brainport Eindhoven, can be considered a knowledge-based ecosystem (Borgh et al., 2012). Within this knowledge-based ecosystem and Brainport Eindhoven, at a smaller scale, ecosystems can be found as well. An example of this is the AI Innovation Center, which is located at the HTCE. The main focus of the AI Innovation Center is to foster and accelerate the adoption of AI in the Brainport region by creating a network of entrepreneurs. This is a clear example of the functioning of an innovation ecosystem: firms' individual offerings and contributions are combined to anticipate a customer-facing solution. The AI Innovation Center is an initiative of the HTCE's Campus Site Management, HTCE, and was co-founded by Philips, ASML, NXP, and Signify (Koelman, 2021). So, as can be derived from this section, Brainport Eindhoven can be considered an entrepreneurial ecosystem in its totality, whereas within the ecosystem, the other conceptual ecosystems are practiced as well.

All in all, since the role of knowledge diffusion became more relevant with the emergence of open innovation and ecosystems, the role of proximity, as a facilitator of knowledge diffusion, became more relevant as well. In the introduction, it was mentioned that the diffusion of knowledge is affected and dependent on multiple forms of proximity (Boschma, 2005). Especially in the context of ecosystems, proximity is a fundamental pillar of an ecosystem's functioning (Robertson et al., 2023). In the descriptions of the conceptual ecosystems by Cobben et al. (2022), the importance of cultural, geographical -and intuitional proximity was addressed. Therefore, in the next chapter, the concept of proximity and its relation to knowledge diffusion will be elaborated. Next to that, the theory on proximity will be assessed in the local case of this research: Brainport Eindhoven and the HTCE.

2.1.2 Proximity

In this sub-chapter, the concept of proximity, in relation to knowledge diffusion and ecosystems, will be elaborated. This chapter focuses on the mechanisms that establish knowledge flows -and diffusion. For knowledge to be transferred among actors, actors need to be able to capture and integrate knowledge. For example, by using language, humans can understand one another. Using language, humans can exchange knowledge. When someone is not able to read, that person is not able to read this thesis and capture the information embodied in this thesis. This is an example of cognitive proximity. In this section, the following proximities, with respect to knowledge diffusion and ecosystems, will be discussed: cognitive proximity, organization proximity, social proximity, institutional proximity, and geographical proximity (Boschma, 2005).

2.1.2.1 Cognitive Proximity

In the above paragraph, an example of cognitive proximity has been given using the example of language. Cognitive proximity is all about the capacity of an actor to capture and integrate the knowledge made available by other actors. In the view of cognitive proximity, there is a potential that people who share a similar knowledge base can learn from one another. However, considering knowledge diffusion and cognitive proximity, Boschma (2005) argues that there is a trade-off. On the one hand, a high cognitive distance opens space for learning, although the absorptive capacity might be lower. On the other hand, when there is no cognitive distance, there is a threat of lock-in and involuntary spillovers. With respect to cognitive proximity at Brainport Eindhoven, it can be argued that the cognitive distance is rather small since Brainport Eindhoven is characterized by high-tech industries. The same reasoning can be applied to the 'High-Tech' Campus. Nevertheless, the high-tech industry covers a lot of different industries and multiple disciplines, implying that the cognitive distance can also be high as well in a high-tech environment.

2.1.2.2 Social Proximity

In addition to cognitive proximity, which focuses on a similar knowledge base between actors, social proximity rather focuses on the embedded relations between actors at the micro-level. Examples are relationships based on trust, experience, and friendship. In general, the more economic relationships among actors are socially embedded, the better the economic performance (Boschma, 2005). However, this linear relationship holds up to a certain threshold, after which the economic performance decreases due to lock-in effects, as displayed in Figure 1. Too many socially embedded relationships will create a lock-in effect because an agent will remain stuck in a particular network of agents, failing to access new knowledge through agents outside the network.

The dynamics displayed in Figure 1 represent the theory and the dynamics of "strong" -and "weak" ties, as often used in network theory. Here, on the one hand, strong ties refer to a strong and socially embedded network. On the other hand, weak ties refer to a social network in which the social proximity is lower. However, in such a network, agents are in a better position to acquire new knowledge. Agents who lack weak ties are more likely to be confined to a few cliques in which innovation is less likely to diffuse (Granovetter, 1973).

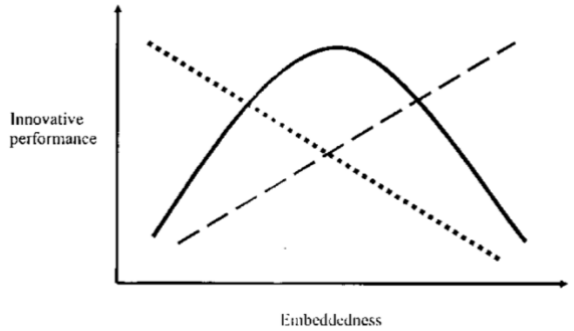


Figure 1. Innovative performance versus socially embedded relationships (Boschma, 2005).

With respect to social proximity at Brainport Eindhoven and the HTCE, it can be argued that the social embeddedness of the actors is high. For example, entrepreneurial activity facilitated by platforms like the HighTech XL and The Gate connects actors from different levels with one another. Next to that, multinational companies like ASML, Signify, and NXP are all spin-offs of Philips. It was found that, by investigating 30 university spin-offs, spin-off companies often keep close contact with the parent firm in the early stage of the spin-off (Rogers et al., 2001). Although these spin-offs all have become multinationals right now, it could still be that the social networks of these firms are socially embedded with one another.

2.1.2.3 Organizational Proximity and Institutional Proximity

Organizational proximity is related to the coordination of transactions within an organization. It refers to interdependencies that are present within -and between organizations. If the organizational proximity is too low, firms might turn to opportunistic behavior, meaning that a firm will take advantage of weak organizational structures (Boschma, 2005). On the other hand, if the proximity is too high, there is a threat of lock-in effects in which dependencies will limit the entrance of novelty within the organizational structures (Boschma, 2005; Talbot, 2021).

Whereas organizational proximity focuses on the meso-level, institutional proximity is related to the macro-level. It includes the idea that actors share the same rules of the game and the same cultural norms and values, e.g., a common language and a law system that secures ownership and navigates intellectual property rights. Again, too much institutionalization creates lock-in effects, which limits novelty, whereas the absence of organizational proximity diminishes the effective allocation of resources.

The fact that Brainport Eindhoven and the HTCE can be considered an ecosystem is due to the presence of organizational structures. As mentioned earlier, ecosystems are not fully hierarchically controlled, however, at the same time, there is some kind of organizational structure to keep the systems intact. Organization-wise, both Brainport Eindhoven and the HTCE have supervising organizations that represent the ecosystem. Institutionally, all companies and actors within Brainport Eindhoven belong to the same municipality in Eindhoven.

2.1.2.4 Geographical Proximity

Finally, this paragraph will discuss the role of geographical proximity with respect to knowledge diffusion. The geographical features of the Netherlands, Belgium, Brainport Eindhoven, and the HTCE are illustrated in Figures 2 – 4. Brainport Eindhoven covers approximately one-third of the province of “Noord-Brabant”. On the other hand, the HTCE covers just an area of approximately one square kilometer. So, spatially, Brainport Eindhoven is way more dispersed whereas the HTCE is very dense. Although Brainport Eindhoven is bigger than the HTCE, the region is still not that big, allowing for the mobilization of entrepreneurs across the region. As mentioned in the introduction, start-ups address the importance of the ecosystem by mentioning platforms like HTCE, TU Eindhoven, The Gate, and Innovation Space, acting as facilitators for start-ups to mobilize their operations (Van Leest

et al., 2022). Therefore, geographical closeness is important for actors to easily access and mobilize throughout these platforms.

Figure 2. Map of the Netherlands and Belgium with the province of Noord-Brabant encircled by the red dotted line.

Figure 3. Map of the province Noord-Brabant including the Brainport region colored in red.

Figure 4. Location of the HTCE in the southeast of Eindhoven.

Although geographical proximity is, intuitively, very important for knowledge to diffuse, as mentioned in the introduction, simply "being there" is not enough for knowledge to diffuse. Geographical proximity facilitates cognitive -and social proximity, which allows knowledge to diffuse. If the spatial distance is short, it is easier for people to physically connect and create social relationships. So besides facilitating other proximities, what is the role of geographical proximity? According to Boschma (2005), spatial proximity is not a prerequisite for interactive learning to take place. To exchange tacit knowledge, face-to-face contact is still required. However, spatial proximity in the sense of permanent co-location is not required, since the co-presence can still be organized by bringing people together through traveling. Although permanent co-location is not required, geographical proximity is a prerequisite for the exchange of tacit knowledge through face-to-face contact.

So, Boschma (2005) argues that spatial proximity is not a sufficient nor necessary condition for the exchange of tacit knowledge. Another argument justifying his claim is the argument that networks are social networks that exclude outsiders. Here, he gives the example of multinational corporations that set up local plants in a host region to get access to the knowledge base of that host region. However, it is hard to become a member of such tight and established networks with the presence of gatekeepers. Likewise, according to Talbot (2021), geographical proximity has a potentially positive effect on innovation, but the form of proximity is not indispensable. Here, Talbot argues that firms often maintain strong ties with partners located at a long distance, to prevent to end up being too constrained by the local bubble.

2.1.2.5 Summary Proximity

In sum, geographical proximity is neither a necessary nor a sufficient condition for interactive social learning (Boschma, 2005). It is not necessary, because other forms of proximity may act as substitutes, and it is not sufficient because learning processes require at least cognitive proximity in addition to geographical proximity. Besides geographical proximity, as can be derived from this chapter, the other forms of proximity are strongly represented within the ecosystems of the HTCE and Brainport Eindhoven. These findings are in line with the results found in the paper by Robertson et al. (2023), in which it is argued that ecosystems do have proximity-related advantages. According to Robertson et al., these proximal advantages contribute to the increased diffusion and creation of knowledge in which actors create, diffuse, and acquire knowledge within the ecosystem. Therefore, in this chapter, the idea is strengthened that the diffusion of knowledge is stickier within -and around ecosystems. For the remainder of

this report, the five proximities will not be studied at the HTCE and Brainport Eindhoven. The sole purpose of this chapter was to illustrate how knowledge diffusion is facilitated by proximity and to reason why knowledge diffusion is likely to be stickier within -and around ecosystems.

2.2 Patent Citations as a Proxy for Knowledge Diffusion

In the previous chapter, it was discussed how the shift towards open innovation led to the emergence of a variety of ecosystems, and specifically, the emergence of Brainport Eindhoven and the HTCE. Besides that, it was elaborated how various forms of proximity affect knowledge diffusion with respect to innovation, ecosystems, Brainport Eindhoven, and the HTCE. As the previous chapter explained how knowledge diffusion and ecosystems are intertwined, this chapter will elaborate on the literature that studied the use of patent citations as a proxy to measure knowledge diffusion.

2.2.1 Patents and citations

2.2.1.1 Patents

First of all, what is a patent? A patent privatizes and makes commercial knowledge, in the form of technical inventions, excludable by conferring exclusive rights to the owners of a patent, which are often the inventors and/or applicants (USPTO, 2022; Wang & Zheng, 2023). In general, a patent contains tacit knowledge (Kijek & Kijek, 2019). In contrast to codified knowledge, tacit knowledge is characterized by know-how, practical experience, and action-oriented knowledge (Kijek & Kijek, 2019; Park et al., 2022).

The owner of a patent has the right to exclude others from their technical innovation. In return, the owner of a patent must fully disclose the content of the invention. After 20 years, the legal protection of the patent will expire. Due to their nature, patents do stimulate both monopolies and knowledge diffusion. Not every invention is patentable. There are strict conditions determining whether an invention is applicable to be patented. The conditions are the following: patentable subject matter (i), industrial applicability (ii), novelty (iii), inventive step (iv), and the disclosure of the invention (v). Especially the latter condition, the disclosure of the invention, has a major impact on knowledge diffusion. Through disclosure, the content of an invention becomes publicly available and accessible. Since a high-end goal of patents is to stimulate knowledge diffusion, the invention should be disclosed in such a way that it is understandable to the public. Otherwise, the invention will remain vague and untouchable to the public. Therefore, the invention should be clearly explained in a way that a person who is skilled in the art can replicate and carry out the invention. Also, the inventor must clearly define the scope of the legal rights and the boundaries of exclusion.

2.2.1.2 Patent Citations

Like in the academic world, where literature is cited to prevent plagiarism and to address novelty, references are made among patents. Here, two different types of citations can be distinguished: *backward -and forward citations.* Imagine a citation link between patent A, filed

in 2002, and patent B, filed in 2010. In this case, a reference is made by patent B to patent A. With respect to patent B, we speak of a 'backward citation', whereas, for patent A, we speak of a 'forward citation'. So, both backward -and forward citations can represent the same citation link, whereas the categorization of being either a backward or forward citation depends on the perspective of the cited and citing patent. All the backward citations contribute to the 'prior art' of a patent, which represents all the published knowledge that is closely related to the invention. In general, those patents with a high number of forward citations are considered "technologically important", since the patent contains knowledge that forms the basis for multiple subsequent inventions (Fontana et al., 2009). Next to that, more than three decades ago, Trajtenberg (1990) studied and found that patent citations are indicative of the value of innovations whereas later Hall et al. (2005) found that patent citations contain significant information on the market value of firms.

All in all, since knowledge flows between inventions do leave a paper trail in the form of patent citations, these knowledge flows become tangible and measurable. In that way, patent citations can act as a proxy to measure knowledge diffusion. In the next chapter, the vast body of literature, that assessed the use of patent citations as a proxy for knowledge diffusion, will be discussed.

2.2.2 Patent Citations as a Proxy to Measure Knowledge Flows

2.2.2.1 Patent Citations as a Proxy for Knowledge Diffusion

Since the pioneering work by Jaffe et al. (1993), patent citations have been utilized extensively as a proxy to measure knowledge diffusion. With the utilization of this proxy, over the years, scholars assessed and researched the quality of this proxy. Perhaps the most important argument made by scholars, in addressing the limitations of patent citations as a measure of knowledge flows, is the influence of examiner citations (Alcácer et al., 2009; Alcácer & Gittelman, 2006; Corsino et al., 2019; Criscuolo & Verspagen, 2008). In the patent's process of examination at a patent office, to judge the degree of novelty, the examiner of the patent is likely to add additional prior art (Criscuolo & Verspagen, 2008). Consequently, since examiners of a patent are likely to add patent citations as well, it can be questioned whether patent citations reflect true knowledge flows (Criscuolo & Verspagen, 2008). Alcácer & Gittelman (2006) were the first to find that, on average, approximately two-thirds of all the patent citations are added by examiners and that for 40% of all the patents, the citations are all added by examiners. Later, by including a specific focus on jurisdiction, Alcácer et al. (2009) found that most of the citations are added by examiners for USPTO patents and Criscuolo & Verspagen (2008) found a similar result for EPO patents. Additionally, using U.S. patent data, Alcácer & Gittelman found that examiner citations are biased as examiners include more self-cations and more proximate citations. On the contrary, using EPO patents, Criscuolo & Verspagen found that the frequency of self-citations is twice as high for the inventors' citations compared to citations from examiners. Also, Criscuolo & Verspagen found that inventor citations are more localized. Besides, Criscuolo & Verspagen found that examiner citations often involve more citations that might compromise novelty and inventiveness.

Besides the influence and biases of examiner citations, Corsino et al. (2019) found that measurement errors of patent citations, originating from firms, are rooted in firms' incentives to cite prior art. These incentives depend on some key factors (Corsino et al., 2019). The first important factor is the patenting system and its regulations. For example, the USPTO is characterized by a doctrine that triggers the inclusion of many backward citations, making USPTO patents a noisier measure of knowledge flows compared to EPO patents (Corsino et al., 2019). A second factor is the technological field in which a patent can be categorized. For instance, applicants in discrete technological fields, such as biotechnology, disclose more prior art compared to applicants in complex industries, such as telecommunications (Corsino et al., 2019). This is because invalidation, due to the incomplete disclosure of prior art, is way more damaging in the field of biotechnology than in the field of telecommunications. Thirdly, larger firms are less likely to withhold prior art since invalidation would hamper production processes or would disallow the commerce of a patent, resulting in high costs. Fourthly, the patenting strategy of a firm has a huge influence on the citing behavior of a firm. For example, patents that are filed to preempt other competing firms are less likely to disclose all relevant prior art. On the other hand, the abundance of prior art rises when a firm is patenting for commercial intentions or to prevent lawsuits. In addition to these findings by Corsino et al., years before that, Jaffe et al. (1998) found that often very basic patents are cited; patents that everyone considers as basic, which are often old and well-known.

To speak of a legitimate knowledge flow, the cited patent must be fully understood and incorporated into the new invention. With these questions in mind, Jaffe et al. (2000) studied the relationship between knowledge spillovers and patent citations. In their study, considering a particular patent, Jaffe et al. surveyed two groups of inventors: the citing inventor and the cited inventor. Here, Jaffe et al. asked a series of questions about the nature of the patent and the communication among the inventors. The cited inventor was asked to read the citing patent and to form a judgment about the likelihood that the citing inventor had used the knowledge embodied in the cited inventor's patent. Though the results included some noise, as perceived by the inventors themselves, the results do show some evidence that citations are correlated with importance. Though not convincing, about 60% of the inventors argued that they benefited from a cited patent, in terms of relevance for the invention by the citing inventor (Jaffe et al., 2000).

In conclusion, as was noted almost three decades ago, patent citations are a valid, but, at the same time, a "noisy" indicator of knowledge diffusion (Jaffe et al., 1998).

2.2.2.2 Licenses and Literature Publications

Besides reflecting on the intrinsic quality of patent citations as a measure of knowledge diffusion, the quality can be assessed by comparing patents to other possible indicators of knowledge diffusion. Licensing agreements and literature publications are two other major indicators of knowledge diffusion. If an actor is willing to use a patented technology for the commercialization of a product, to prevent infringement, the actor is obliged to license the patented technology. In case the actor wants to innovate on the patented technology, the actor is obliged to make a reference. Below, using the paper by Nelson (2009), these other two

indicators of knowledge diffusion will be examined and compared to the proxy of patent citations. The purpose of this section is to briefly introduce that, besides patent citations, other proxies of knowledge diffusion can be measured and used. For the remainder of this research, however, these other proxies will not be studied.

In his research, Nelson (2009) studied the differences, similarities, and dynamics between patents, licenses, and scientific citations as measures of knowledge diffusion. At the beginning of the paper, Nelson reasons the limitations of patent citations as an indicator of knowledge diffusion. Besides the critiques, similar to the issues addressed in the previous paragraph, Nelson argues that patent citations fail to capture and indicate a big majority of the knowledge that is being diffused among organizations and companies. In his research, Nelson performed a case study on the patented technology of rDNA, consisting of multiple patents. He found that 135 organizations hold direct patent citations to the focal rDNA patents, whereas 464 organizations hold license agreements to the focal rDNA patents. Since 55 organizations hold both direct patent citations and license agreements, patents fail to capture 409 organizations that are present in the licensing sample, which corresponds to 88.1% of the organizations holding patent citations (Nelson, 2009).

The same holds for literary publications. The most common knowledge flows within the domain of literature are scientific citations. Patent citations do miss out on a lot of 'knowledge-sharing organizations' when solely focusing on patent citations (Nelson, 2009). Like patent citations, scientific citations are not a perfect indicator of knowledge diffusion. Scientific citations are biased and authors do only cite 30% of their formal influences (Nelson, 2009). Besides these issues, literature publications are a significant indicator of knowledge flows. Figure 5 illustrates the number of organizations that are captured by each measure of knowledge diffusion. As can be seen, compared to licensing, patents do miss out on a lot of organizations, firms, and firms with products.

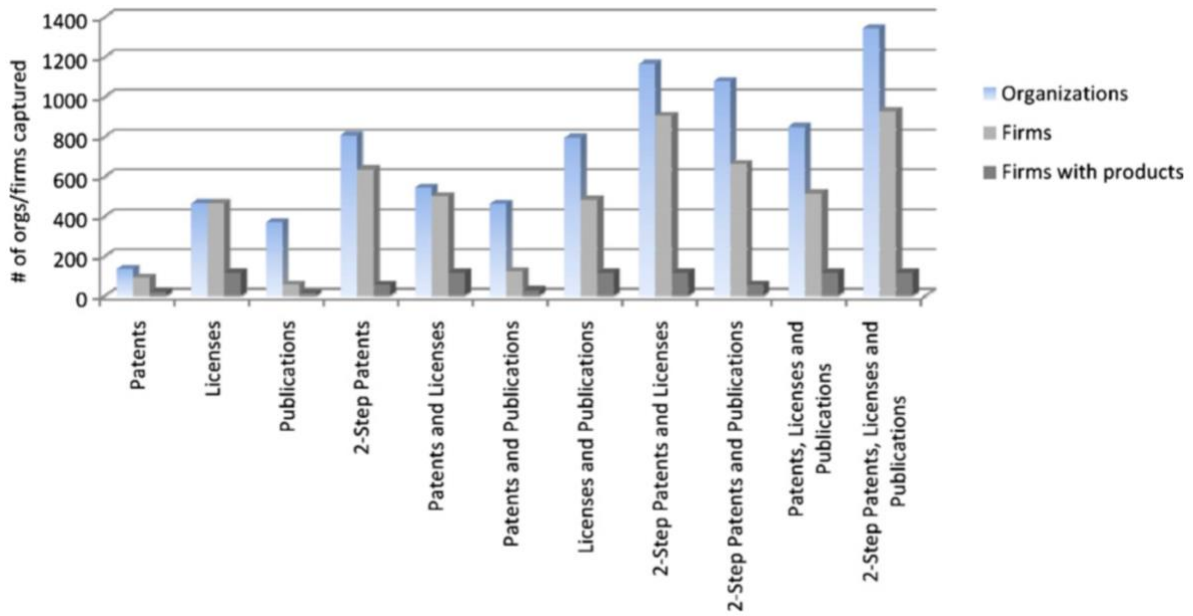


Figure 5. The number of unique firms captured by different measures of knowledge diffusion (Nelson, 2009).

2.2.3 Patent Citations to Measure the Localization of Knowledge Flows.

As the previous chapter assessed the use of patent citations as a proxy for knowledge diffusion, this chapter focuses on 'how' patent citations are used in research. It will do so by reflecting on a specific selection of literature that used patent citations as a measure to study the localization of knowledge flows. In this area of research, the research by Jaffe et al. (1993) is considered to be a pioneer and a seminal for the research field (Abramo et al., 2020; Belenzon & Schankerman, 2013; Breschi & Lissoni, 2009; Criscuolo & Verspagen, 2008). The research by Jaffe et al. has become a seminal by introducing the rationale that knowledge spillovers do leave a paper trail in the form of patent citations. Next to that, Jaffe et al. were the first to study whether knowledge flows, measured by patent citations, are spatially localized.

2.2.3.1 Jaffe et al. (1993)

In their study, Jaffe et al. (1993) collected two datasets of patents. One cohort consisting of 950 patents filed in 1975 received 4750 citations by the end of 1989. The second cohort, consisting of 1450 patents, all filed in 1980, received a total of 5200 citations by the end of 1989. To study whether citation links are spatially localized, Jaffe et al. made use of control patents to create a baseline of reference. To be more specific, for each citing patent, Jaffe et al. found a "twin" patent that is almost identical to the focal citing patent. This control patent was found by looking for all patents within the same patent class and the same application year as the citing patent. From this set of patents, Jaffe et al. chose the patent whose grant date came closest to the citing patent. As a result, Jaffe et al. created a dataset of patents including the cited patents, citing patents and control patents.

Regarding the validity of the control patents, the grant date of the control patent must be as close as possible to the citing patent. If the difference between the grant date of the control -and

citing patent is too large, it can be argued that the technological landscape is very different for both patents. For example, the control patent was granted two years later than the citing patent. In two years of time, the technological landscape has changed by additional patents being granted, which might influence the probability that the cited patent is cited by a control patent. If the grant date of the citing -and control patent are not in line, the comparison between the two becomes less trustworthy, it's comparing apples and oranges.

Next to the issue of time, Jaffe et al. (1993) address the issue of self-citations. Self-citations occur when an inventor uses one of his patents for a new invention. So, the inventor, or the examiner, uses one of the inventor's own inventions as prior art. These citations cannot be considered knowledge spillovers. However, it is important to mention that Jaffe et al. adopt a different interpretation of self-citations. In their research, Jaffe et al. exclude those citations that have the same origin in terms of the organization. So, if a citation comes from the same firm or organization, Jaffe et al. considered the citation a self-cited citation.

2.2.3.1.1 Results and Findings

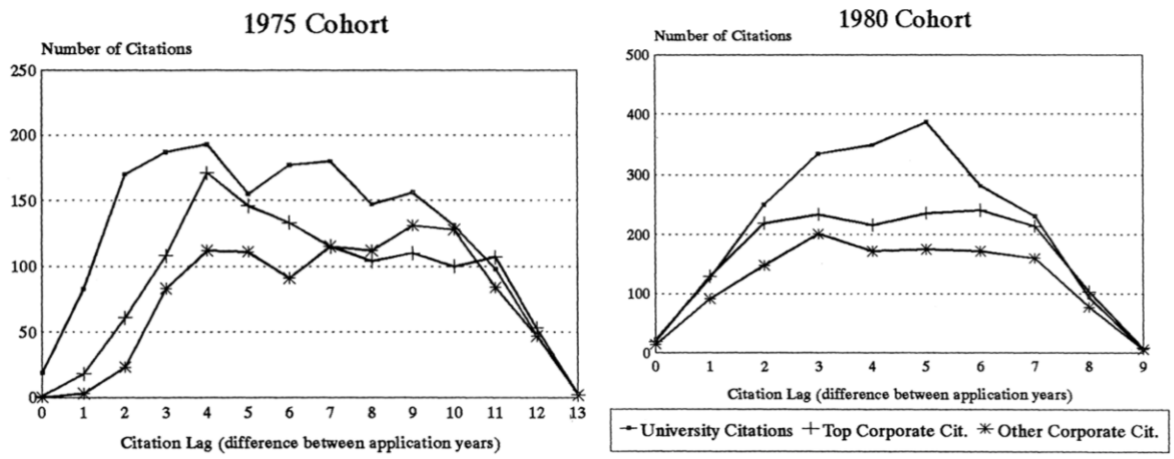


Figure 6. Citation lag in both cohorts (Jaffe et al., 1993).

Figure 6 displays the citation lag of the cohorts. For each citation link, the citation lag represents the difference in years concerning the timing of filing between the cited -and citing patent. Figure 6 provides an overview of the number of patents for different values of the citation lag. As can be seen, Figure 6 shows that the number of citations is at its peak around approximately 5 years and eventually diminished over time.

GEOGRAPHIC MATCHING FRACTIONS

	1975 Originating cohort			1980 Originating cohort		
	University	Top corporate	Other corporate	University	Top corporate	Other corporate
Number of citations	1759	1235	1050	2046	1614	1210
		Matching by country				
Overall citation matching percentage	68.3	68.7	71.7	71.4	74.6	73.0
Citations excluding self-cites	66.5	62.9	69.5	69.3	68.9	70.4
Controls	62.8	63.1	66.3	58.5	60.0	59.6
t-statistic	2.28	−0.1	1.61	7.24	5.31	5.59
		Matching by state				
Overall citation matching percentage	10.4	18.9	15.4	16.3	27.3	18.4
Citations excluding self-cites	6.0	6.8	10.7	10.5	13.6	11.3
Controls	2.9	6.8	6.4	4.1	7.0	5.2
t-statistic	4.55	0.09	3.50	7.90	6.28	5.51
		Matching by SMSA				
Overall citation matching percentage	8.6	16.9	13.3	12.6	21.9	14.3
Citations excluding self-cites	4.3	4.5	8.7	6.9	8.8	7.0
Controls	1.0	1.3	1.2	1.1	3.6	2.3
t-statistic	6.43	4.80	8.24	9.57	6.28	5.52

Figure 7. Results showing the proportion of patents that come from the same geographical area (Jaffe et al., 1993)

Figure 7 summarizes the results of the localization of citation links. For each geographical area, it summarizes the proportion of the citing patents that come from the same geographical area as the cited patents (Jaffe et al., 1993). Figure 7 displays three categories: overall citation matching percentage, citations excluding self-citations, and the controls. Looking at the 1980 cohort, it can be seen that for every geographical area, the percentages of the citations excluding self-citations are significantly higher than the percentages of the controls (Jaffe et al., 1993). Compared to the control patents, citations excluding self-citations are two to six times more likely to come for the same SMSA. Citations are twice as likely to come from the same state. The 1975 cohort displays the same pattern, though the pattern is weaker. Based on these findings and results, Jaffe et al. concluded that patent citations are indeed localized.

2.2.3.2. Critique Towards Jaffe et al. (1993)

Even though the study by Jaffe et al. (1993) is considered a seminal -and pioneer in the research field of knowledge diffusion and patents, over the years, scholars expressed their critique towards the research. In their research, Breschi & Lissoni (2009) provide a detailed overview of the criticism and doubts that have been raised by scholars in response to the paper by Jaffe

et al. Two types of critiques and doubts have been raised. Firstly, it is questioned whether localized knowledge flows can be considered true knowledge spillovers, referring to the distinction between knowledge diffusion through social informal ties and those traveling through transactions in the market economy (causality). The second line of critique is related to the methodology that has been used by Jaffe et al. Next to that, Breschi & Lissoni stress the importance of understanding the mechanisms through which knowledge is transmitted from the origin (cited patent) to a new destination (citing patent).

In the previous section, it has been mentioned that Jaffe et al. (1993) acknowledged that contractual agreements can diminish the likelihood of the presence of true knowledge externalities. As a response to this example by Jaffe et al, Breschi & Lissoni (2009) introduce the concept of 'partial spillovers', which might be present in contractual agreements. Therefore, Breschi & Lissoni stress the importance to understand and study relationships between knowledge-exchanging agents in order to be able to justify whether a knowledge flow is a true externality or a partial spillover. Both Almeida & Kogut (1999) and Zucker et al. (1998) touch upon these mechanisms by illustrating localization effects due to the interfirm mobility of patent holders. So, knowledge is transferred by individuals who move from one organization to the other (Breschi & Lissoni, 2009). Moreover, Almeida & Kogut did find that localization effects, as studied by Jaffe et al., are more likely to be found in regions with a strong labor market. Next to that, Mowery & Ziedonis (2015) found that citations, having a contractual license agreement, are more localized compared to non-market spillovers. Although the sample solely used university patents of a limited sample, Breschi & Lissoni recognize that the findings by Mowery & Ziedonis are relevant insights.

Besides the concerns about whether the citations used in the study by Jaffe et al. (1993) can be considered true externalities, Breschi & Lissoni (2009) question whether there are any flaws in the methodology used by Jaffe et al. (1993). Thompson & Fox-Kean (2005) show that the control patents in the study by Jaffe et al., which were selected on having a similar USPTO 3-digit code, have little resemblance with the citing patents. According to Thompson & Fox-Kean, because of the 3-digit code, the within-class heterogeneity is rather large. As a result, the control patents do have little resemblance with the citing patents. Thompson & Fox-Kean show that the localization effects diminish when the control -and citing patents are matched at a finer level than the USPTO 3-digit code. On the other hand, in response to Thompson & Fox-Kean, Henderson et al. (2005) argue that a too-tight technological match limits the size of the patent samples. In addition, Henderson et al. argue that technological distance is required for knowledge spillovers to occur since pure imitation and replication cannot produce any follow-up patent, since novelty is an essential requirement for patents to be approved.

2.2.3.3 The Mobility of Inventors and Technological Distance

As mentioned before, in response to the paper by Jaffe et al. (1993), Breschi & Lissoni (2009) address the need to examine the relationships that are embedded in citation links to understand the true nature of the knowledge flow. Therefore, in their research, Breschi & Lissoni examine these relationships by looking at the level of inventors. To be more specific, Breschi & Lissoni

try to trace the links between inventions by looking for so-called “mobile inventors”. The definition of these “mobile inventors” can be best explained using Figure 8.

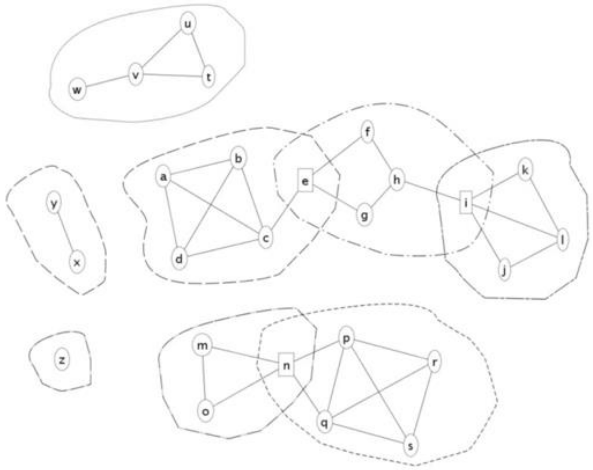


Figure 8. Network of inventors (Breschi & Lissoni, 2009).

Figure 8 displays a schematic overview of a network of inventors. In this network, each node represents an individual inventor. The edges in the network represent the connections among inventors in the network. Each tie corresponds to inventors that have been on the same team on one or more patents. In other words, they have at least co-invented one patented innovation (Breschi & Lissoni, 2009). The dashed lines surrounding the nodes represent organizations and firms. In the network displayed in Figure 8, Breschi & Lissoni describe the presence of three possible relationships among inventors. Firstly, Breschi & Lissoni introduce the relationship of the “mobile inventor”. A mobile inventor is assigned to at least two different patent applicants. In other words, the inventor has worked for at least two different organizations, and at both organizations, the inventor co-invented a patented invention. These inventors might be employees or consultants that move from one company to the other. As can be seen in Figure 8, inventor “e” is a good example of a mobile inventor, since the inventor has co-invented an invention at two different organizations. These mobile inventors act as bridges in a social network, connecting employees of different organizations indirectly with one another. These connections constitute the second type of relationship as described in the research by Breschi & Lissoni. For each inventor, Breschi & Lissoni calculate the path length to reach another inventor through the ties of co-inventions. For example, inventor “l” and inventor “a” are connected through inventor “e” and the path length is $d(a,l)$ equal to 6. The third type of relationship describes the actors that are not connected. For example, in Figure 8, inventors “a” and “w” are not connected to one another in a formal matter.

In terms of cited -and citing patents, Breschi & Lissoni (2009) refer to mobile inventors as an inventor that is both assigned to a cited patent and a citing patent. Furthermore, the cited -and citing patents have two different applicants, two different organizations. Considering the second relationship, for each citing patent, at least one of the inventors assigned to the patent has to be connected to one of the other inventors which are assigned to the patent (Breschi & Lissoni,

2009). The connection must already be present before the new patent is filed. For example, considering the network in Figure 8, a new invention is patented by inventors “a”, “b” and “w”. Inventors “a” and “b” already co-patented an invention before, whereas inventor “w” is new to the network.

With the use of this knowledge, Breschi & Lissoni (2009) performed a similar study to the research by Jaffe et al. (1993). In addition, Breschi & Lissoni wondered what will happen to the results when both mobile inventors and connected inventors are removed from the patent sample. Will the results still show significant localization effects like the findings by Jaffe et al.? The results will be discussed in the next section.

2.2.3.3.1 Results

Since the methodology used by Breschi & Lissoni (2009) is similar to the methodology used by Jaffe et al. (1993), there is no need to further explain the methodology used by Breschi & Lissoni in great detail. Figure 9 displays the results of the study by Breschi & Lissoni. For different combinations of patent pairs, the percentage of localization is given for both the citing -and the control patents at the MSA level and the state level. In the first row, the percentages of localization are given for all patent pairs, according to the JTH experiment. Breschi & Lissoni refer to Jaffe et al. as “JTH”. In the second row, all patent pairs including mobile inventors are removed. As can be seen here, the localization effects for the citing patents strongly diminished, while the percentage of the control patents remained similar, even though only 182 patents were removed from the sample. This effect even gets stronger when other connected pairs are removed from the sample as well. Based on these findings, Breschi & Lissoni conclude that the localization of patent citations is explained by the mobility of inventors. In that way, geography matters because the mobility of inventors and technologists are bounded in space.

	Number of observations	Citing cited	Control cited	z-test ($P > z$)	Odds ratio
MSA level					
All patent pairs (JTH experiment)	3700	17.4	10.8	8.1 (0.00)	1.73
All pairs except those linked by mobility (social distance = 0)	3512	13.5	10.8	3.5 (0.00)	1.29
All pairs except those connected at social distance ≤ 5	3217	11.8	9.9	2.5 (0.01)	1.22
Only not connected pairs	2299	11.7	9.4	2.5 (0.01)	1.21
State level					
All patent pairs (JTH experiment)	3700	20.8	14	7.8 (0.00)	1.62
All pairs except those linked by mobility (social distance = 0)	3512	17.2	13.8	3.9 (0.00)	1.3
All pairs except those connected at social distance ≤ 5	3217	15.5	13	3.0 (0.00)	1.24
Only not connected pairs	2299	15.1	12.5	2.5 (0.01)	1.24

Figure 9. The percentage of citation links within the same geographical area (Breschi & Lissoni, 2009).

2.2.3.4 Is There a Flaw in the Methodology by Jaffe et al. (1993)?

In the previous sections, the critique on the paper by Jaffe et al. (1993) was addressed. In their paper, Breschi & Lissoni (2009) illustrate the critique by other scholars who reflected on the

methodology used by Jaffe et al. For example, Thompson & Fox-Kean (2005) showed that, when increasing the technological match between the control -and the citing patents, the localization effects as introduced by Jaffe et al. disappear. As a response to Thompson & Fox-Kean, Breschi & Lissoni revisited the research and performed the analysis again but this time they used different sets of control patents with different levels of technological closeness. The results of this analysis are displayed in Figure 10. As can be seen here, the more the control patents and the citing patents are related, the more the localization effect is diminished. Eventually, there is even no longer a localization effect. Does this mean that knowledge spillovers and patent citations are no longer spatially localized?

	Number of obs.	Citing cited	Control cited	z-test ($P > z$)	Odds ratio
MSA level					
Matching at primary 4-digit (JTH experiment)	3700	17.4	10.8	8.1 (0.0)	1.73
Matching at primary 12-digit	3168	17.4	13.7	4.1 (0.0)	1.33
Matching at primary 12-digit and secondary 4-digit	2458	17.5	15	2.4 (0.0)	1.2
Matching at primary 12-digit and secondary 12-digit	1882	16.7	16.6	0.1 (0.9)	1
State level					
Matching at primary 4-digit (JTH experiment)	3700	20.8	14	7.8 (0.0)	1.62
Matching at primary 12-digit	3168	20.8	18.4	2.4 (0.0)	1.16
Matching at primary 12-digit and secondary 4-digit	2458	21	18.6	2.1 (0.0)	1.17
Matching at primary 12-digit and secondary 12-digit	1882	20.2	20.1	0.0 (1.0)	1

Figure 10. The localization of patent citations and technological closeness (Breschi & Lissoni, 2009).

Although the results displayed in Figure 10, Breschi & Lissoni still do believe that patent citations can indeed be localized, following the rationale of mobile inventors. If this is the case, the percentage of mobile inventors and co-invention networks in citation linkages of the control patents should increase when the technological focus gets narrower. The rationale here is that the narrower the technological focus, the easier it is for an inventor to cite a patent from its network of co-inventors. Breschi & Lissoni tested their assumption, the results are displayed in Figure 11. As can be seen in Figure 11, the narrower the technological focus, the higher the percentage of connected patents among the cited control patents. So, the fact that the localization effects diminish is explained through the underlying mechanism of connected patents. However, although connected inventors do matter, the matter of spatial distance has not yet been proven. As can be seen in the right column in Figure 11, when the technical focus gets narrower, the average path length between connected patents diminishes. In that way, the distance in a network does play a significant role.

	Percentage of connected patents by (mobility or co-invention nw.)		Average path length between connected patents	
	Citing-cited	Control-cited	Citing-cited	Control-cited
Matching at primary 4-digit (JTH experiment)	30.9	19.8	9.2	10.5
Matching at primary 12-digit	28.5	22.1	8.3	9.8
Matching at primary 12-digit and secondary 4-digit	31.1	25.2	8	8.5
Matching at primary 12-digit and secondary 12-digit	31.4	27.1	7.1	7.8

Figure 11. Technological focus and the mobility of inventors (Breschi & Lissoni, 2009).

2.3 Summary Literature Review

The literature related to this thesis was divided into two chapters. In the first chapter, the focus was laid on the emergence of open innovation and ecosystems. Here, the empirical case of this research; Brainport Eindhoven and the HTCE, were studied. Besides that, it was elaborated on how various forms of proximity affect knowledge diffusion concerning innovation, ecosystems, Brainport Eindhoven, and the HTCE. The key message here is that proximity-related conditions at Brainport Eindhoven and the HTCE enhance knowledge diffusion within the ecosystem.

The second chapter elaborated on patent citations as a measure of knowledge flows. Although patent citations are frequently used as a measure of knowledge flows, scholars addressed the limitations and downsides of patent citations as a proxy to measure knowledge diffusion. The large share of examiner citations in patent citation pools, as found by Alcácer & Gittelman (2006), Criscuolo & Verspagen (2008), and Alcácer et al. (2009), make it questionable whether patent citations represent true knowledge spillovers. Next to that, it was found that patent citations are biased for both inventor -and examiner citations (Alcácer & Gittelman, 2006; Corsino et al., 2019). The last part of the chapter elaborated on the literature that studied the localization of patent citations. Here, the key message is that the findings in the pioneering work by Jaffe et al. (1993) are mediated by the mobility of inventors (Breschi & Lissoni, 2009).

3. Methodology

3.1 Method

3.1.1 Research Design

This section will elaborate on the method that was used to study whether patent citations, that originate from cited ecosystem patents, are spatially localized. This research did so by comparing the geographical distances of patent citations, originating from cited ecosystem patents to the geographical distances of patent citations that originate from cited 'non-ecosystem' patents. Patent data from Brainport Eindhoven and the HTCE represented two different groups of 'ecosystem' patents, also referred to as "treated patents". The set of control patents, which were matched to the ecosystem patents, consisted of similar patents originating from places elsewhere within the Netherlands and Belgium. Like in the studies by Jaffe et al. (1993), Breschi & Lissoni (2009), and Castaldi et al. (2015), the control patents were matched on a similarity in the technological classification -and year of application to the peer ecosystem patents. Finally, in addition to the descriptive statistics, with the use of OLS regression modeling, it was tested whether patent citations that originate from cited ecosystem patents were spatially more localized compared to their peer control patents. In the next sections, it will be further elaborated on how the data was collected and treated. The research design that is described above is schematically illustrated in Figure 12. On the next page, based on the research design, for each hypothesis that was described in the Chapter 1.2, a visual illustration is presented (Figures 13 – 15).

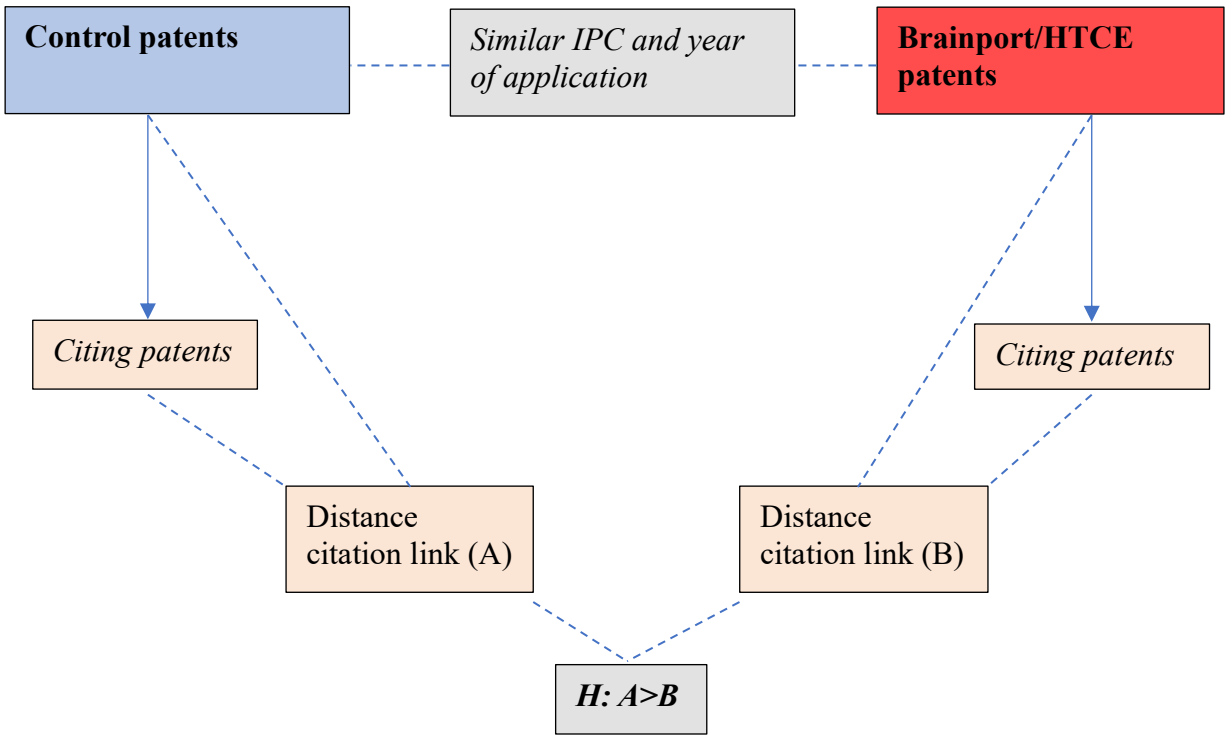


Figure 12. Schematic representation of the research design (H = Hypothesis).

3.1.1.1 Research Design: Hypotheses

Figures 13 – 15 display a visual representation of the hypotheses in correspondence with the research design. Each arrow represents a citation link of a cited -and citing patent. The red arrows represent patent citations of the cited patents that either originate from Brainport Eindhoven or the HTCE, whereas the blue arrows represent the citation links that originate from the cited control patents.

Hypothesis 1: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized compared to patent citations originating from cited 'non-ecosystem' patents.

Figure 13. Visual representation of hypothesis 1.

Hypothesis 2: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized within the ecosystem compared to patent citations originating from cited 'non-ecosystem' patents.

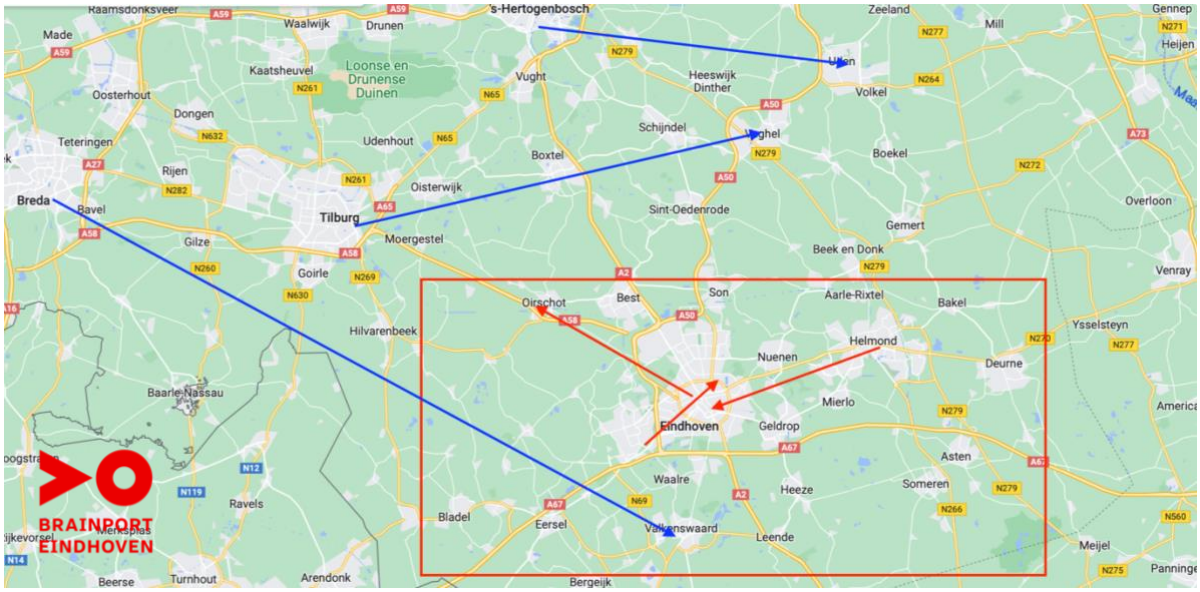


Figure 14. Visual representation of hypothesis 2.

Hypothesis 3: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized within -and around the ecosystem compared to patent citations originating from cited 'non-ecosystem' patents.

Figure 15. Visual representation of hypothesis 3.

3.1.2 Research Design: Measuring and Interpreting Distance

With respect to the method, it is acknowledged and considered that various methods can be applied to study whether ecosystem citations are localized since distance and localization are relative concepts. Keeping this in mind, it was chosen to test the three hypotheses using three different distance metrics. In Table 1, for each distance metric, it is explained for what hypothesis the distance metric was used in the empirical analysis.

Measure	**Hypothesis**	**Focus**
Average distance of citation links	Hypothesis 1	Diffusion of ecosystem knowledge in general.
Shortest citation link	Hypothesis 1	Diffusion of ecosystem knowledge in general with a focus on the most local citation link.
dist_30, dist_50	Hypothesis 2	Diffusion of knowledge within ecosystems.
dist_100 – dist_1000	Hypothesis 3	Diffusion of knowledge within -and around ecosystems.

Table 1. Overview of the distance measures.

***Measure I**. Average distance of all citations per patent.*

In this approach, for each treated -and control patent, the average distance of all citation links was calculated. In that way, a comparison could be made between the average distance of citation links for each pair of treated -and control patents. The downside of this approach is that the effect of localization is overshadowed by the presence of distant citation links. For example, consider a treated patent that has three citation links. The first two citation links cover a distance of approximately 20 kilometers, e.g., the distance from Eindhoven to Tilburg, whereas the third citation link represents a citation link from Eindhoven to China, e.g., 6000 kilometers. When computing the average distance of all three citation links, the average distance will be relatively large, even though a local citation is present. The localization effect of the two localized citations gets overshadowed by the distant citation link. Nevertheless, the presence of local citations will decrease the average distance of all citation links. This measure was used to test hypothesis 1 (Table 1), to examine whether ecosystem citations are in general more localized.

***Measure II**. The shortest distance of all citation links.*

With respect to the second measure, for each treated -and control patent, the shortest citation link of the patent was found. The shortest citation link indicates the shortest path of the knowledge diffused from one patent to the other. Compared to the average distance of all citation links, this approach is a better proxy to study the localization of patent citations, since distant citation links are excluded in this approach. Nevertheless, if a particular patent does not have any relatively short citation, the shortest citation can still comprehend a relatively long distance. Next to that, by picking only the shortest citation link, the other more 'distant' citation links are neglected in the data. Like the previous measure, this measure was used to test hypothesis 1 (Table 1).

***Measure III**. Track and flag the number of citations within a specific range.*

In the third approach, the number of citation links within a particular range of distance was quantified. For example, in this research, for both the treated -and control patents, the number of citation links within a range of 30, 50, 100, 200, 500, 1000, 2500, and 5000 kilometers were counted. In that way, a comparison could be made between the treated and control groups with respect to the number of citation links within a specific range. For each range a variable was created, named *dist_**. Here, *dist* represents the word distance and the star indicates within what range the citations were counted. The ranges of 30 and 50 kilometers were chosen since they represent, to a large extent, knowledge flows within the ecosystem of Brainport Eindhoven. However, they could also include knowledge flows that originate from the ecosystem and move elsewhere outside the ecosystem. The measure to track and flag the number of citations within the range of 30 and 50 kilometers was used to test hypothesis 2.

Thirdly, the ranges up to a thousand kilometers were chosen since the corresponding data about the citation links would allow for a detailed examination of localization effects around the ecosystems. Therefore, these ranges were used to test hypothesis 3. To be more specific, the

ranges of 100, 200 and 500 kilometers were studied in most detail. The radius of 200 kilometers around the city of Eindhoven captures the entire Benelux and some major German cities (Figure 16). The range of 2500 kilometers was chosen since this radius represents, more or less, the geographical borders of Europe. The threshold of 5000 kilometers was chosen to exclude patent data from the Unites States of America (USA) and China.

Figure 16. Radius of 200 kilometers around Eindhoven.

3.2 Data Collection Methods

3.2.1 Overview of Datasets and Software

Before elaborating on the methods used concerning the collection and treatment of data, first, it will be briefly reasoned what datasets and software were used. To obtain the necessary geographic information about patents, the dataset by de Rassenfosse et al. (2019) was used. To obtain other relevant information about patents, it was chosen to use the database of PATSTAT. Since de Rassenfosse et al. (2019) made use of a unique application identity number for each patent, which corresponds to the PATSTAT database, it was possible to link these two databases. The software of Microsoft SQL was used to treat the data. Through the Department of IE&IS at the University of Eindhoven, I got access to a remote computer, named "REX", which gave me access to SQL in which the datasets of de Rassenfosse et al. (2019) and PATSTAT 2021 were imported. Appendix A covers a transcript of all the lines of code that were used during the treatment of data in SQL. The structure of Appendix A resembles and corresponds with the structure of this report.

3.2.2 Initial Dataset

For the collection of data, the objective was to collect all the patents that have been filed within Brainport Eindhoven and the HTCE. Among the options, either search the data manually or use

an existing dataset, it was chosen to use an existing dataset. In this research, the dataset by de Rassenfosse et al. (2019) was used. In their study, de Rassenfosse et al. had the following objective: "The geocoding of worldwide patent data". In more detail, for each first filing of a patent, de Rassenfosse et al. found the corresponding geographical location (GPS coordinates) of the inventor(s). The dataset consists of 18.8 million first filings, invented in 46 countries for the period 1980-2014. The dataset has a high accuracy: it covers 81 percent of all the first filings worldwide for the period mentioned above. Due to the availability of GPS coordinates in the dataset by de Rassenfosse et al., there was an opportunity for this research to use these coordinates to study the localization of patent citations. Alternative to the dataset by de Rassenfosse et al., the option was to collect the ecosystem patents, control patents, citing patents and the coordinates, manually. However, due to time constraints, the usage of the dataset by de Rassenfosse et al. was the best option for this research to study its objective. De Rassenfosse et al. created multiple data files, each having a different focus. For this research, the data file "geoc_app.txt" had been used[1].

For this research, it was important to distinguish between the location of the inventor and the applicant since they are not the same. For example, it might be the case that the location of any applicant is in the Netherlands, let's say Eindhoven, whereas the location of the inventors is abroad. This implies that, although the invention is filed in the Netherlands, the invention itself is not invented in the Netherlands.

Since this research is about knowledge diffusion, it was important to consider the location of the inventors. De Rassenfosse et al. determined the geographical location of the patents' origin using the location of the inventors. Therefore, the dataset was applicable and suitable for this study. In the process of finding the corresponding addresses of the inventors, de Rassenfosse et al. first used the inventors' location of the first priority. If no information was found, the earliest equivalent was used, which is a second filing that refers to the first filing. Thirdly, if no information was found in that way, other equivalents of the same patent family were used. If no information was found there, the location of the applicant was used. Eventually, de Rassenfosse et al. converted the addresses of the inventors into GPS coordinates.

3.2.3 Collection of the Ecosystem Patents and Citing Patents

3.2.3.1 Identification of the Area

Since the dataset by de Rassenfosse et al. (2019) contains the GPS coordinates of all the patents present in the dataset, it was possible to filter out the patents of Brainport Eindhoven and the HTCE very accurately. In Figures 17 and 18, the border of the HTCE and Brainport Eindhoven are indicated by the black squares. Within this square, all the patents that originate from Brainport Eindhoven and the HTCE can be found. To be more specific, within these black squares, all the patents that have at least one inventor, whose coordinates are within either Brainport Eindhoven or the HTCE, can be found. Using Google Maps, the coordinates of the 4 corners were identified. These coordinates are marked red in Figures 17 and 18.

[1] https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/OTTBDX/VWIK0D&version=5.1

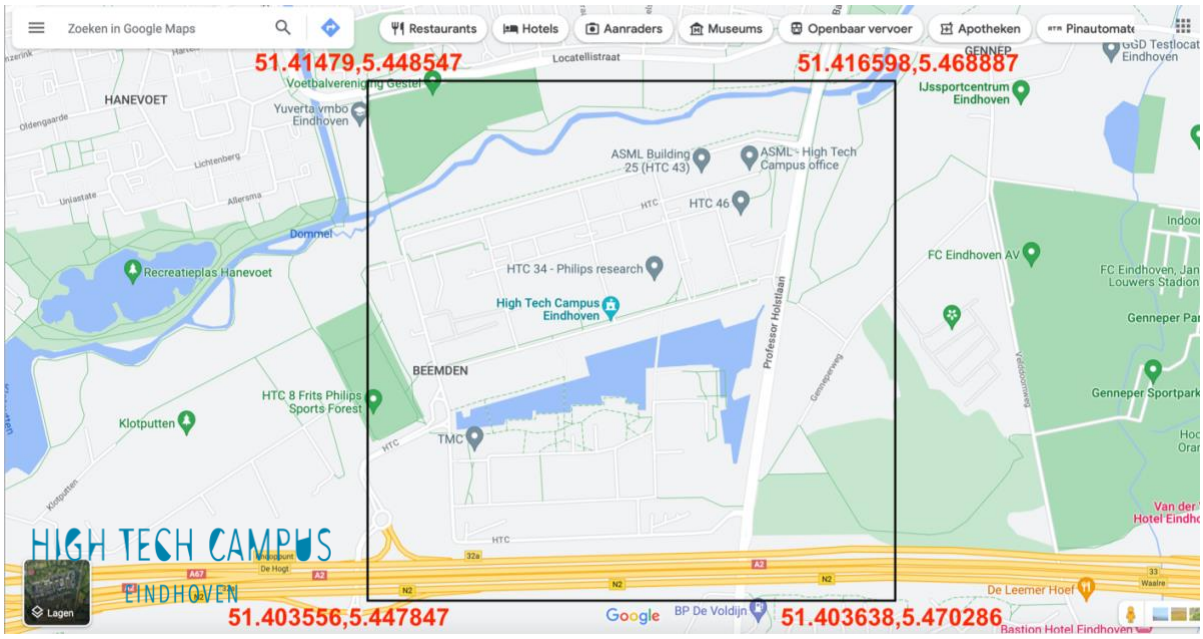


Figure 17. Geographical borders of the HTCE.

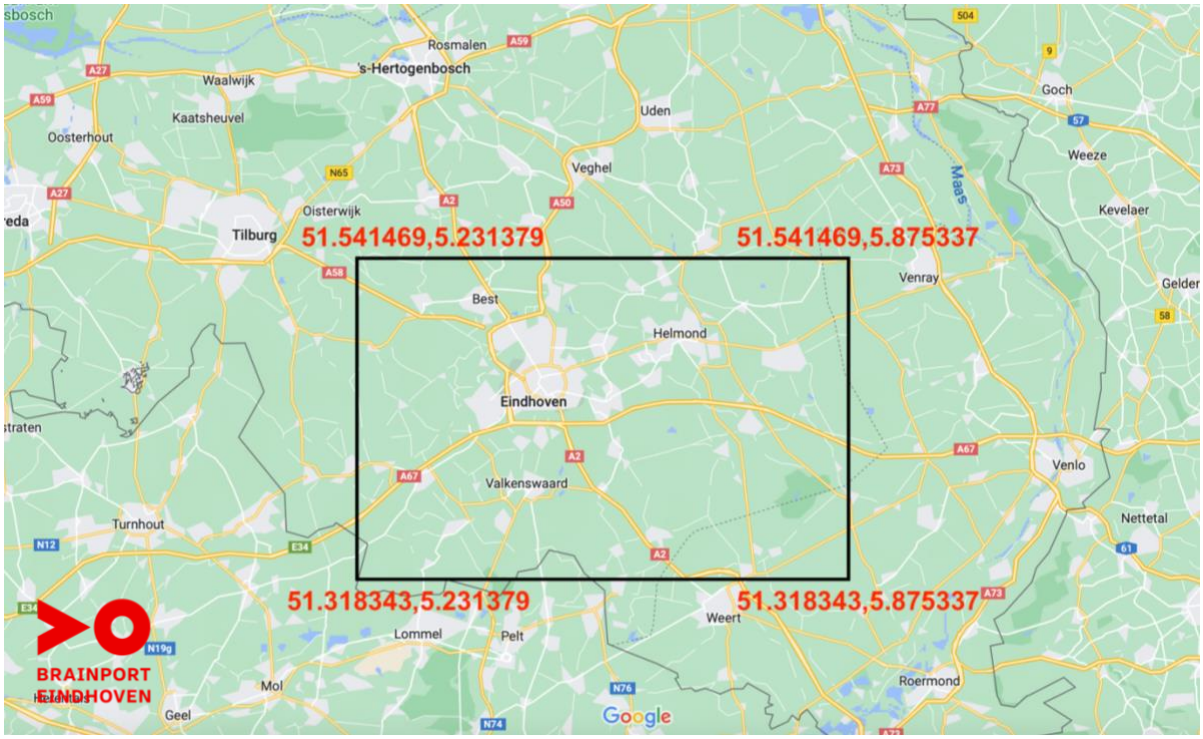


Figure 18. Geographical borders of Brainport Eindhoven.

3.2.3.2 Collection of the Ecosystem Patents

With the use of these coordinates, a conditional search query (Appendix A) was created to retrieve the patents of Brainport Eindhoven and the HTCE from the dataset by de Rassenfosse et al. (2019). Here, it had to be considered that the HTCE was founded in the year 2003, whereas the dataset by de Rassenfosse et al. contains all the first filings from the period 1980-2014. Besides, the citation lag of patent citations was considered. Therefore, patent data from Brainport Eindhoven and the HTCE was collected for the years 2003-2010. In that way, the citation lag was set for a minimum of 4 years, up to 2014. Any citing patent that was filed after 2014 could not be included. With a citation lag of 4 years, it was assumed that approximately half of the citing patents will be covered in the data (Jaffe et al., 1993).

After the ecosystem patents were found, a query was formulated to retrieve the citing patents of the corresponding ecosystem patents. To find this information, the PATSTAT 2021 database was used. PATSTAT is the worldwide patent statistic database of the European Patent Office (EPO). After the citing patents were found, self-citations were identified and removed from the data. Similar to Jaffe et al. (1993) and Breschi & Lissoni (2009), in this research, a citation is considered a self-citation when both the cited -and the citing patent are filed by the same organization.

Eventually, using the database of de Rassenfosse et al. (2019), the geocoordinates of the citing patents were found and connected to the geocoordinates of the cited patents. In that way, the distance of a citation link could be calculated. Appendix A displays statistics about the number of patents and citation links that are present in the data. As can be seen, due to the exclusion of self-citations, and the limited availability of data in the dataset by de Rassenfosse et al. (2019), the number of unique ecosystem patents, for both the Brainport Eindhoven group and the HTCE group, decreased.

3.2.4 Collection of the Control Patents and Citing Patents

As mentioned earlier in this chapter, the control patents were matched to the ecosystem patents based on a similarity in technological classification and year of application. To be more specific, concerning technological classification, the match is based on a 4-digit IPC (International Patent Classification) code, like the study by Breschi & Lissoni (2009). With respect to the number of digits, a trade-off was faced. On the one hand, the higher the number of IPC digits, the narrower the technological closeness between the treated -and the control patent. However, the higher the technological closeness, the higher the difficulty to find a pairing match for each treated patent. Therefore, a match based on similarity in a 4-digit IPC code was chosen to find an appropriate balance between technological closeness and the likelihood to find a control patent for each ecosystem patent.

To facilitate a match between the ecosystem patents and the control patents, for each patent, the most frequent corresponding 4-digit IPC code was found. Since one patent can have more than one IPC code, the most frequent 4-digit IPC was counted, selected, and allocated to the patent. In case a patent had multiple codes with the same frequency, either one of the codes was randomly selected. This part of the data treatment is displayed in Appendix A.

Since the dataset by de Rassenfosse et al. (2019) consists of patent data about first filings, with respect to the similarity of the application year between the treated -and control patents, it was chosen to use the filing year of the earliest application, a variable that is featured in the PATSTAT dataset.

Before the control patents were matched to the ecosystem patents, first, a pool of possible control patents was created. As mentioned at the beginning of this chapter, all other patents that originate from either the Netherlands or Belgium, though outside the Brainport region, founded a pool of possible control patents for the Brainport Eindhoven patents and the HTCE patents. Regarding the pool of control patents for the HTCE patents, it was chosen to select patents outside the Brainport region as well since otherwise some control patents could originate from the Brainport region, and thereby ecosystem patents would be part of the control group for the HTCE.

With the use of a search query (Appendix A), the possible pool of patents was created. After the creation of the patent pool, the 4-digit IPC code and the earliest year of application were assigned to both the ecosystem patents and the control patents. For some of the ecosystem patents, multiple possible control patents were found. Eventually, among all the possible control patents, for each ecosystem patent, one control patent was selected at random and allocated to the ecosystem patent. In that way, for each ecosystem group, a set of control patents was created. After, the citing patents and other relevant information of the control patents were found. In Appendix A, the number of observations of the dataset for both the treated -and control patents can be found.

3.2.5 Matching and Weights

As mentioned in the previous paragraph, for each Brainport or HTCE patent, a control patent was matched at random. However, due to the decay of data along the process, the balance between the treated -and control patents was not intact. For example, there could be a control patent present in the final set of control patents that has no longer a peer ecosystem patent, because the ecosystem patent got excluded because of either the presence of self-citations or because missing data in the dataset by de Rassenfosse et al. (2019). Therefore, with the final set of treated -and control patents, for both the Brainport Eindhoven and the HTCE group, it was examined to what extent the control patents resembled with the treated patents. To be more specific, for each unique combination of the 4-digit IPC code and the earliest year of application, the number of corresponding ecosystem -and control patents were counted. In that way, a frequency weight could be created to balance the share of the treated -and the control patents per unique combination of the 4-digit IPC code and the earliest year of application. For example, consider the SQL output displayed in Figure 19. For the combination “H04H and 2003”, displayed in row 76 in Figure 19, the number of corresponding Brainport patents is one, whereas the number of controls present is two. Therefore, the Brainport patents with the combination “H04H and 2003” received a frequency weight of ‘1’, whereas the control patents with the combination “H04H and 2003” received a frequency weight of ‘0.5’.

100 %

Results | Messages

	IPC	earliest_filing_year	fq_combination_brainport	fq_combination_control	weight_control
76	H04H	2003	1	2	0,5
77	H04J	2003	2	2	1
78	H04L	2003	15	18	0,833333333
79	H04M	2003	2	1	2
80	H04N	2003	17	9	1,888888889
81	H04...	2003	16	7	2,285714286
82	A47J	2004	2	3	0,666666667
83	A47L	2004	2	1	2
84	A61B	2004	8	8	1
85	A61F	2004	3	2	1,5
86	A61N	2004	3	2	1,5
87	B23B	2004	1	1	1
88	B29C	2004	1	2	0,5
89	B32B	2004	2	2	1

Figure 19. Assessing weights to the control patents.

3.2.6 Calculation of the Distance of the Citation Links

3.2.6.1 Calculating Distance

The distance of a citation link was calculated using the formula of Haversine. With this formula, the distance between two coordinates can be calculated. This formula was used since the Haversine formula considers the curvature of the earth and thereby the distance between two points on a sphere could be calculated. Before the Haversine formula could be applied, first, the values of latitude and longitude that were found in the data by de Rassenfosse et al. (2019), had to be converted from degrees into radians (Equation 1). Down below, the Haversine formula is illustrated (Equation 2). The symbols of φ and λ represent the latitude and longitude in radians. For each citation link, the distance was calculated using Microsoft Excel.

$$rad = \frac{degrees * \pi}{180}$$

Equation 1. Converting degrees into radians.

$$d = 2r \arcsin\left(\sqrt{sin^2\left(\frac{\varphi 2 - \varphi 1}{2}\right) + \cos(\varphi 1) \cdot \cos(\varphi 2) \cdot sin^2\left(\frac{\lambda 2 - \lambda 1}{2}\right)}\right)$$

$$\varphi = latitude$$
$$\lambda = longitude$$
$$r = radius\ of\ the\ earth = 6371\ kilometers$$

Equation 2. Haversine formula.

3.2.6.2 Shortest distance per citation link

Since the geocoordinates of patents in the dataset by de Rassenfosse et al. (2019) are based on inventor locations, per patent, multiple locations can be found. This implies that for each citation link, multiple distances can be calculated. Next to that, it could be the case that for some patents the inventors are located outside the Brainport region or even outside Belgium and the Netherlands. For example, a particular patent is both assigned to an inventor that is located within the Brainport region and to an inventor that is located outside the Brainport region. Due to the likelihood of this, in the end, it has been checked that only the inventor locations of patents within the Brainport Region, or Belgium and Netherlands, would be included in the final dataset. In the end, for each citation link, having multiple calculated distances, the shortest distance was picked since it represents the shortest path of the knowledge flow.

With this approach, it had to be kept in mind that the shortest citation link that was picked does not necessarily represent the 'true' shortest citation link. It could be the case that the true shortest citation link of a particular patent originated from an inventor location that was excluded. For example, imagine a patent that is filed by a Chinese company and is invented by five Chinese inventors and one Dutch inventor from Eindhoven. Since the patent is to a large extent Chinese, with the rationale of localized knowledge diffusion, it is more likely that the forward citations originate from China as well. However, when calculating the distances of the citation links using the location of the Dutch inventor, the distances of the citation links are relatively large, which might produce a distorted and "noisy" picture of the diffused knowledge.

3.3 Variables

This section on variables illustrates what variables were used in the empirical part of this research. Table 3 shows a snapshot of the dataset that was created for the Brainport Eindhoven patents and the corresponding control patents. Table 2 only includes the variables that were used for the descriptive statistics and multivariate regressions. Table 3, on the other hand, does include other demographic features of the patents that were not used for any of the results discussed later in this thesis.

In Table 2, each variable is explained by a short description. The role of each variable in the empirical research is indicated by the column *type*. In the statistical analysis, the distance variables were used as dependent variables. The dummy variable *treated* acted as the independent variable. This dummy variable indicates whether a patent is an ecosystem patent or not.

Besides the dependent -and independent variables, control variables were considered and included. These control variables were important to validate the relationship between the independent -and dependent variables. To be more specific, variables were included that were likely to have a possible relationship with the dependent variables. In that way, the effect of the ecosystems was examined while controlling for other factors that might influence the prediction

on distance. Most of the control variables are based on descriptive information that is found in patent applications.

Variable	***Description***	***Type***
avg_haversine	The sum of the average distance of all forward citation links per unique patent.	**DV**
shortest_citation	The shortest forward citation link per patent.	**DV**
*dist_**	A flag that indicates the number of forward citation links within a particular range per patent.	**DV**
*Dist_larger_**	A flag that indicates the number of forward citation links per patent outside a particular range.	**DV**
nb_applicants_companies	The number of companies registered as applicant.	**DV**
treated	A flag (dummy variable) indicating whether a patent is a treated patent or not. The value of '1' indicates that a patent is an ecosystem patent. The value '0' indicates that a patent is a control patent.	**IV**
name	The name of patent's applicant.	Control
earliest_filing_year	The earliest year of filing for the patent application.	Control
nb_citations	The number of forward citation links per patent.	Control
nb_foreign_inventors	The number of inventors outside the Netherlands and Belgium.	Control
nb_outside_inventors	The number of inventors outside the HTCE/brainport.	Control
IPC	The 4-digit IPC code of the patent.	Control
nb_ipc_codes	The number of IPC codes assigned to a patent.	Control
earliest_filing_year	The earliest year of filing for the patent application.	Control
nb_backward_citations	The number of backward citations.	Control
nb_forward_citations	The number of forward citations, including those forward citations after 2014.	Control
nb_cited_literature	The number of cited literature documents.	Control
nb_publn_claims	The number of publication claims.	Control

DV = Dependent variable
IV = Independent variable

Table 2. Description of the variables.

3.4 Statistical Analysis

Appendix A reports the process of how the final datasets, the HTCE -and control patents (i) and the Brainport Eindhoven -and control patents (ii), were assembled. After the assembly of the final datasets, the basis of the analytical model was founded. First, for each hypothesis and measure of distance (Table 1), the means of the treated -and control group were compared with the use of a t-test. However, since a t-test does not allow the inclusion of other predictor variables, it was chosen to test the hypotheses with the use of Ordinary Least Squares (OLS) regression. As this research includes multiple dependent variables and multiple predictor variables (Table 2), the model in this research can be considered a Multivariate Multiple Regression (MMR). Equation 3 illustrates the formula of the MMR model. The use of MMR is essential to examine the potential effects of other predictor variables. Therefore, MMR is essential to validate any of the results found in this research. Among all the predictor variables, listed in Table 2, the variable *treated* is the variable of interest. The variable *treated* measures the effect of the ecosystem on the distance of patent citation links. For the data analysis, the statistical software *Stata ME* was used.

$$\gamma = \beta 0 + \beta 1x1 + \beta 2x2 + .. + \epsilon$$
$$\gamma 1 = \beta 0 + \beta 1x1 + \beta 2x2 + .. + \epsilon$$
$$\gamma 2 = \beta 0 + \beta 1x1 + \beta 2x2 + .. + \epsilon$$

$$\beta 0 = y - intercept\ (constant\ term)$$
$$\beta = slope\ coefficent\ for\ each\ explanatory\ variable$$
$$x = independent\ variable$$
$$y = dependent\ variable$$
$$\epsilon = residuals$$

Equation 3. Formula multivariate multiple regression.

3.4.1 Fixed Firm Effect.

By including the names of the companies in the analytical model, the effect of firms on the distance of citation links could be examined. It could be the case that a particular firm, located in Brainport Eindhoven, has a lot of local forward citation links. Consequently, it could be that the effect of localization is not explained by the ecosystem, but rather by the influence of the individual firm. This phenomenon is also referred to as the “fixed firm effect”. Therefore, the effect of the individual firm was incorporated into the MMR model. In Stata, all the patents that were filed by the same company were grouped (Appendix B). In that way, the effect of each group, representing a single firm, was tested. The same technique was applied to the variables *earliest_filing_year* and *IPC,* to account for the fixed effect of a particular technology class and the year of application.

In sum, three different control groups were created representing the effect of the individual firm, the effect of the year of application, and the effect of the 4-digit IPC code. In the next chapter,

first, relevant descriptive statistics of both datasets will be presented. After that, the results of the regression models will be presented.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	appln_id	avg_haversin	nb_citations	shortest_cita	dist_0	dist_30	dist_50	dist_100	dist_200	dist_500	dist_1000	dist_2500	dist_5000	dist_larger_2500	dist_larger_5000	appln_auth	nb_inventor	nb_applican	name	ctry_code	province	nb_foreign_i	nb_outside_i	IPC	nb_ipc_code	earliest_filing	nb_backward	nb_forward_	nb_cited_lite	publn_claim	fq_matches	treated	weight	
2	8873	4898	4	421,6096	0	0	0	0	0	1	2	2	2	2	2	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G01R	1	2006	4	12	2	0	3	1	1	
3	60832	708	1	707,64917	0	0	0	0	0	0	1	1	1	0	0	WO	11	1	Janssen Pharmaceutica N.V.	NL	Noord-Braba	4	8	A61K	2	2007	6	28	0	0	2	1	1	
4	76089	213	1	213,14382	0	0	0	0	0	1	1	1	1	0	0	WO	3	2	PHILIPS INTELLECTUAL PROPER	NL	Noord-Braba	1	1	H03F	2	2007	2	1	0	0	1	1	1	
5	120179	6680	1	6680,3071	0	0	0	0	0	0	0	0	0	1	1	WO	2	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G01S	1	2007	3	5	1	0	7	1	1	
6	143190	2	1	1,6433811	0	1	1	1	1	1	1	1	1	0	0	EP	4	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	1	2002	3	3	0	11	7	1	1	
7	190871	583	1	583,37311	0	0	0	0	0	0	1	1	1	0	0	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H01J	3	2007	6	2	0	14	5	1	1	
8	208855	5873	3	577,46306	0	0	0	0	0	0	1	1	1	2	2	EP	2	1	NXP B.V.	NL	Noord-Braba	0	1	H01L	4	2003	7	4	2	16	37	1	1	
9	237297	9063	1	18716,126	0	0	0	0	0	0	0	0	0	1	1	WO	2	2	Koninklijke Philips Electronics	NL	Noord-Braba	2	2	G06T	1	2007	5	8	4	0	9	1	1	
10	237299	4004	2	156,88109	0	0	0	0	1	1	1	1	1	1	1	WO	1	2	Koninklijke Philips Electronics	NL	Noord-Braba	1	1	A61N	1	2007	4	7	0	0	4	1	1	
11	271491	8	13	0	1	12	12	13	13	13	13	13	13	0	0	EP	5	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	1	2003	9	25	1	21	15	1	1	
12	305102	3481	6	477,14377	0	0	0	0	0	1	4	4	4	2	2	EP	5	0	NULL	NL	Noord-Braba	1	1	A61F	2	2005	3	16	3	16	3	1	1	
13	370363	9067	1	9067,0003	0	0	0	0	0	0	0	0	0	1	1	WO	2	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H05K	6	2007	9	4	0	0	3	1	1	
14	441389	9351	1	9351,0842	0	0	0	0	0	0	0	0	0	1	1	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	A61B	2	2007	6	4	0	0	15	1	1	
15	6935621	6706	11	5938,5673	0	0	0	0	0	0	0	0	0	11	11	CN	2	1	KONINKLIJKE PHILIPS ELECTRO	NL	Noord-Braba	0	0	H04W	8	2003	0	22	0	0	7	1	1	
16	6946853	9004	1	9003,8542	0	0	0	0	0	0	0	0	0	1	1	CN	3	1	KONINKLIJKE PHILIPS ELECTRO	NL	Noord-Braba	0	0	H04W	12	2003	0	3	0	0	7	1	1	
17	7792154	9258	1	9257,5326	0	0	0	0	0	0	0	0	0	1	1	CN	3	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H04W	1	2003	0	5	0	0	7	1	1	
18	7799838	8557	2	7856,46	0	0	0	0	0	0	0	0	0	2	2	CN	1	1	KONINKLIJKE PHILIPS ELECTRO	NL	Noord-Braba	0	0	H04B	4	2003	0	7	0	0	7	1	1	
19	14928642	394	4	312,53308	0	0	0	0	0	0	3	4	4	4	0	DE	2	1	Robert Bosch GmbH	NL	Noord-Braba	1	1	G01C	3	2004	0	5	0	0	3	1	1	
20	15024140	550	1	549,83732	0	0	0	0	0	0	1	1	1	0	0	DE	0	3	Merck Patent GmbH	NL	Noord-Braba	3	3	G02F	4	2006	0	6	0	0	1	1	1	
21	16016156	7837	5	7836,6185	0	0	0	0	0	0	0	0	0	5	5	EP	0	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	1	2003	3	9	0	11	15	1	1	
22	16017160	443	1	443,42274	0	0	0	0	0	1	1	1	1	0	0	EP	1	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G21K	9	2003	4	5	2	12	2	1	1	
23	16017258	2	4	1,6433811	0	4	4	4	4	4	4	4	4	0	0	EP	2	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G02B	9	2003	3	22	0	24	4	1	1	
24	16017826	6363	1	6362,522	0	0	0	0	0	0	0	0	0	1	1	EP	8	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	1	2002	2	11	1	43	7	1	1	
25	16020183	4535	10	0	4	4	4	4	4	4	4	4	4	6	6	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H04N	4	2003	5	62	0	14	9	1	1	
26	16022623	2	2	1,6433811	0	2	2	2	2	2	2	2	2	0	0	EP	3	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	H01L	5	2002	5	12	0	29	9	1	1	
27	16022893	8826	1	8826,3744	0	0	0	0	0	0	0	0	0	1	1	EP	0	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	2	2003	3	4	0	11	15	1	1	
28	16025431	79	2	79,074505	0	0	0	2	2	2	2	2	2	0	0	EP	2	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	2	2003	3	11	1	8	15	1	1	
29	16110299	1461	7	1,2580734	0	3	3	3	3	6	6	6	6	1	1	EP	9	2	CARL ZEISS SEMICONDUCTOR M	NL	Noord-Braba	1	1	G01N	2	2003	6	23	0	29	4	1	1	
30	16110801	223	1	223,37764	0	0	0	0	0	0	1	1	1	1	0	EP	3	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	1	2003	5	1	1	11	15	1	1	
31	16110848	4684	2	442,96875	0	0	0	0	0	0	1	1	1	1	1	EP	2	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G03F	1	2003	5	8	0	16	15	1	1	
32	16111033	8697	1	8696,759	0	0	0	0	0	0	0	0	0	1	1	EP	1	1	AutoMedic B.V.	NL	Noord-Braba	0	0	A61F	1	2004	6	6	0	15	2	1	1	
33	16111784	176	5	1,6433811	0	3	3	3	3	5	5	5	5	0	0	EP	2	1	ASML Netherlands B.V.	NL	Noord-Braba	0	0	G02B	2	2003	4	20	0	30	4	1	1	
34	16113739	7942	2	7942,1914	0	0	0	0	0	0	0	0	0	2	2	EP	0	1	Koninklijke Philips Electronics,	NL	Noord-Braba	0	0	G06T	1	2004	3	2	1	14	4	1	1	
35	16205674	8591	2	7845,6841	0	0	0	0	0	0	0	0	0	2	2	EP	1	0	NULL	NL	Noord-Braba	0	0	E04G	2	2005	8	4	0	19	1	1	1	
36	16213900	2	1	1,9779624	0	1	1	1	1	1	1	1	1	0	0	EP	0	2	FEI COMPANY	NL	Noord-Braba	1	1	G01N	1	2005	2	3	1	21	6	1	1	
37	16304865	9256	1	9256,2501	0	0	0	0	0	0	0	0	0	1	1	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	D06F	1	2006	8	7	0	15	7	1	1	
38	16311751	581	1	580,94157	0	0	0	0	0	0	1	1	1	0	0	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	D06F	1	2006	3	14	0	13	7	1	1	
39	16314559	1582	1	1582,1973	0	0	0	0	0	0	0	1	1	0	0	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G02F	3	2006	9	7	1	19	1	1	1	
40	16405641	565	1	564,74527	0	0	0	0	0	0	1	1	1	0	0	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H02M	1	2007	4	1	1	9	2	1	1	
41	16406277	721	2	720,18183	0	0	0	0	0	0	2	2	2	0	0	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	A61K	5	2007	11	18	13	17	2	1	1	
42	16406843	714	1	714,19254	0	0	0	0	0	0	1	1	1	0	0	EP	0	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	C12Q	1	2007	11	4	6	21	2	1	1	
43	17530723	6541	11	1540,447	0	0	0	0	0	1	2	3	3	8	8	WO	11	2	Koninklijke Philips Electronics	NL	Noord-Braba	9	9	G10L	2	2006	1	64	2	0	2	1	1	
44	20939965	3214	1	3213,6003	0	0	0	0	0	0	0	0	1	1	0	WO	2	1	France Telecom	NL	Noord-Braba	2	2	H04W	2	2006	3	3	0	0	2	1	1	
45	24015127	7701	2	6767,9594	0	0	0	0	0	0	0	0	0	2	2	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G06F	2	2002	6	4	0	0	2	1	1	
46	24016480	290	1	289,63341	0	0	0	0	0	0	1	1	1	1	0	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G11B	3	2002	5	9	5	0	2	1	1	
47	24029518	16531	1	16530,964	0	0	0	0	0	0	0	0	0	1	1	WO	2	2	Koninklijke Philips Electronics	NL	Noord-Braba	2	2	A61B	3	2003	3	27	3	0	10	1	1	
48	24033660	8894	1	8894,0273	0	0	0	0	0	0	0	0	0	1	1	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H01J	2	2003	6	4	1	0	5	1	1	
49	24034498	5627	1	5627,0536	0	0	0	0	0	0	0	0	0	1	1	WO	5	1	Koninklijke Philips Electronics	NL	Noord-Braba	1	1	H03F	3	2003	4	2	0	0	3	1	1	
50	24034566	8420	7	6800,8533	0	0	0	0	0	0	0	0	0	7	7	WO	3	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H01L	4	2003	2	20	0	0	37	1	1	
51	24034849	9230	1	9230,0248	0	0	0	0	0	0	0	0	0	1	1	WO	5	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	A61B	2	2003	4	4	0	0	10	1	1	
52	24035232	4685	2	311,30367	0	0	0	0	0	1	1	1	1	1	1	WO	2	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G01R	2	2003	3	11	0	0	3	1	1	
53	24035451	8791	3	8603,4299	0	0	0	0	0	0	0	0	0	3	3	WO	3	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G01N	2	2003	5	3	5	0	4	1	1	
54	24035568	8871	1	8870,8923	0	0	0	0	0	0	0	0	0	1	1	WO	2	2	PHILIPS INTELLECTUAL PROPER	NL	Noord-Braba	1	1	H01L	2	2003	4	3	0	0	37	1	1	
55	24035702	812	3	811,7067	0	0	0	0	0	0	3	3	3	0	0	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H03K	5	2003	6	11	0	0	2	1	1	
56	24036187	568	1	567,97034	0	0	0	0	0	0	1	1	1	0	0	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H01J	4	2004	6	3	1	0	8	1	1	
57	24040865	8884	5	8868,7222	0	0	0	0	0	0	0	0	0	5	5	WO	4	2	Koninklijke Philips Electronics	NL	Noord-Braba	1	1	G09G	2	2004	1	16	0	0	2	1	1	
58	24041082	9259	1	9259,2348	0	0	0	0	0	0	0	0	0	1	1	WO	3	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H04B	4	2004	4	1	4	0	6	1	1	
59	24041097	7835	2	7832,1906	0	0	0	0	0	0	0	0	0	2	2	WO	3	1	Koninklijke Philips Electronics	NL	Noord-Braba	3	3	H05B	1	2004	7	2	0	0	1	1	1	
60	24041348	5404	2	1540,447	0	0	0	0	0	0	0	1	1	1	1	WO	3	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	A61B	3	2004	5	6	0	0	8	1	1	
61	24041350	573	3	572,96177	0	0	0	0	0	0	3	3	3	0	0	WO	1	2	PHILIPS INTELLECTUAL PROPER	NL	Noord-Braba	2	2	H01J	3	2004	6	5	1	0	8	1	1	
62	24041535	3118	2	367,80794	0	0	0	0	0	1	1	1	1	1	1	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G06T	1	2004	2	6	3	0	4	1	1	
63	24041665	8876	1	8875,5736	0	0	0	0	0	0	0	0	0	1	1	WO	1	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	H04N	2	2004	2	7	4	0	10	1	1	
64	24041769	8898	1	8898,2137	0	0	0	0	0	0	0	0	0	1	1	WO	3	2	PHILIPS INTELLECTUAL PROPER	NL	Noord-Braba	2	2	H04M	2	2004	0	3	4	0	3	1	1	
65	24041786	9353	2	9352,5205	0	0	0	0	0	0	0	0	0	2	2	WO	2	1	Koninklijke Philips Electronics	NL	Noord-Braba	0	0	G11B	1	2004	2	3	0	0	17	1	1	

Table 3. Snapshot of the dataset of the Brainport Eindhoven sample.

4. Results

This chapter is structured as follows. First, a section on descriptive statistics will provide basic information about the variables in the dataset. Next to that, these descriptive statistics are used to get a first insight into the relationship between the distance of citation links and ecosystems. In addition to the descriptive statistics, the results of the linear regression models are presented. Again, it is important to keep in mind that the HTCE -and Brainport datasets are not completely different samples. All the HTCE patents are present in the Brainport sample as well. The Brainport region is larger and thereby includes all the firms that are present at the HTCE plus other firms from the region, e.g., ASML, located in Veldhoven. This means that both datasets are partially similar and thereby are likely to have a form of similarity in the results. All the results that will be presented in this chapter were created using the software Stata and Excel. Appendix B reports the lines of code that were used in Stata chronologically with the results displayed in this chapter.

4.1 Descriptive Statistics

4.1.1 Distribution of the Patents per Firm

Figures 20 – 23 display the distribution of the applicants in each separate sample. Figures 20 and 21 represent the HTCE patents and the corresponding control patents. Figures 22 and 23 represent the Brainport patents and the corresponding control patents. Previously in the method, it was explained to account for the effect of firms while studying the relation between the distance of citation links and ecosystems. Since many patents within the HTCE and Brainport Eindhoven do have the same applicant, it is important to understand the distribution of the applicants in the datasets.

As can be seen in Figure 20, almost all patents in the first sample do originate from either Philips or NXP, whereas the composition of the control patents, displayed in Figure 21, is rather dispersed. The same pattern can be found in Figures 22 and 23. The distribution of the applicants in the Brainport sample is to a large extent represented by Philips, ASML, and NXP, whereas the distribution of the firms of the corresponding control patents is more dispersed.

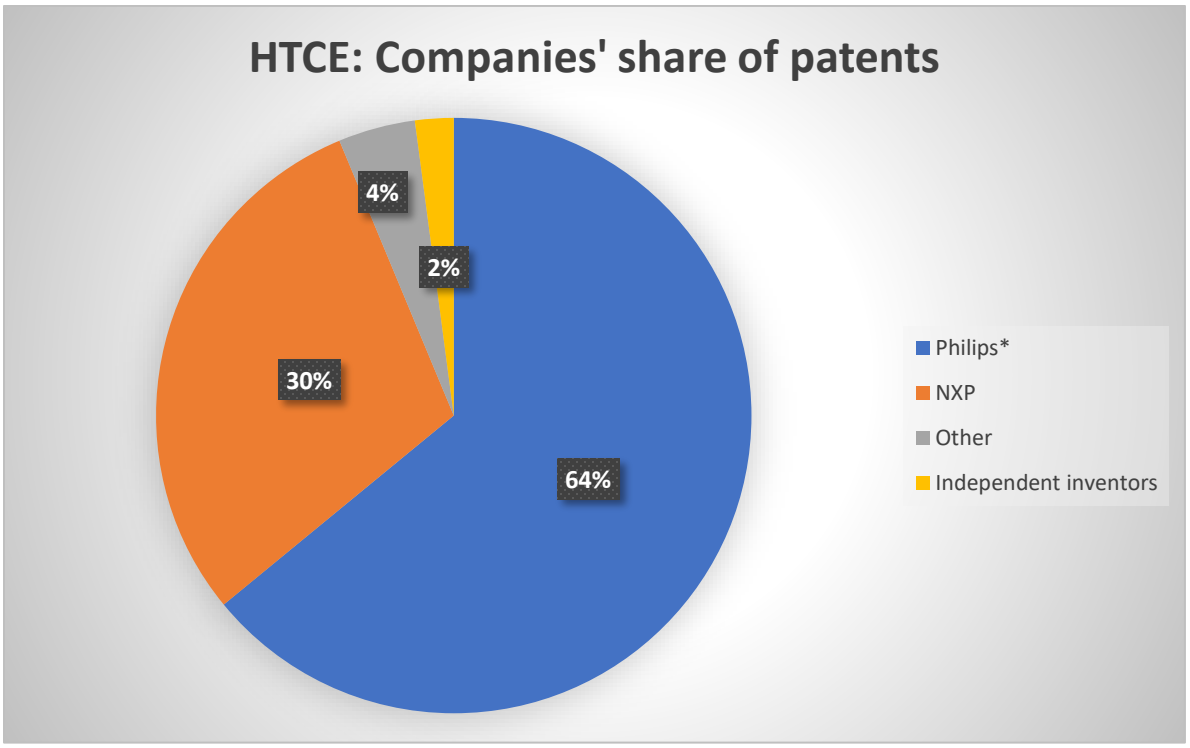


*Figure 20. Distribution of firms in the HTCE sample (N=765). *(all departments of Philips)*

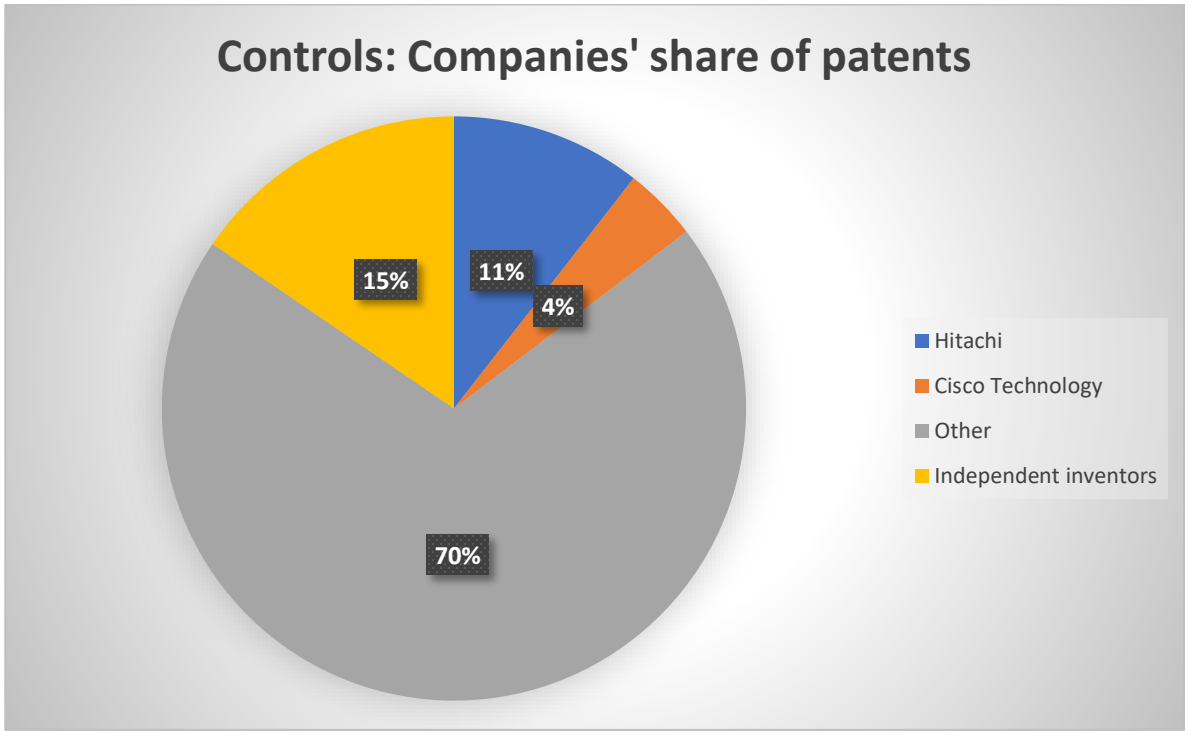


Figure 21. Distribution of firms in the control of group of the HTCE patents (N=853).

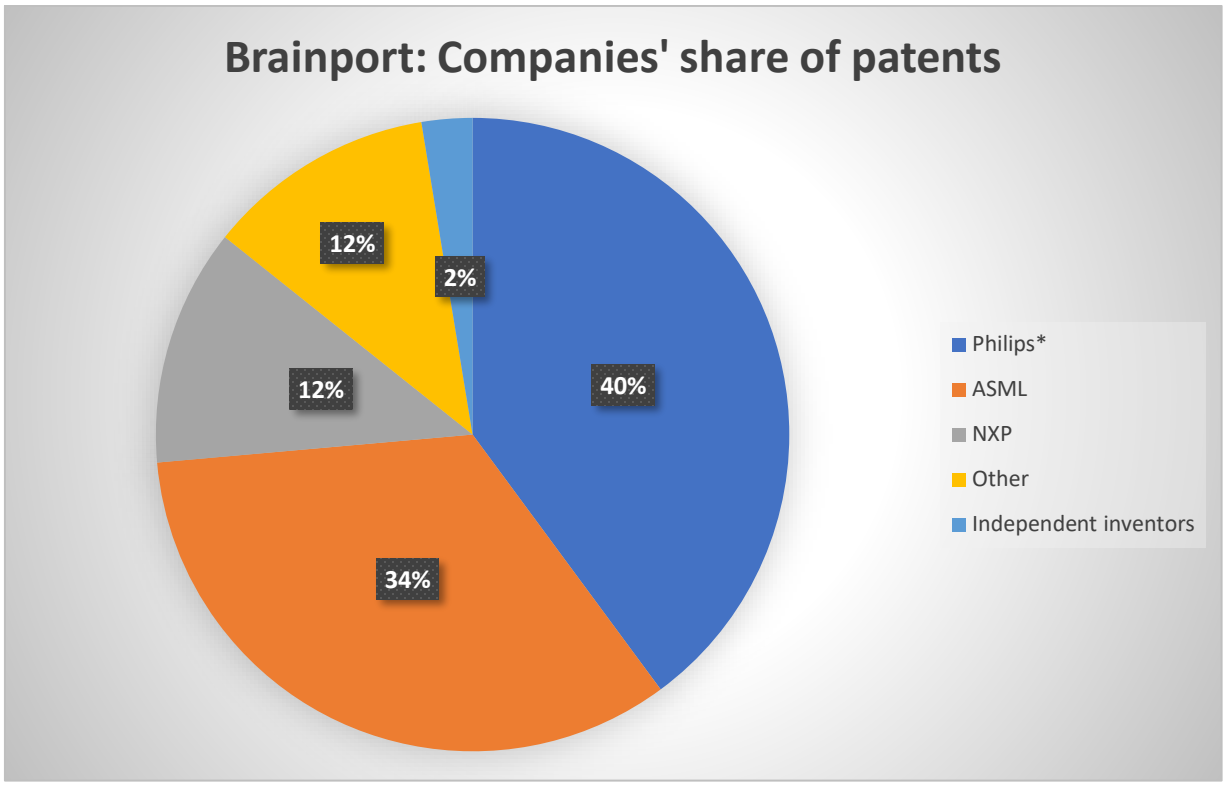


*Figure 22. Distribution of firms in the Brainport sample (N=2821). *(all departments of Philips)*

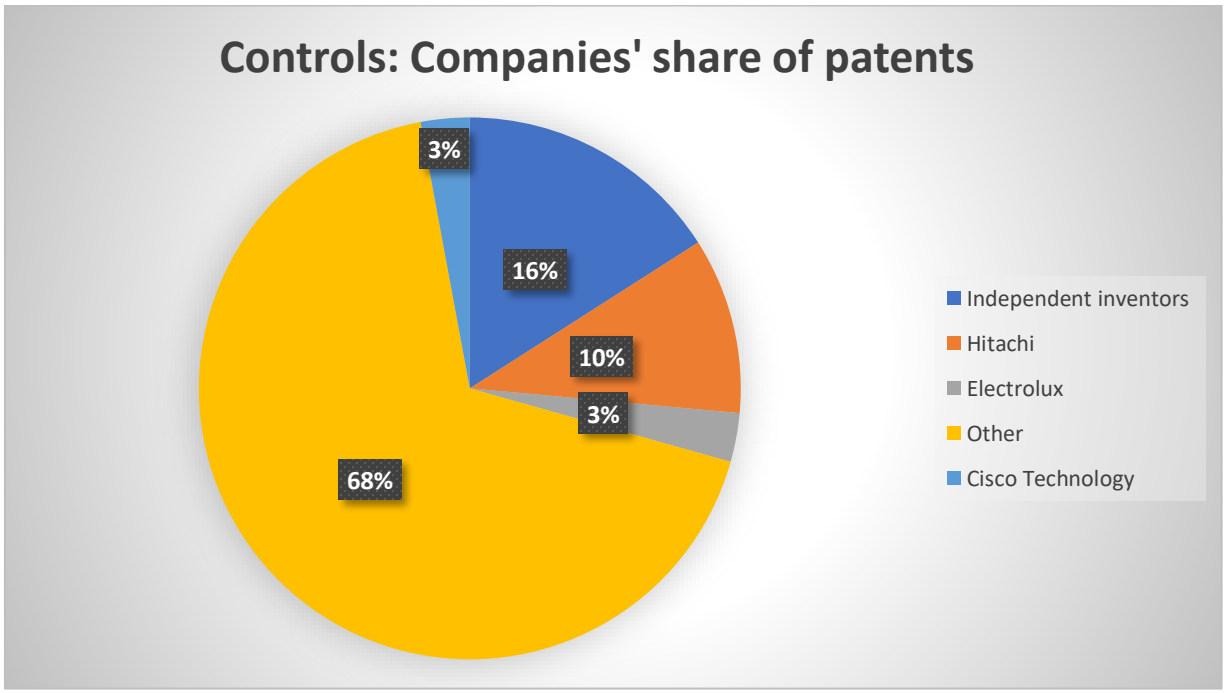


Figure 23. Distribution of firms in the control group of the Brainport patents (N=1818).

4.1.2 Technological Classification

For both ecosystem groups, Tables 4 and 5 represent the distribution of the technological diversity of the patents, indicated by the corresponding IPC codes. For each 4-digit IPC code, a description of the technology class is provided (WIPO, 2023). As can be seen, the majority of the patents originate either from class "G", representing "physics", or technology class "H", representing "electricity". Although most of the patents are allocated to a handful of companies, as described in the previous section, the distribution of the IPC codes is more dispersed. This means that these highly dominating firms do not solely innovate on one specific IPC code. When comparing the distributions of both Tables 4 and 5, code "G03B" is the most frequent code in the Brainport sample, whereas the same code is not in the top 10 most frequent codes in the HTCE sample. The fact that "G03B" is the most frequent IPC code is explained by the presence of ASML in the Brainport sample. It was found that of all the patents allocated to ASML, which are 844 patents, approximately half the patents do have the code "G03B". So, in this case, there is a very strong relationship between the company and the sort of IPC code.

IPC	Frequency	Percentage	Description
G06F	227	14.03	Electrical digital data processing
H01L	163	10.07	Semiconductor devices
H04L	145	8.96	Transmission of digital information
A61B	125	7.73	Diagnosis; surgery; identification
H04N	89	5.50	Pictorial communication
G06T	55	3.40	Image data processing or generation, in general
H04B	49	3.03	Transmission
G11B	45	2.78	Information storage based on relative movement between record carrier and transducer
H04W	43	2.66	Wireless communication networks
G01N	42	2.60	Investigating or analyzing materials by determining their chemical or physical properties

Table 4. Distribution of the IPC codes in the HTCE group.

IPC	Frequency	Percentage	Description
G03B	485	10.45	Apparatus or arrangements for taking photographs or for projecting or viewing them
H01L	466	10.05	Semiconductor devices
G06F	401	8.64	Electrical digital data processing
H04L	246	5.30	Transmission of digital information
G03F	213	4.59	Photomechanical production or textured or patterned surfaces
A61B	202	4.35	Diagnosis; surgery; identification
H04N	175	3.77	Pictorial communication
G02B	144	3.10	Optical elements, systems or apparatus
G11B	125	2.69	Information storage based on relative movement between record carrier and transducer
G01N	118	2.54	Investigating or analyzing materials by determining their chemical or physical properties

Table 5. Distribution of the IPC codes in the Brainport Eindhoven group.

4.1.3 Year of Application

For both ecosystem groups, the distribution of the year of application is shown in the histograms presented in Figures 24 and 25. As can be seen, except for the year 2010, the distribution of the Brainport sample is more equally dispersed. Next to that, for the Brainport Eindhoven group, in the years 2004 and 2005 most of the patents were filed, whereas for the HTCE group, most patents were filed in the years 2007 and 2008. In a way, this is important since it is given in this research that the earlier the filing of the patent, the longer the citation lag will be as the data only includes citing patents that are filed up to the year 2014. So, the earlier the filing of the cited patent, the more likely that the number of forward or citing patents will be higher.

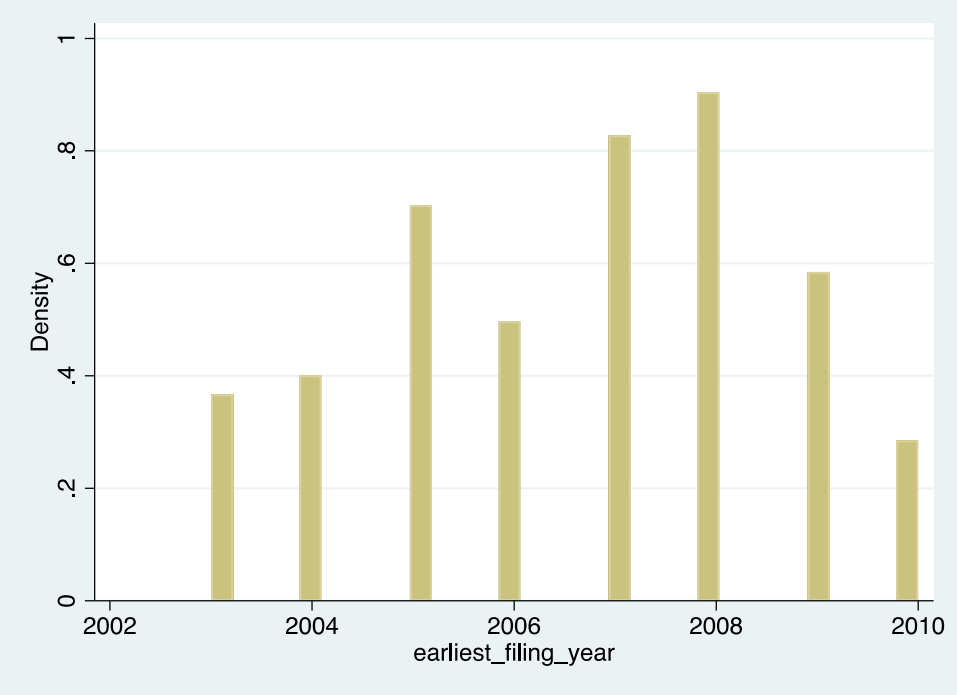


Figure 24. Earliest year of filing, HTCE group.

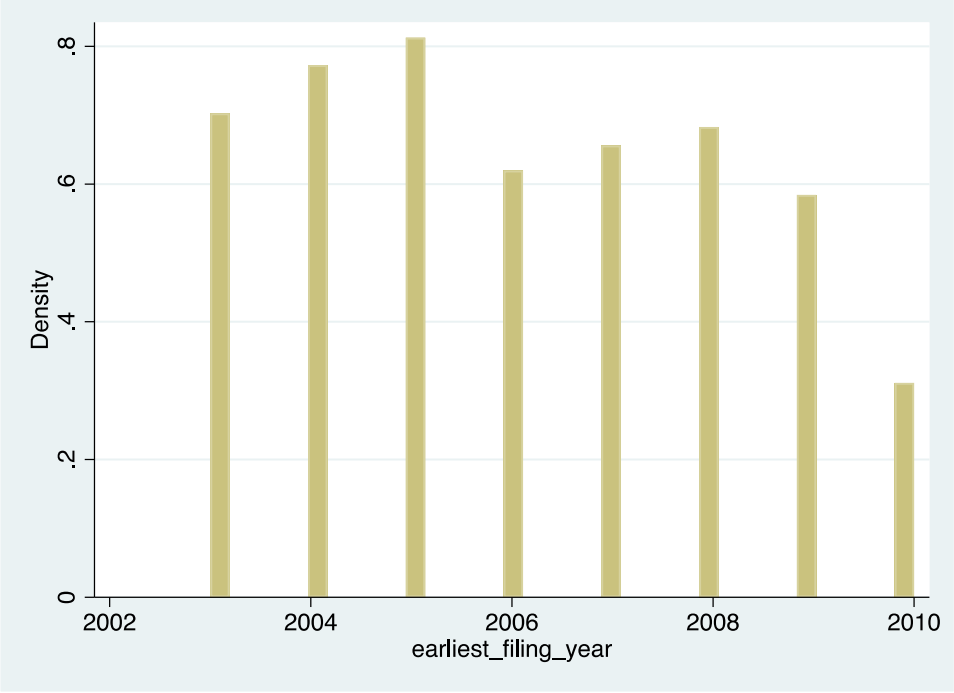


Figure 25. Earliest year of filing, Brainport Eindhoven group.

4.1.4 Geographical Distribution of the Citing Patents

Figures 26 – 31 present the distribution of the citing patents for both the ecosystem and the peer control patents. For each ecosystem -and control group, the distribution of the citing patents is visualized by the hand of three geographical narratives. Figures 26 and 27 display the distribution of the citing patents worldwide, Figures 28 and 29 display the distribution of the citing patents within Europe and Figures 30 and 31 display the distribution of the citing patents within the Benelux. Figures 26 – 31 were made using the software 'Flourish.Studio'.[2] As input for these figures, all the geographical locations of the citing patents that were found throughout the data treatment were used (Appendix A). In these figures, the geographical locations of the citing patents are untreated. This means that for each citing patent all the corresponding inventor locations are incorporated in Figures 26 – 31. Therefore, the red and blue dots do represent inventor locations instead of the locations of unique citing patents. However, to simplify, the inventor locations of the citing patents are referred to as the location of the citing patents. For the HTCE group, 5312 citing patents were found whereas for the peer control group 17269 citing patents were found (Appendix A). For the Brainport Eindhoven group, 46468 citing patents were found whereas for the peer control group 35256 citing patents were found (Appendix A). These differences must be considered when interpreting Figures 26 – 31.

Figures 26 and 27 illustrate that the citing patents are distributed among three major regions: The USA, Europe, and Asia. Especially in the HTCE group, the share of control citing patents seems to be larger in the USA. With respect to Asia, both figures illustrate that the density of the citing patents is most frequent in the eastern part of China and Korea. In the Middle East, citing patents are hardly found, except for Israel, where a significant share of citing patents is found.

Considering Europe, Figures 28 and 29 show that for both ecosystem groups, most of the citing patents do originate from the United Kingdom (UK), the Benelux, and Germany. Figure 29 displays a concentrated proportion of 'red labeled' citing patents around the city of Eindhoven. The same pattern is observed in Figures 30 and 31. The visualizations in Figures 30 – 31 do not show that citation links, based on cited patents from ecosystems, are spatially more localized to their controls. However, Figures 30 and 31 do show that the diffusion of knowledge within -and around the region of Eindhoven is spatially sticky.

[2] https://flourish.studio

Figure 26. Worldwide distribution of the citing patents for both the HTCE -and control patents.

Figure 27. Worldwide distribution of the citing patents for both the Brainport Eindhoven -and control patents.

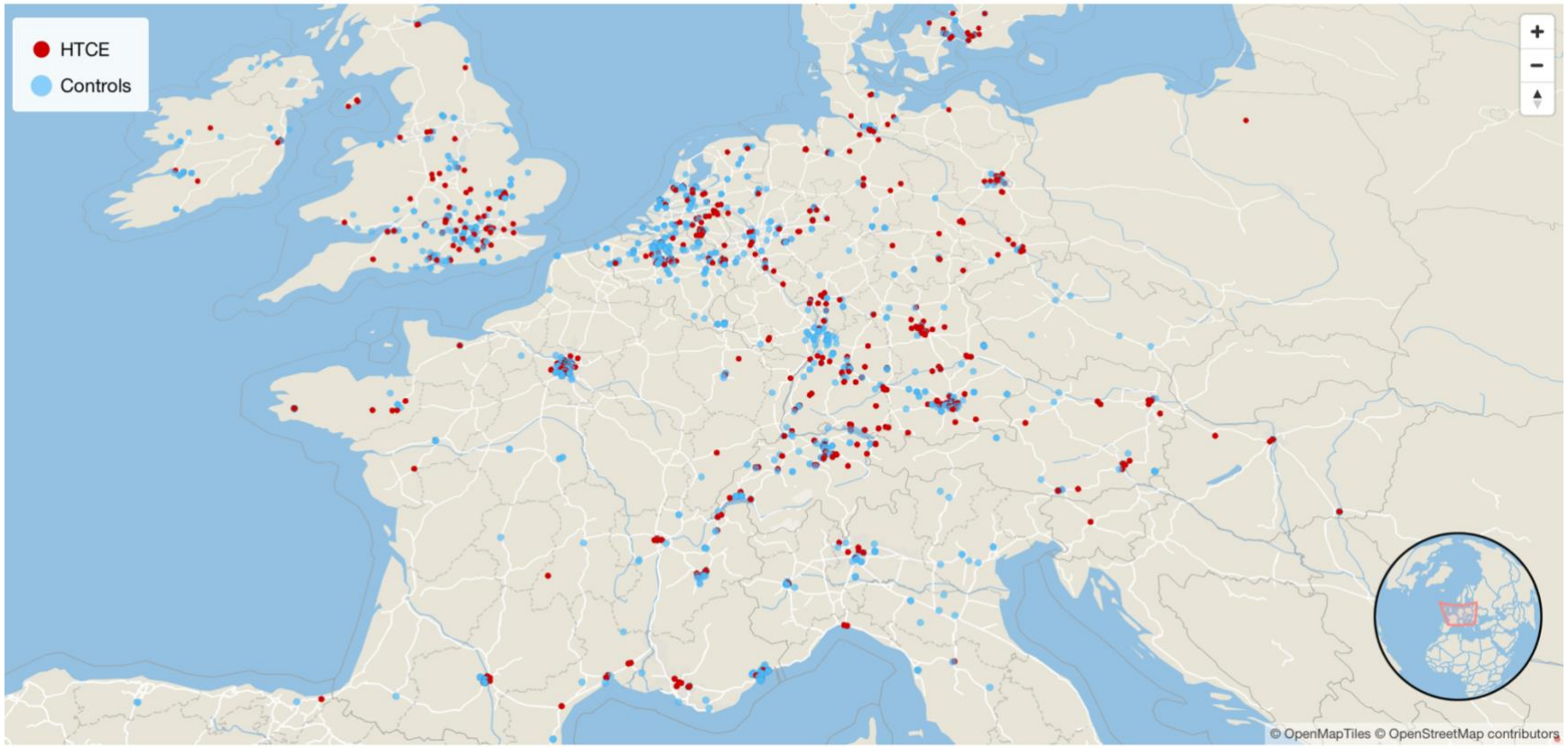


Figure 28. Distribution of the citing patents for both the HTCE -and control patents within Europe.

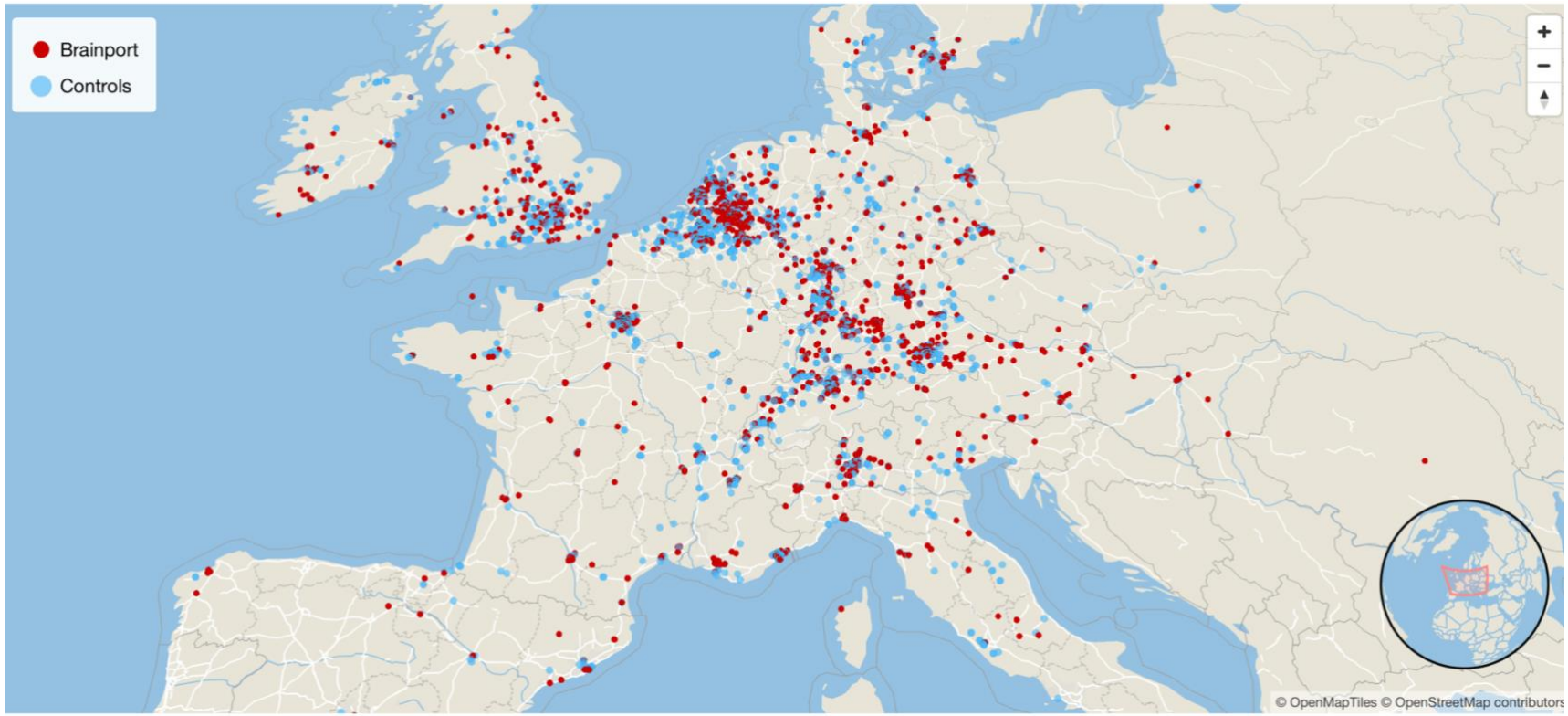


Figure 29. Distribution of the citing patents for both the Brainport Eindhoven -and control patents within Europe.

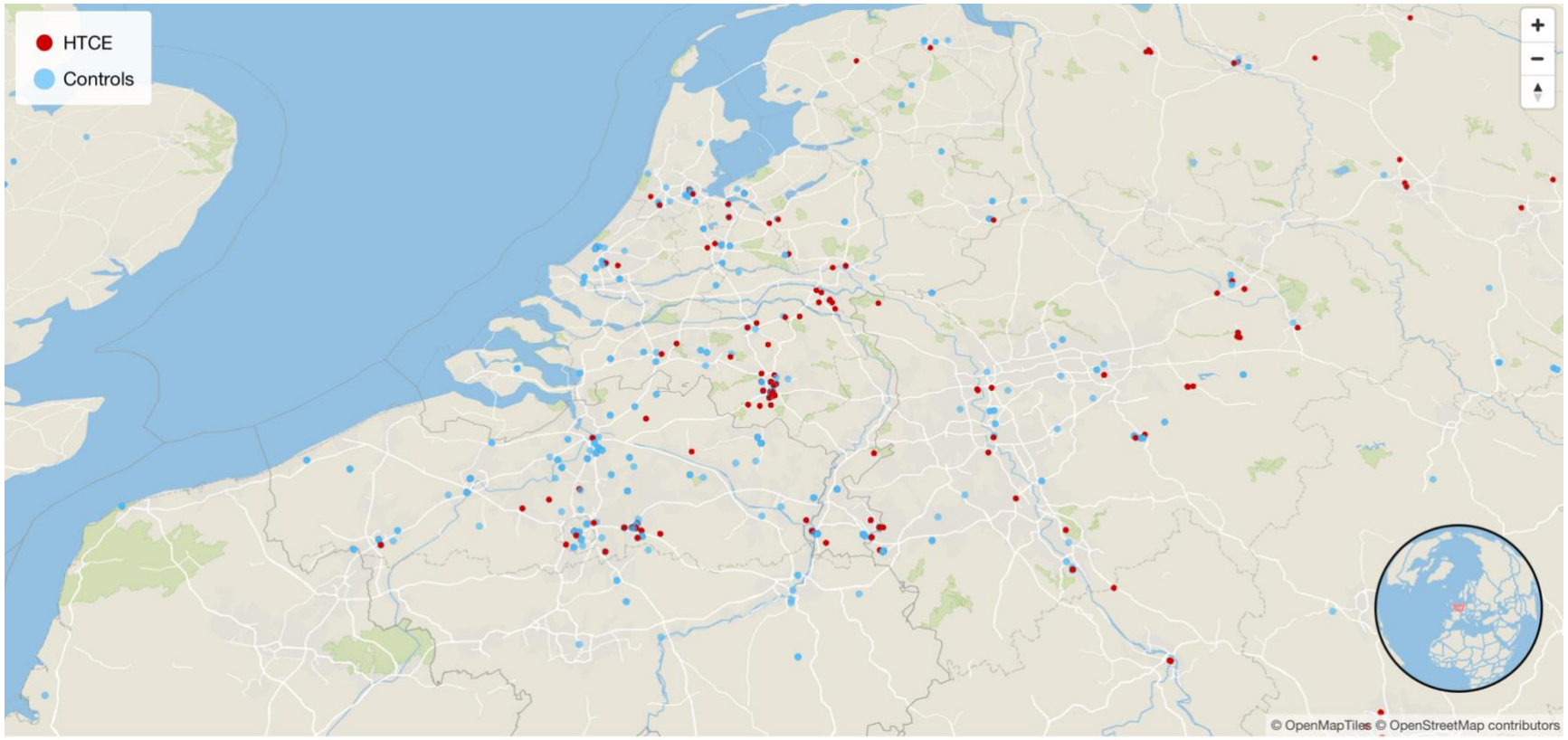


Figure 30. Distribution of the citing patents for both the HTCE -and control patents within the Benelux.

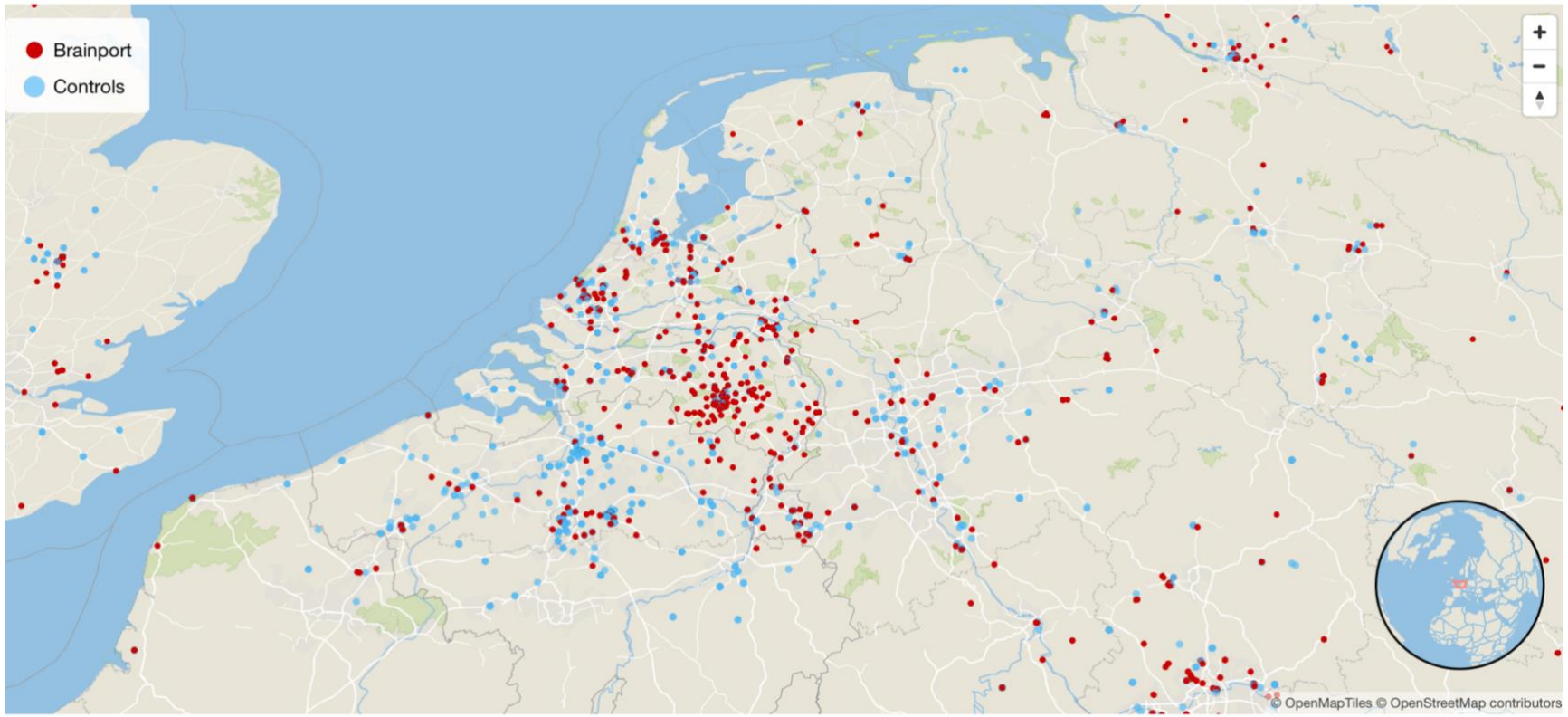


Figure 31. Distribution of the citing patents for both the Brainport Eindhoven -and control patents within the Benelux.

4.1.5 Distance Metrics

In the Methodology in Chapter 3, three different forms of distance metrics were formulated. Tables 6 and 7 display a comparison of the means between the ecosystem -and the control groups with respect to the following distance metrics: 1. The average of all citation links per patent, represented by the variable *average_haversine*. 2. The shortest citation link per patent, represented by the variable *shortest_citation*. 3. Multiple dummy variables, indicating whether a patent has a forward citation within, or outside, a specific range. This measure is represented by the variable *dum_**.

	Treated (N=765)	**Controls (N=853)**	**Mean difference**
mean average_haversine (km)	4770.414	5222.674	452.3** (2.77)
mean shortest_citation (km)	3957.111	3655.684	-301.4 (-1.61)
mean dum_0	.0352941	.1313013	0.0960*** (6.98)
mean dum_30	.0810458	.1746776	0.0936*** (5.63)
mean dum_50	.0849673	.1817116	0.0967*** (5.72)
mean dum_100	.103268	.2004689	0.0972*** (5.45)
mean dum_200	.1267974	.2344666	0.108*** (5.64)
mean dum_500	.3215686	.4150059	0.0934*** (3.90)
mean dum_1000	.4888889	.5334115	0.0445 (1.79)
mean dum_2500	.5346405	.5720985	0.0375 (1.51)
mean dum_5000	.5581699	.5978898	0.0397 (1.62)
mean dum_larger_2500	.6745098	.7854631	0.111*** (5.07)
mean dum_larger_5000	.6535948	.7819461	0.128*** (5.80)

Treated == 1, Control == 0
diff = mean(0) - mean(1)
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 6. Comparison of means for several distance metrics in the HTCE group.

In Stata, for each of the *dist_** variables, a dummy variable was created and named *dum_** (Appendix B). Instead of using the absolute number of citations within a particular range *(dist_*),* the dummy variables indicate whether a patent has a forward citation within a particular range by returning the value "1" in case the patent has one or more citation links within a particular range

and returning the value "0" in case a patent has no citation links within that range. These dummy variables make the models less sensitive to outliers. In the first two columns, for each measure, the means of both the treated -and the control patents are provided. In the third column, the difference between the means is calculated and complemented with a t-statistic.

Firstly, as can be seen, both the average distance of the average citation link and the shortest citation link is relatively large for both the ecosystem -and the control group. To support the interpretation of Tables 6 and 7 visually, Figure 32 illustrates the radius of the average shortest citation link for the HTCE -and control patents. The circle around Eindhoven is not entirely round due to the curvature of the earth. In the figure, the radius of the average shortest citation link extends beyond the borders of Europe. The purpose of Figure 32 is purely to visualize the lengths of the average citation links displayed in Tables 6 and 7. It is not the intention here to present any differences between ecosystem -and control patents.

	Treated (N=2821)	**Controls (N=1818)**	**Mean difference**
mean average_haversine (km)	5227.677	4923.1	-304.6** (-3.18)
mean shortest_citation (km)	3898.492	3392.118	-506.4*** (-4.49)
mean dum_0	.1088267	.160066	0.0512*** (5.10)
mean dum_30	.1751152	.2112211	0.0361** (3.07)
mean dum_50	.1779511	.2156216	0.0377** (3.18)
mean dum_100	.2041829	.2326733	0.0285* (2.31)
mean dum_200	.224743	.2684268	0.0437*** (3.40)
mean dum_500	.3807161	.4532453	0.0725*** (4.92)
mean dum_1000	.5086849	.5720572	0.0634*** (4.23)
mean dum_2500	.5359801	.6028603	0.0669*** (4.49)
mean dum_5000	.5547678	.6226623	0.0679*** (4.59)
mean dum_larger_2500	.7653314	.7623762	-0.00296 (-0.23)
mean dum_larger_5000	.7561149	.7557756	-0.000339 (-0.03)

Treated == 1, Control == 0
diff = mean(0) - mean(1)
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 7. Comparison of means for several distance metrics in the Brainport Eindhoven group.

With respect to the differences between the first two distance metrics, in Table 6, the mean of the shortest citation link is higher for the HTCE patents in comparison to their controls, whereas the mean of the average of all citation links is shorter for the HTCE patents. With respect to the Brainport Eindhoven group, both the means are larger for the ecosystem patents compared to the corresponding control group. Considering the dummy variables, firstly, for both ecosystems, it can be seen that the peer control patents do have a higher share of citation links with zero kilometers which are found to be statistically significant. Secondly, for both ecosystem groups, up to a range of 500 kilometers, the controls are found to be more localized as the means of the control patents are higher and the differences are statically significant. Based on these results, it can be concluded that the control patents are more localized compared to their peer-treated patents. In Tables 6 and 7, the last two rows indicate whether the means of both groups differ with respect to the number of distant citation links. For the HTCE group, it can be seen that the control patents are estimated to have more distant citations, as the difference in the means is statistically supported (Table 6).

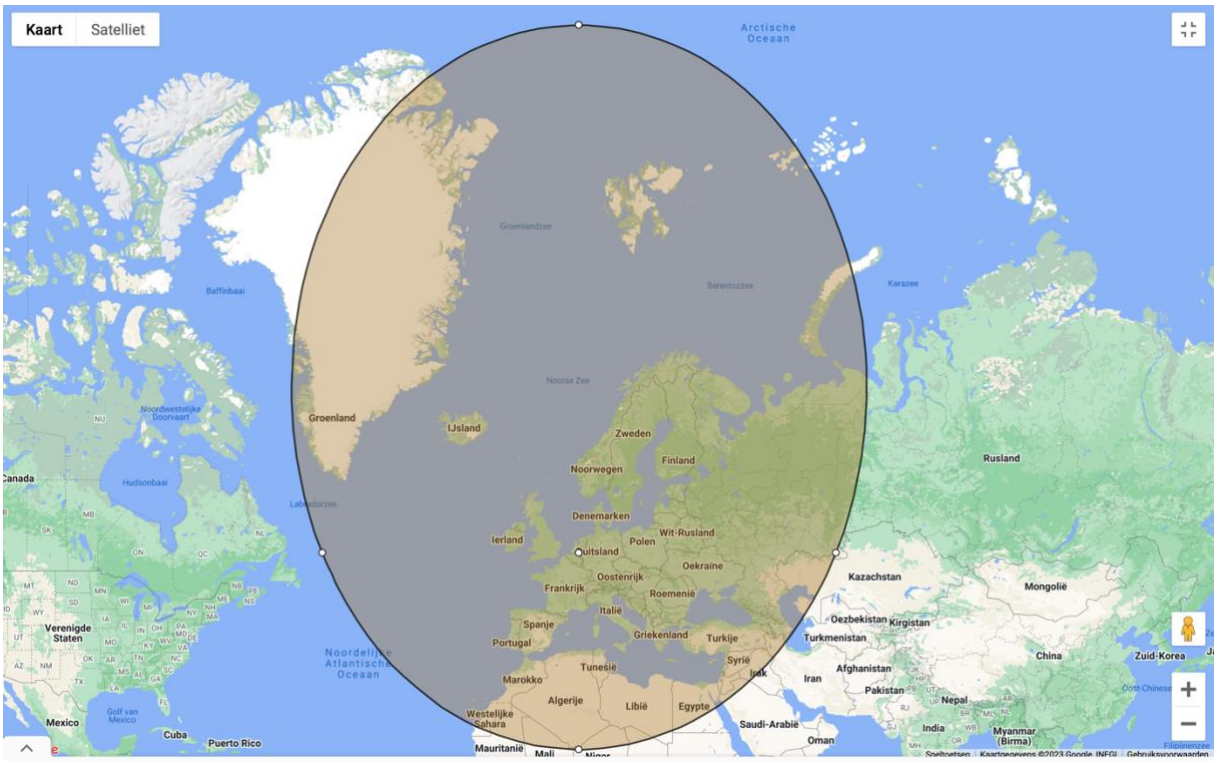


Figure 32. Radius of the average shortest citation link for both Brainport -and control patents (r = 3800 km).

4.1.6 Distribution of the Density of Citation Links Over Distance

In the previous section, the effects of localization were studied by comparing the means of different distance metrics. In addition to the comparison of the means, the relationship between ecosystems and the distance of citation links can be visualized in a graph. Figures 33 – 38 display the distribution of citation links of both ecosystem groups. To be more specific, the density of patents is plotted along the distance of the shortest citation link. Appendix B reports the commands that were used in Stata to create Figures 33-38. Appendix B reports the transcript of all the lines of code that were used in Stata during the statistical analysis. Although no control variables are used for the plotting of the graphs, the graphs do include analytical weights. For each ecosystem group, three different graphs were plotted.

Figures 33 and 34 present the distribution of patents over the distance of the shortest citation link. As can be seen here, both graphs are skewed to the right substantially due to the high frequency of citation links with a distance of zero kilometers. Concerning the HTCE group, it can be seen that the control patents do have a larger share of these citation links compared to their peer HTCE patents. Since the first of group of graphs are highly skewed, in the 2nd and 3rd group of the plots, citation links with a distance of zero kilometers were excluded. The 2nd group (Figures 35 and 36) displays the distribution of the shortest citation link within a range of 200 kilometers, whereas the 3rd group (Figures 37 and 38) displays the distribution of the shortest citation link within a range of 2500 kilometers. As can be seen, with the exclusion of citation links with zero kilometers, the distributions display new information.

For both the HTCE -and the Brainport Eindhoven group, a similar pattern can be observed. For both ecosystem groups, within a range of 50 kilometers, the density of the ecosystem patents is larger in comparison to the controls, as displayed in Figures 35 and 36. These observations are in line with the assumption and the hypothesis that the share of ecosystem patents within a range of 50 kilometers is larger compared to the control groups. Besides a higher density of ecosystem patents within the short range of 50 kilometers, for both ecosystem groups, the density of ecosystem patents is larger compared to the controls at the distance ranges of approximately 400-750 kilometers (Figures 37 and 38). In the range of approximately 50-500 kilometers, for both ecosystem groups, the density of the control patents is larger compared to the ecosystem patents. In addition, Figures 37 and 38 indicate that for all groups the share of patents peaks at a range of approximately 400 kilometers. This is probably due to the large share of citing patents that originate from firms in the UK, France, and Germany at the perimeter of 400 kilometers around Eindhoven (Figures 28, 29, and 39). Figure 39 shows that within a radius of 400 kilometers around Eindhoven, major European cities like London, Paris, and Hamburg can be found.

All in all, within a range of 1000 kilometers, the effects of localization fluctuate. The visualizations and results displayed in Figures 33 – 38 are purely indicational and do not have any statistical significance. In the next paragraph, it will be examined whether the indicated effects of localization are statistically significant.

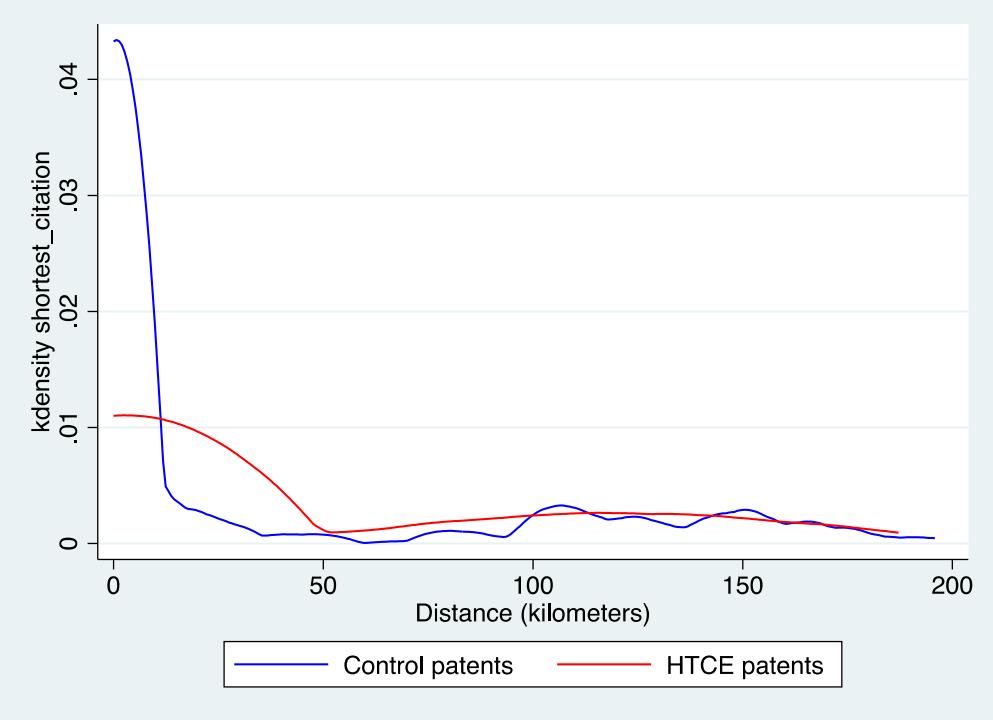


Figure 33. Density distribution of the shortest citation link (range < 200 kilometers) of the HTCE group.

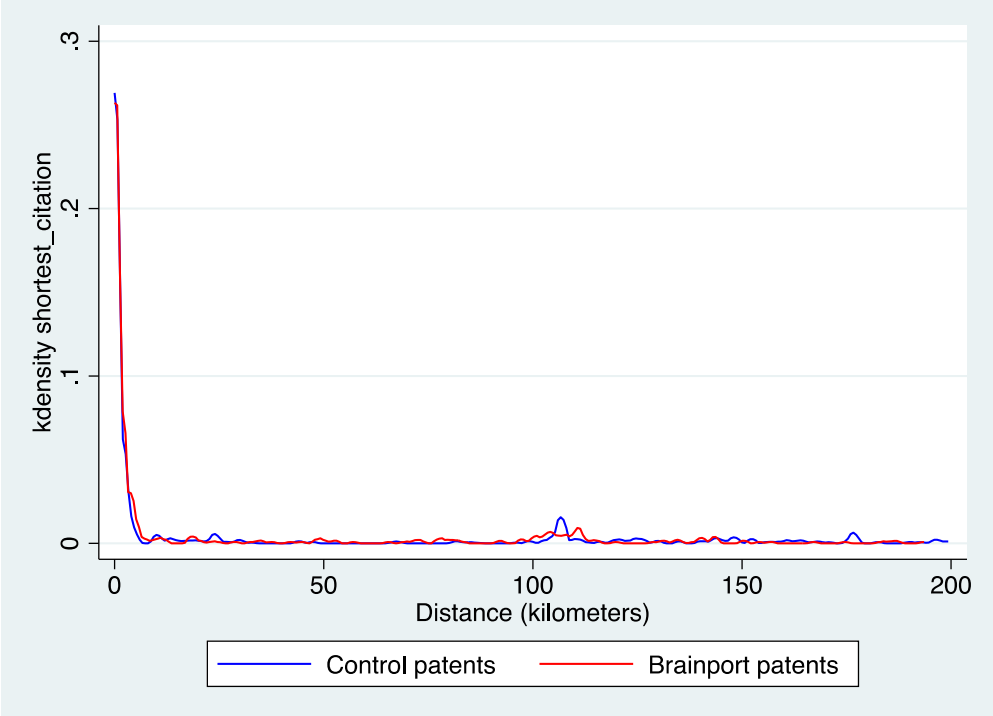
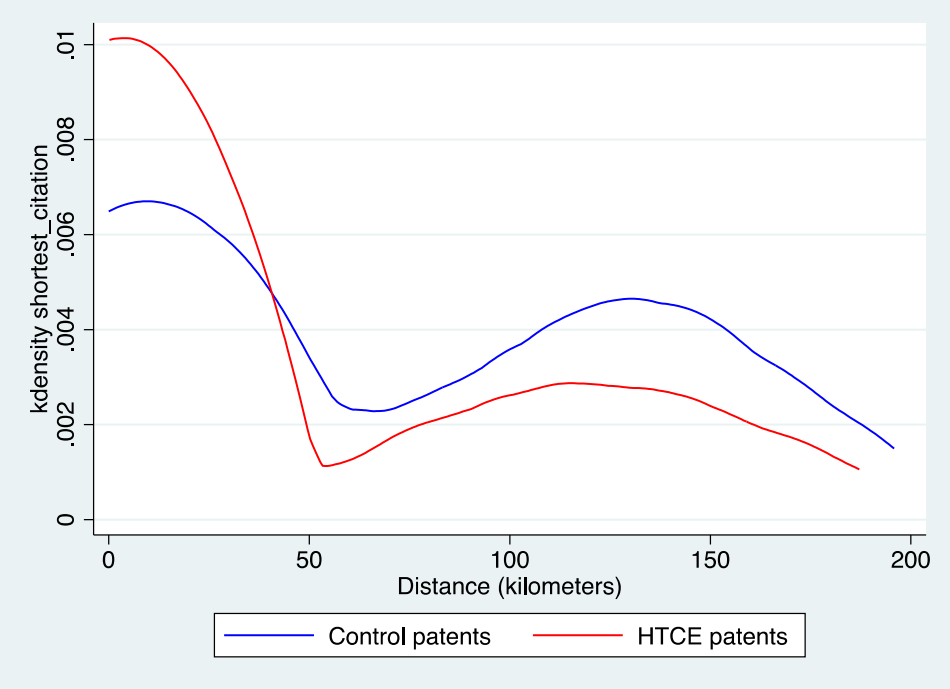
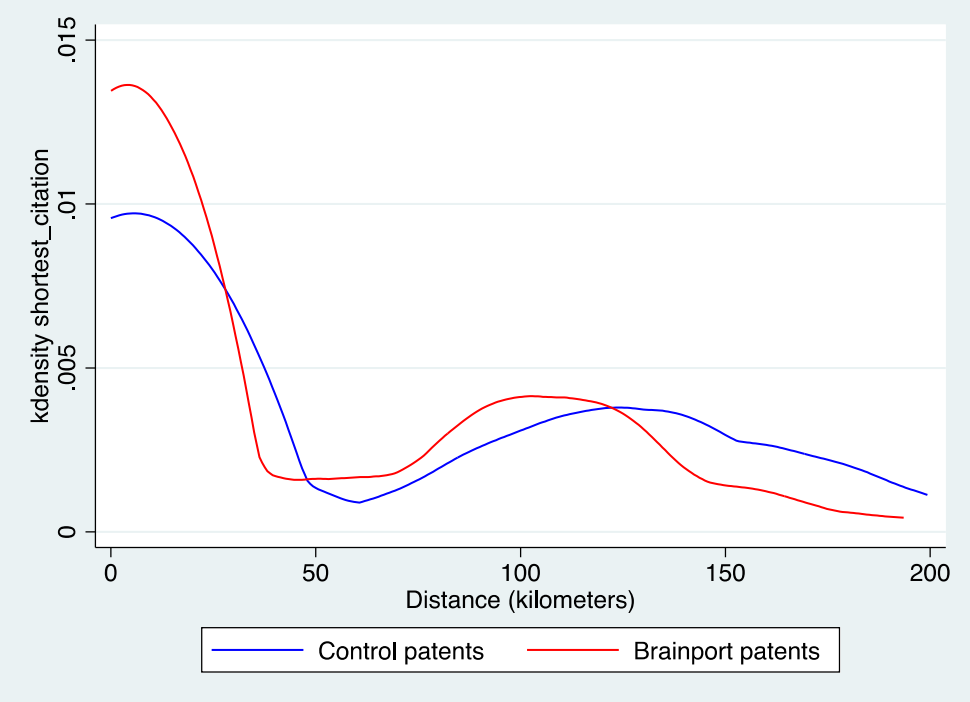


Figure 34. Density distribution of the shortest citation link (range < 200 kilometers) of the Brainport Eindhoven group.



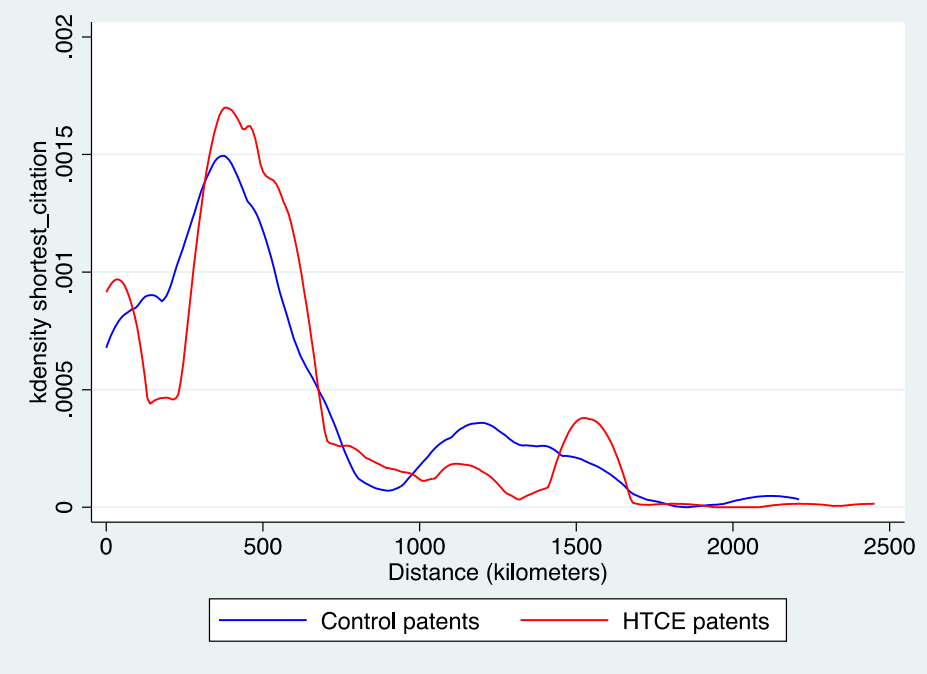
Condition: "shortest_citation != 0"

Figure 35. Density distribution of the shortest citation link (range < 200 kilometers) of the HTCE group.



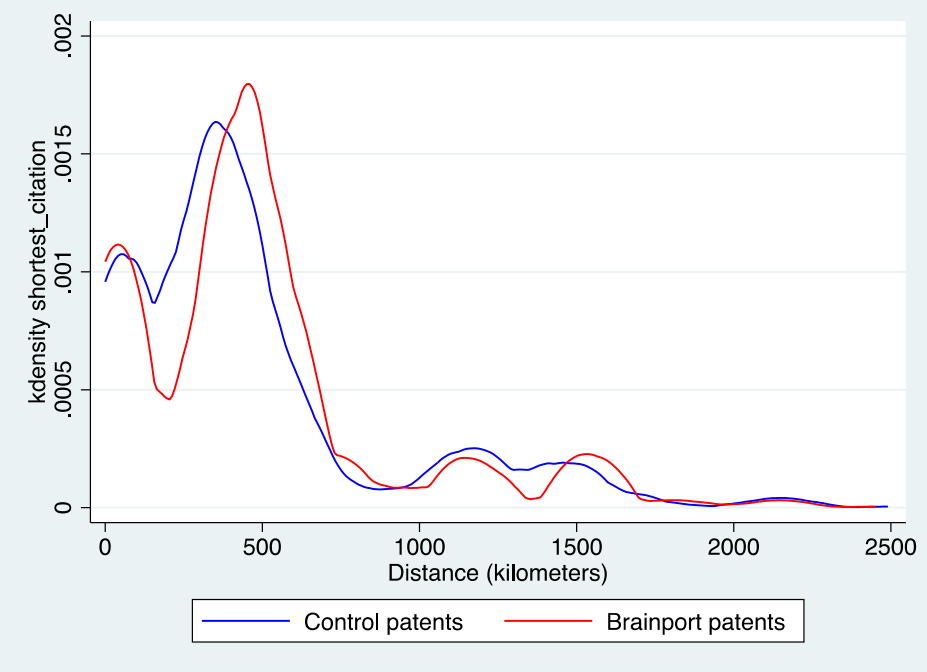
Condition: "shortest_citation != 0"

Figure 36. Density distribution of the shortest citation link (range < 200 kilometers) of the Brainport Eindhoven group.



Condition: "shortest_citation != 0"

Figure 37. Density distribution of the shortest citation link (range < 2500 kilometers) of the HTCE group.



Condition: "shortest_citation != 0"

Figure 38. Density distribution of the shortest citation link (range < 2500 kilometers) of the Brainport Eindhoven group.

Figure 39. Radius of 400 kilometers around the city of Eindhoven.

4.2 Regression Results

The descriptive statistics in section 4.1.5 do show rather counterintuitive results. The comparison of the means between the treated -and control group indicate that the citation links, originating from cited control patents, are spatially more localized. However, as mentioned in the Methodology in Chapter 3, with the use of t-statistic, the relationship between ecosystems and the distance of citation links was not studied while controlling for other predictor variables. Hence, it was chosen to use OLS regression.

With respect to the citation links with a distance of zero kilometers, based on the issues described in section 4.1.6, it was chosen to exclude citation links with zero kilometers in the regression models. As can be derived from Figures 33 and 34, the dominant share of citation links with zero kilometers overshadows potential signs of localization elsewhere. It might be the case that these citation links, which were found to have a significantly higher share in the control group (Tables 6 – 7), mediate the localization effects in support of the control patents. Besides that, the share of these citation links was very high (Figures 33 and 34), and the nature of these citation links with zero kilometers is somewhat peculiar. Namely, since they represent self-citations in some kind of form, either a self-citation from an inventor or an organization. Although self-citations from organizations were excluded from the data, it was found that some self-citations by organizations were still present. For example, the sub-departments of some firms were each provided with a unique company name in PATSTAT. For example, a citation link was found from 'Philips N.V.' to 'Philips Intellectual Property GmbH'. Although this citation link within one overarching organization is a self-citation, is not considered a self-citation in the data treatment, and thereby remains present in the data.

For the remainder of this chapter, the results of the linear regression models are structured as follows. Tables 9 and 10 display the outcomes of the linear regression models in anticipation of testing the hypotheses. Table 8 shows for each hypothesis what dependent variable is measured. Table 9 presents the outcomes of the regression models of the HTCE group, whereas Table 10 presents the Brainport Eindhoven group. In addition to Tables 9 and 10, Tables 11 – 13 can be found. The reason to include Tables 11 – 13 can be considered twofold. First, Table 11 assesses the sensitivity and vulnerability of the regression models that were used to create Tables 9 and 10. Secondly, besides the main aim of this research, Tables 12 and 13 present additional findings. Appendix B reports the transcript of the lines of code that were used in Stata to run the regressions corresponding to Tables 9 – 13. In the caption of Tables 9 – 13, additional information on the regression tables is explained. Here, a distinguishment is made between control variables and control groups. The control variables are listed in Table 2. The control groups are described in Chapter 3.4.1.

Measure	Hypothesis	Focus
avg_haversine	Hypothesis 1	Diffusion of ecosystem knowledge in general.
shortest_citation	Hypothesis 1	Diffusion of ecosystem knowledge in general with a focus on the most local citation link.
dum_50	Hypothesis 2	Diffusion of knowledge within ecosystems.
dum_100, dum_200, dum_500	Hypothesis 3	Diffusion of knowledge within -and around ecosystems.

Table 8. Measured dependent variable for each hypothesis.

4.2.1 Main Results

In both Tables 9 and 10, each column represents a separate regression. For each regression, the same control variables and settings were used, only the dependent variables differ (Appendix B, Tables 9 and 10). In each column, the dependent variable, representing a measurement of distance, is displayed. In the left column, the independent variable *treated* is displayed. This variable is a dummy variable; taking the value "1" in case a patent originates from Brainport Eindhoven or the HTCE and taking the value "0" in case a patent is a control patent. The statistics of the other control -and predictor variables that were used in the regression are not displayed in Tables 9 – 13.

In column 2, for both the HTCE group -and the Brainport Eindhoven group, the distance of the shortest citation link is estimated to be larger for the control patents. To illustrate, for the Brainport Eindhoven group, it is estimated that the distance of the shortest citation link of a control patent is approximately 760 kilometers longer compared to an ecosystem patent (Table 10). This implies that the shortest citation link of patents, originating from cited ecosystem patents, is estimated to be shorter compared to the control patents. For both the HTCE -and the Brainport Eindhoven group, this result is significant. With respect to the average distance of all citation links, represented in column 1, a similar result is found, although only significant in the HTCE dataset (Table 9).

In columns 3 – 6 (Tables 9 and 10), the results are shown with respect to the dummy variables of citation links within a specific range. In both Tables 9 and 10, within a range of 50 kilometers, indicated by *dum_50* in column 3, it is found that control patents are estimated to have fewer citation links within a range of 50 kilometers compared to the ecosystem patents. However, this result is not statistically supported.

	(1)	(2)	(3)	(4)	(5)	(6)
	avg_haversine	*shortest_citation*	*dum_50*	*dum_100*	*dum_200*	*dum_500*
treated	-3103.3***	-4308.2***	0.0441	0.208**	0.187*	0.424***
	(-3.73)	(-4.51)	(0.70)	(2.99)	(2.35)	(3.56)
_cons	1577.6	1965.7	0.278*	0.431***	0.417**	0.502*
	(1.04)	(1.13)	(2.41)	(3.38)	(2.86)	(2.31)
N	1504	1504	1504	1504	1504	1504
R^2	0.391	0.406	0.435	0.448	0.441	0.386

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Control variables: *nb_citations, nb_inventors, nb_foreign_inventors, nb_ipc_codes, nb_backward_citations, nb_forward_citations, publn_claims, treated*.

Control groups: fixed firm effect, fixed year effect, fixed IPC effect (Chapter 3.4.1, Appendix B).

Table 9. OLS regression of multiple distance measures for the HTCE group.

On the other hand, with respect to the distances of 100 kilometers, 200 kilometers, and 500 kilometers, for both the HTCE and Brainport Eindhoven group, a similar -and significant result is found (columns 4 – 5). Both interesting and contrasting, the results displayed in Tables 8 and 9 are opposite to the patterns which can be found in the graphs above in Figures 35 and 36. Here, within a radius of 50 kilometers, the density of ecosystem patents was found to be higher compared to the control patents, whereas in Tables 9 and 10 no statistical support can be found for this observation. There is only a statistically significant effect for the distance ranges above 50 kilometers where the direction of the coefficients is opposite to the patterns in Figures 35 and 36 in which the density of the control patents increases over the density of the ecosystem patents after the range of 50 kilometers.

	(1)	(2)	(3)	(4)	(5)	(6)
	avg_haversine	*shortest_citation*	*dum_50*	*dum_100*	*dum_200*	*dum_500*
treated	-446.2	-761.1*	0.00452	0.0717*	0.0797*	0.149***
	(-1.53)	(-2.18)	(0.17)	(2.44)	(2.47)	(3.32)
_cons	4257.7	3984.5	0.183	0.167	0.146	-0.115
	(0.92)	(0.72)	(0.44)	(0.36)	(0.28)	(-0.16)
N	4084	4084	4084	4084	4084	4084
R^2	0.394	0.349	0.380	0.368	0.374	0.352

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Control variables: *nb_citations, nb_inventors, nb_foreign_inventors, nb_ipc_codes, nb_backward_citations, nb_forward_citations, publn_claims, treated*.

Control groups: fixed firm effect, fixed year effect, fixed IPC effect (Chapter 3.4.1, Appendix B).

Table 10. OLS regression of multiple distance measures for the Brainport Eindhoven group.

4.2.2 Sensitivity and Additional Findings

In this section, the aim is to address the sensitivity of the models that were used for the main results (Table 11). Next to that, additional findings will be presented in Tables 12 and 13. Table 11 is divided into two parts. The left side of the table displays the results for the HTCE group whereas the right side of the table displays the results for the Brainport Eindhoven group. These tables are, to a certain extent, not self-explanatory. Therefore, the results in each column are carefully elaborated.

4.2.2.1 Sensitivity

With respect to the models that were used for the main results (Tables 9 – 10), it is found that the models were very sensitive to the following three circumstances: the fixed firm effect (i), the in -or exclusion of citation links with a distance of zero kilometers (ii), the in -or exclusion of patents with only one citation link (iii).

i. The fixed firm effect.

As mentioned in the Methodology, in the regression models, it was controlled for the fixed firm effect. Here it was checked whether the relationship between a patent and the distance of the citation links is mediated by a particular firm. In columns 1 of Table 11, the results of the exact same regression as the regression used for the results in columns 4 of Tables 9 and 10 are displayed, except for the control on firms, which is excluded. As can be seen, for both the HTCE and Brainport Eindhoven groups, the results of the variable *dum_100* are no longer significant. This indicates that, the significance of the model used for the main results is very sensitive to either an in -or exclusion of the fixed firm effect.

ii. The in -or exclusion of citation links with a distance of zero kilometers.

Earlier in this chapter, it was elaborated and reasoned why patents with a citation link of zero kilometers were excluded from the data. All in all, the in -or exclusion of citation links with a distance of zero kilometers had a huge impact on the significance of the models that were used for the main results. In columns 2 of Table 11, the results of the exact same regression as the regression used for the results in column 5 of Tables 9 and 10 are displayed, except that citation links with a distance of zero kilometers are included. As can be seen, for both the HTCE group, the results of the variable *dum_200* are no longer significant, whereas, for the Brainport Eindhoven group, the results do remain statistically significant. With respect to the HTCE group, it indicates that the significance of the model used for the main results is sensitive to either an in -or exclusion of citation links with a distance of zero kilometers.

iii. The in -or exclusion of patents with only one citation link.

At last, it was tested whether the regression models were sensitive to the in -or exclusion of patents with only one citation link. In the data, it was found that, for both the HTCE -and the Brainport Eindhoven group, the majority of the patents do only have one citation link. This group of patents accounts for approximately 41 percent of the HTCE group and approximately 35 percent in the Brainport Eindhoven sample. The issue is that for this group of patents, it is more likely that the value of the distance of the citation link is an outlier. Ideally, for each cited ecosystem -or control patent, it is desirable to have data about multiple citation links. In that way, the nature of the citations and the corresponding distance can be better understood. Next to that, if each cited ecosystem -and control in the dataset would have multiple citation links, the distance of measures of the 'shortest citation link' and the 'average of all citation links' would be more powerful. For example, in case a patent has only one citation link with a corresponding distance of 5000 kilometers, consequently, the distance of the shortest citation link and the average distance of all citation links will both be 5000 kilometers as well. So, the higher the number of citation links, the better the quality of a patent in the context of research.

In columns 3 of Table 11, the results of the exact same regression as the regression used for the results in column 5 of Tables 9 and 10 are displayed, except that patents with only one citation link are excluded. As can be seen, for the Brainport Eindhoven group, the result of the variable *dum_200* is no longer significant (Table 10), whereas for the HTCE group, the results do remain significant (Table 9). This indicates that the significance of the model used for the main results is sensitive to either an in -or exclusion of patents with only one citation link.

	(1)	(2)	(3)
	dum_100	*dum_200*	*dum_200*
treated	0.00799	0.158	0.287*
	(0.45)	(1.82)	(2.52)
applicant			
_cons	0.306**	0.467**	0.372
	(2.93)	(2.84)	(1.94)
N	1504	1618	860
R^2	0.117	0.455	0.523

	(1)	(2)	(3)
	dum_100	*dum_200*	*dum_200*
treated	0.00918	0.0725*	0.0751
	(0.94)	(1.99)	(1.38)
applicant			
_cons	0.595*	0.151	1.083**
	(2.47)	(0.25)	(2.89)
N	4084	4639	2495
R^2	0.091	0.399	0.419

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Control variables: *nb_citations, nb_inventors, nb_foreign_inventors, nb_ipc_codes, nb_backward_citations, nb_forward_citations, publn_claims, treated*.
Control groups: fixed firm effect, fixed year effect, fixed IPC effect (Chapter 3.4.1, Appendix B).

Exceptions: Column (1) does not include fixed firm effect. Column (2) does include citation links with a distance of zero kilometers. Exceptions: Column (3) does not include patents with only one citation link.

Table 11. OLS regression: Sensitivity results for the HTCE group (left) and Brainport Eindhoven group (right).

4.2.2.2 Additional Findings

Next to the analysis of the sensitivity of the regression models, this section will elaborate on the additional findings that were found. These results are displayed in Tables 12 and 13. Down below, the additional findings will be elaborated one by one.

i. The opposite effect: Patents with a citation link that is larger than 2500 kilometers.

In this research, the main aim is to examine whether citation links that originate from cited ecosystem patents are more localized compared to their control patents. Besides the study of localization, the effect opposite to localization can be studied as well. To be more specific, it can be studied whether the control patents do have more distant citation links compared to the ecosystems. In this case, as shown in columns 1 and 2, it is studied whether control patents do have more citation links within the group of citation links having a distance larger than 2500 kilometers. As can be seen, this assumption is statistically significant in the HTCE group, however, only when the group of patents with only one citation link is included in the regression (column 2).

ii. Patents with two or more firms as an applicant.

As can be seen in column 3, it was examined whether ecosystem patents are more likely to have two or more firms registered as the applicants on a patent. Since ecosystems are characterized by cooperation among different industry sectors and organizations, it was reasoned and assumed that ecosystems patents are more likely to have collaborations among firms with respect to innovation. As shown in Tables 12 and 13, only for the Brainport Eindhoven group a significant result was found (Table 13). In addition, the relationship between these firms, having two or more firms as applicants, and the distance of citation links was examined. However, as can be seen in column 4, no results were found that are statistically supported.

iii. The number of inventors.

Back in the literature review, the paper by Breschi & Lissoni (2009) was elaborated. Breschi & Lissoni had shown that the localization of patent citations was mediated through the mobility of inventors. Concerning ecosystems, it can be interesting to study whether the mobility of inventors is more present within -and around ecosystems compared to non-ecosystems. Although the mobility of inventors is not studied in this research, it is examined whether patents that originate from ecosystems do have more inventors compared to the control patents. As can be seen in column 5, this result is only statistically significant for the Brainport sample (Table 13). Besides the number of all inventors, it was examined whether the control patents are more likely to have a higher number of foreign inventors on a patent application compared to the ecosystem patents. However, columns 6 show that no statistical support was found.

	(1)	(2)	(3)	(4)	(5)	(6)
	dum_larger_2500	*dum_larger_2500*	*applicant*	*shortest_citation*	*nb_inventors*	*nb_foreign_inventors*
treated	-0.0386	-0.342**	0.0227	-4306.1***	-0.540	-0.183
	(-0.34)	(-3.13)	(0.41)	(-4.51)	(-1.27)	(-0.61)
applicant				-91.65		
				(-0.17)		
_cons	0.000212	0.145	0.000790	1965.8	1.389	-0.00379
	(0.00)	(0.73)	(0.01)	(1.13)	(1.79)	(-0.01)
N	860	1504	1504	1504	1504	1504
R^2	0.480	0.424	0.701	0.406	0.401	0.651

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Control variables: *nb_citations, nb_inventors, nb_foreign_inventors, nb_ipc_codes, nb_backward_citations, nb_forward_citations, publn_claims, treated*.

Control groups: fixed firm effect, fixed year effect, fixed IPC effect (Chapter 3.4.1, Appendix B).

Exceptions: Column (1) does not include patents with only one citation link. Column (4) includes predictor variable *applicant*. Columns (5) and (6) exclude predictor variables *nb_inventors* and *nb_foreign_inventors*.

Table 12. OLS regression: Additional research for the HTCE group.

	(1)	(2)	(3)	(4)	(5)	(6)
	dum_larger_2500	*dum_larger_2500*	*applicant*	*shortest_citation*	*nb_inventors*	*nb_foreign_inventors*
treated	0.0152	-0.0717	0.0547**	-755.2*	0.805***	-0.0560
	(0.34)	(-1.88)	(2.76)	(-2.16)	(4.16)	(-0.43)
applicant				-107.7		
				(-0.34)		
_cons	-0.0177	0.339	-0.173	3965.8	6.730*	-1.186
	(-0.06)	(0.56)	(-0.55)	(0.71)	(2.18)	(-0.57)
N	2495	4084	4084	4084	4084	4084
R^2	0.441	0.426	0.668	0.349	0.339	0.567

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Control variables: *nb_citations, nb_inventors, nb_foreign_inventors, nb_ipc_codes, nb_backward_citations, nb_forward_citations, publn_claims, treated*.

Control groups: fixed firm effect, fixed year effect, fixed IPC effect (Chapter 3.4.1, Appendix B).

Exceptions: Column (1) does not include patents with only one citation link. Column (4) includes predictor variable *applicant*. Columns (5) and (6) exclude predictor variables *nb_inventors* and *nb_foreign_inventors*.

Table 13. OLS regression: Additional research for the Brainport Eindhoven group.

5. Conclusion

In this research, it is studied whether the diffusion of knowledge is spatially stickier within -and around ecosystems. Since ecosystems do have proximity-related advantages, which contribute to the increased diffusion and creation of knowledge, it was hypothesized that the diffusion of knowledge is spatially stickier within -and around these regions. Down below, a repetition of the research questions and corresponding hypotheses can be found.

Main research question*:* Is the diffusion of knowledge within -and around ecosystems spatially stickier compared to non-ecosystems?

Sub-question: Are patent citations, originating from cited High Tech Campus Eindhoven (HTCE) -and Brainport Eindhoven patents, more localized compared to patent citations originating from cited 'non-ecosystem' patents?

Hypothesis 1: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized compared to patent citations originating from cited 'non-ecosystem' patents.

Hypothesis 2: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized within the ecosystem compared to patent citations originating from cited 'non-ecosystem' patents.

Hypothesis 3: Patent citations originating from cited HTCE -and Brainport Eindhoven patents are spatially more localized within -and around the ecosystem compared to patent citations originating from cited 'non-ecosystem' patents.

With the use of the dataset by de Rassenfosse et al. (2019), patent data about the HTCE and Brainport Eindhoven was collected for the years 2003 – 2014. Next to the ecosystem patents, a set of control patents was created by matching patents elsewhere from the Netherlands and Belgium to the ecosystem patents on a similarity in IPC and year of filing. With the use of OLS regression, the distances of the citation links originating from cited ecosystem -and control patents were compared.

The first hypothesis was formulated to examine the more general relationship between patent citations and ecosystems. The first hypothesis was tested using the distance metrics of the shortest citation link and the average of all citation links (Table I). In both ecosystem groups, it was found and statistically supported that the distance of the shortest citation link of ecosystem patents is estimated to be shorter compared to their control patents. Concerning the metric of the average

distance of all citation links, the hypothesis was only statistically supported for the HTCE group. Therefore, H1 is partially accepted.

The second hypothesis was formulated to test whether the diffusion of patent citations, diffusing within the ecosystem, is more localized compared to those from non-ecosystems. The second hypothesis was tested by creating a dummy variable that indicated whether a patent had a citation link within the range of 50 kilometers (Table 8). Although the density plots indicated a higher density of ecosystem systems within a range of 50 kilometers (Figures 35 and 36), no statistical support was found in favor of hypothesis 2. Therefore, H2 is rejected.

The third hypothesis was formulated to test whether the diffusion of patent citations, originating from cited ecosystem patents, is more localized within -and around ecosystems compared to the control patents. To test H3, the dependent variables *dum_100, dum_200, and dum_500* were used (Table 8). It was found that, for both ecosystem groups, patent citations originating from cited ecosystem patents were estimated to have more citation links within the range of 100, 200, and 500 kilometers. Therefore, H3 is accepted.

Based on the tested hypotheses, it can be concluded that patent citations originating from cited HTCE -and Brainport patents are localized. However, the localization is not driven by firms within the ecosystems, but rather by firms outside the ecosystems. With respect to the main research question, it can be concluded that the diffusion of knowledge is spatially stickier within -and around ecosystems, however, the effects of localization are not explained by the diffusion of knowledge within the ecosystems.

As can be concluded from the results as well, the regression models were very sensitive. Therefore, although the effects of localization are present in the results, they must be interpreted carefully. Chapter 6, the chapter on Discussion, will further discuss how the results should be interpreted. Additional to the main results of this research, for both ecosystem groups, it was found that the control patents were estimated to have more citation links with a distance larger than 2500 kilometers. However, since the relationship was no longer significant with the exclusion of patents with only one citation link, the evidence is considered less robust. For the Brainport Eindhoven group, it was found that ecosystem patents were estimated to have more collaborations among firms concerning the application of a patent. Next to that, in the Brainport Eindhoven group, it was found that ecosystem patents were estimated to have a higher number of inventors per patent application.

6. Discussion

6.1 Interpretation of the Results

The results that were found in this research should be interpreted considering the following factors. First, the limitations of patent citations as a proxy for knowledge diffusion should be considered. Secondly, the features of the data, the methodology that was chosen, and the trade-offs that had to be made, should be considered.

In Chapter 2, the use of patent citations as a proxy for knowledge diffusion was reviewed. Here, it was addressed that most of the patent citations are all added by examiners (Alcácer et al., 2009; Criscuolo & Verspagen, 2008). Besides the influence and biases of examiner citations, Corsino et al. (2019) found that measurement errors of patent citations, originating from firms, are rooted in firms' incentives to cite prior art. Although patent citations were considered the best option for this research to measure knowledge diffusion, the results must be interpreted with the cautions that come along with the proxy of patent citations.

With respect to the features of the data, in the results, it was found that most of all ecosystem patents were filed by a handful of large firms. Especially, in the HTCE dataset, Philips and NXP accounted for 94% of all patent applications. Consequently, it can easily be argued that these firms represent the entire ecosystem. Although it was controlled for the effect of the individual firm in the regression models, the results found in the HTCE group are explained by the data of two large firms. Next to that, the significance of the regression models was dependent on the anticipation to control for the effect of the individual firm. All in all, the point of whether the results found in this research are explained by the ecosystem, or by the individual firms, remains a point of discussion.

Concerning the Methodology (Chapter 4) of this research, it was chosen to exclude citation links with no geographical distance (zero kilometers) from the final regressions. Before a decision was taken to exclude this group of citation links, a trade-off had to be made. On the one hand, it was found that self-citations by organizations were still present in the data and the share citation links with no spatial distance was very high in the group of local citations (Figures 33 and 34), and thereby, having a major impact on the estimations of localization. On the other hand, by excluding citation links with no spatial distance, the dataset got manipulated and other forms of self-citations got excluded as well. When the distance of a citation link is zero kilometers, it is most likely that the citation link is either a self-citation by an organization or a self-citation by an inventor. For the latter one, it could be that a self-citing inventor is a mobile inventor, as introduced by Breschi & Lissoni (2009). Consequently, it is likely that these mobile inventors are not included in the empirical part of this research. This is unfortunate since mobilize inventors are likely to be a large contributor to the localized diffusion of knowledge. However, counterintuitively, the effects of localization in support of ecosystems were only significant when citation links with a distance of zero kilometers were excluded (Figure 11).

Nevertheless, although citation links with a distance of zero kilometers were excluded from the data, it doesn't necessarily mean that all mobile inventors are excluded from the empirical research. In case the location of the mobile inventor is not similar on the cited -and citing patent, the citation link is included in the research.

Considering the method, it was chosen to collect the cited patents that were filed between the years 2003-2010. Since the dataset by de Rassenfosse et al. (2019) included patents that were filed between the years 1980 – 2014, any citing patent that was filed after 2014, was not included in this research. With a relatively short citation lag of 4 years, a lot of citing patents and thereby citation links are missing in the dataset. As can be seen in Appendix A, the decay of the number of citation links is very high. Additionally, not all the citing patents found in PATSTAT were covered in the dataset by de Rassenfosse et al. (2019). Consequently, as illustrated in the results, the share of patents with only one citation link was very high. In the results, it was discussed how patents with only citation link are of lower quality and more likely to add noise to the data. Again, a trade-off had to be made on whether to include patents with only one citation link in the final dataset. Moreover, eventually, it was found that this group of citation links was decisive for the statistical significance of the model. In the end, it was chosen to include patents with only one citation link because otherwise a lot of data would be excluded. Nevertheless, it must be considered that patents with only one citation link are likely to add noise to the data.

For the HTCE group, it was found that ecosystems are estimated to have more citation links within the range of 200 kilometers, even when patents with only one citation link were excluded (Table 11). This means that, within the range of 200 kilometers, there is a significant and robust result supporting the hypothesis that HTCE patents are estimated to have more citation links within this range. As can be seen in Figure 16, in the chapter on the Methodology, the range of 200 kilometers represents the exact area of the Benelux. Therefore, within the Benelux, there is a significant and robust result supporting the hypothesis that HTCE patents are estimated to have more citation links within this range.

6.2 Limitations

First of all, it has to be noted that the data used in this research does not reflect the contemporary diffusion of knowledge at the HTCE and Brainport Eindhoven. The patent data from these ecosystems originates from the years 2003 – 2014. So, the results reflect the diffusion of knowledge over a period of more than 10 years ago. In that period, Philips was found to have the most patent applications in both ecosystem groups (Figures 20 and 22). Nowadays, it could be that both ecosystems have changed a lot. Nevertheless, in 2022, Philips was still the leading firm concerning the number of patent applications in that year.[3] Secondly, it must be mentioned that patent citations, which were used as a proxy to measure knowledge diffusion, only capture a small proportion of all the knowledge that is diffused within the ecosystem. Patent citations

[3] https://www.businessinsider.nl/octrooi-innovatie-nederland-philips-signify-2022/

are just one proxy in a wide spectrum of possible indicators to measure knowledge diffusion. Within the boundaries and limitations of this master thesis, patent citations were considered a good starting point to study the diffusion of knowledge within -and around ecosystems.

Concerning the second hypothesis, this research was not able to ensure that only patent citations that were both produced and captured within the ecosystem, were considered when the second hypothesis was tested. Concerning the variables *dist_30* and *dist_50*, both variables do include citation links that might diffuse beyond the borders of the ecosystem, even though the citation links are very local. To ensure that all forward citations were located in the ecosystem as well, a check had to be made concerning the location of the citing patent. However, this was not considered in this research.

In Chapter 3, the chapter on Methodology, the issue was addressed that each patent can have multiple geographical locations. This means that every citation link between two patents can have multiple distances. Next to that, it could be that an international patent appears to feature an inventor that is located within Brainport Eindhoven. However, if all the other inventors are, e.g., located in China, it is reasonable to assume that the heart of the invention is found in China. Therefore, when using the location of the inventor located in Brainport Eindhoven, the distance of the citation link that is eventually calculated might produce a distorted picture. Lastly, in this research, the effect of the patent office has not been included in the OLS regression models. As mentioned by Corsino et al. (2019), the number of citing patents linked to a focal patent is affected by the corresponding patent office. However, within the scope and limits of this research, the influence of patent offices has not been considered.

6.3 Implications

6.3.1 Contribution to the Literature

In this section, the implications of this research will be provided. First, the results that are found in this research do matter and are scientifically relevant. Although the imperfections and sensitivities of the method and the regression model, the findings in this research indicate that the knowledge diffused from ecosystems is spatially stickier compared to the knowledge diffused from non-ecosystems. Since the stickiness of knowledge diffusion and ecosystems has barely been studied empirically up to the time of writing in this thesis, this research complements the current literature about ecosystems and knowledge diffusion. In the recent literature, a lot is written about ecosystems and the proximal advantages that can be found in ecosystems, however, empirical findings on how knowledge diffuses within these ecosystems are rare. Therefore, this research places an important contribution to the current scientific literature on ecosystems and knowledge diffusion.

6.3.2 Policy and Managerial Implications

Since the HTCE and Brainport Eindhoven are of high importance to the national economy, as mentioned in Chapter 1, any insights that help to gain an understanding of the behavior of these ecosystems are of importance and should be of interest to national policy on innovation, the municipality of Eindhoven and the HTCE and Brainport Eindhoven itself. For example, with

results showing that patents originating from cited HTCE patents are estimated to have more citations links within the Benelux, it can be concluded that the knowledge produced at the HTCE is very important, even more, important than similar knowledge within the Netherlands and Belgium, concerning innovation within the Benelux. Besides the effects of localization, for both ecosystems, it is valuable to understand to what actors and regions their knowledge are diffused and captured. From a managerial point of view, with the use of data about patent citations, potential alliances, and competitors can be identified.

6.4 Recommendations

Lastly, in this section, recommendations for future research will be provided. Firstly, the methodology that was used in this research can be applied to study all ecosystems worldwide. In this research, it was chosen to examine the HTCE and Brainport Eindhoven. However, to find a profound effect of the spatial effect of knowledge diffusion and ecosystems, it is desirable to apply the methodology of this research to other ecosystems. Secondly, although the dataset by de Rassenfosse et al. (2019) facilitated this research, it would be ideal to have more recent geographical data about patents. This would make the results more significant and relevant for policymakers and managers.

Thirdly, concerning the method, it would be recommended to increase the range of the citation lag. In that way, more citing patents can be included, which increases the quality of the patents. The issue of patents with only one citation link was found to be a serious hindrance in this research. Fourthly, besides patent citations, it would be recommended to explore other aspects of localized knowledge diffusion within -and around ecosystems, e.g., the diffusion of literature and licensing agreements.

Lastly, in this research, it was found that Brainport Eindhoven patents are estimated to have more inventors registered on a patent application (Table 13). As a recommendation for future research, it would be very interesting to study the networks of inventors within -and around ecosystems, like the study by Breschi & Lissoni (2009), to find out whether these networks are different for ecosystems compared to networks in non-ecosystems. Moreover, it is worthwhile to study whether the localization effects, that were found in this research, are mediated by the mobility of inventors.

7. References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2020). The role of geographical proximity in knowledge diffusion, measured by citations to scientific literature. *Journal of Informetrics*, *14*(1), 101010. https://doi.org/10.1016/j.joi.2020.101010

Adner, R. (2006). *Match Your Innovation Strategy to Your Innovation Ecosystem*. www.hbrreprints.org

Alcácer, J., & Gittelman, M. (2006). Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. *Review of Economics and Statistics*, *88*(4), 774–779. https://doi.org/10.1162/rest.88.4.774

Alcácer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy*, *38*(2), 415–427. https://doi.org/10.1016/j.respol.2008.12.001

Almeida, P., & Kogut, B. (1999). Localization of Knowledge and the Mobility of Engineers in Regional Networks. *Management Science*, *45*(7), 905–917. https://doi.org/10.1287/mnsc.45.7.905

Belenzon, S., & Schankerman, M. (2013). Spreading the Word: Geography, Policy, and Knowledge Spillovers. *Review of Economics and Statistics*, *95*(3), 884–903. https://doi.org/10.1162/REST_a_00334

Borgh, M., Cloodt, M., & Romme, A. G. L. (2012). Value creation by knowledge-based ecosystems: evidence from a field study. *R&D Management*, *42*(2), 150–169. https://doi.org/10.1111/j.1467-9310.2011.00673.x

Boschma, R. (2005). Proximity and Innovation: A Critical Assessment. *Regional Studies*, *39*(1), 61–74. https://doi.org/10.1080/0034340052000320887

Brainport Eindhoven. (2022). *Toekomstgericht en vooruitstrevend*.

Breschi, S., & Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, *9*(4), 439–468. https://doi.org/10.1093/jeg/lbp008

Capello, R., & Varga, A. (2013). Knowledge creation and knowledge diffusion in space and regional innovation performance: introductory remarks. *The Annals of Regional Science*, *51*(1), 113–118. https://doi.org/10.1007/s00168-013-0569-x

Castaldi, C., Frenken, K., & Los, B. (2015). Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting. *Regional Studies*, *49*(5), 767–781. https://doi.org/10.1080/00343404.2014.940305

Chesbrough, H. W. (2006). *Open Innovation: The New Imperative for Creating and Profiting from Technology*.

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. *Scientometrics*, *126*(7), 5431–5476. https://doi.org/10.1007/s11192-021-04001-1

Cobben, D., Ooms, W., Roijakkers, N., & Radziwon, A. (2022). Ecosystem types: A systematic review on boundaries and goals. *Journal of Business Research*, *142*, 138–164. https://doi.org/10.1016/j.jbusres.2021.12.046

Corsino, M., Mariani, M., & Torrisi, S. (2019). Firm strategic behavior and the measurement of knowledge flows with patent citations. *Strategic Management Journal*, *40*(7), 1040–1069. https://doi.org/10.1002/smj.3016

Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy*, *37*(10), 1892–1908. https://doi.org/10.1016/j.respol.2008.07.011

de Rassenfosse, G., Kozak, J., & Seliger, F. (2019). Geocoding of worldwide patent data. *Scientific Data*, *6*(1), 260. https://doi.org/10.1038/s41597-019-0264-6
Dealroom. (2022). *The next generation of tech ecosystems*.
Fontana, R., Nuvolari, A., & Verspagen, B. (2009). Mapping technological trajectories as patent citation networks. An application to data communication standards. *Economics of Innovation and New Technology*, *18*(4), 311–336. https://doi.org/10.1080/10438590801969073
Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469
Granstrand, O., & Holgersson, M. (2020). Innovation ecosystems: A conceptual review and a new definition. *Technovation*, *90–91*, 102098. https://doi.org/10.1016/j.technovation.2019.102098
Hall, B., Trajtenberg, M., & Jaffe, A. (2005). Market Value and Patent Citations. *The RAND Journal of Economics* , *36*(1), 16–38.
Henderson, R., Jaffe, A., & Trajtenberg, M. (2005). Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Comment. *American Economic Review*, *95*(1), 461–464. https://doi.org/10.1257/0002828053828644
High Tech Campus Eindhoven. (2023). *HighTechXL*. Https://Hightechcampus.Com/Nl/Hightechxl.
Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, *39*(8), 2255–2276. https://doi.org/10.1002/smj.2904
Jaffe, A. B., Fogarty, M. S., & Banks, B. A. (1998). Evidence from Patents and Patent Citations on the Impact of NASA and Other Federal Labs on Commercial Innovation. *The Journal of Industrial Economics*, *46*(2), 183–205. https://doi.org/10.1111/1467-6451.00068
Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors. *American Economic Review*, *90*(2), 215–218. https://doi.org/10.1257/aer.90.2.215
Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, *108*(3), 577–598. https://doi.org/10.2307/2118401
Kijek, & Kijek. (2019). Knowledge Spillovers: An Evidence from The European Regions. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(3), 68. https://doi.org/10.3390/joitmc5030068
Koelman, K. (2021). Orchestrating the AI Ecosystem in Brainport Eindhoven: the role of the AI Innovation Center. *IE&IS*. https://research.tue.nl/en/studentTheses/orchestrating-the-ai-ecosystem-in-brainport-eindhoven
Marshall, A. (1920). *Principles of Economics* (8th ed.). MacMillan .
Moore, J. F. (1993). Predators and prey: a new ecology of competition. *Harvard Business Review*, *71*(3), 75–86.
Mowery, D. C., & Ziedonis, A. A. (2015). Markets versus spillovers in outflows of university research. *Research Policy*, *44*(1), 50–66. https://doi.org/10.1016/j.respol.2014.07.019
Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy*, *38*(6), 994–1005. https://doi.org/10.1016/j.respol.2009.01.023
Paci, R., Marrocu, E., & Usai, S. (2014). The Complementary Effects of Proximity Dimensions on Knowledge Spillovers. *Spatial Economic Analysis*, *9*(1), 9–30. https://doi.org/10.1080/17421772.2013.856518

Park, C., Ghauri, P. N., Lee, J. Y., & Golmohammadi, I. (2022). Unveiling the black box of IJV innovativeness: The role of explicit and tacit knowledge transfer. *Journal of International Management*, *28*(4), 100956. https://doi.org/10.1016/j.intman.2022.100956

Robertson, J., Caruana, A., & Ferreira, C. (2023). Innovation performance: The effect of knowledge-based dynamic capabilities in cross-country innovation ecosystems. *International Business Review*, *32*(2), 101866. https://doi.org/10.1016/j.ibusrev.2021.101866

Rogers, E. M., Takegami, S., & Yin, J. (2001). Lessons learned about technology transfer. *Technovation*, *21*(4), 253–261. https://doi.org/10.1016/S0166-4972(00)00039-0

Romme, S. (2022, July 6). *Succes van Brainport Eindhoven is uniek en niet eenvoudig te repliceren*. Https://Esb.Nu/Succes-van-Brainport-Eindhoven-Is-Uniek-En-Niet-Eenvoudig-Te-Repliceren/.

Stam, E. (2015). Entrepreneurial Ecosystems and Regional Policy: A Sympathetic Critique. *European Planning Studies*, *23*(9), 1759–1769. https://doi.org/10.1080/09654313.2015.1061484

Talbot, D. (2021). Proximity – Impacts of Geographic, Organizational and Cognitive Proximities on Innovation. In *Innovation Economics, Engineering and Management Handbook 1* (pp. 293–298). Wiley. https://doi.org/10.1002/9781119832492.ch36

Thompson, P., & Fox-Kean, M. (2005). Patent Citations and the Geography of Knowledge Spillovers: A Reassessment. *American Economic Review*, *95*(1), 450–460. https://doi.org/10.1257/0002828053828509

Trajtenberg M. (1990). A Penny for Your Quotes: Patent Citations and the Value of Innovations. *The RAND Journal of Economics* , *21*(1), 172–187.

USPTO. (2022, June 13). *Trademark, patent, or copyright* . USPTO.

Van Leest, E., Janssen, C., & Brouwers, B. (2022). *Brainport als sleutel tot startup succes*.

Wang, Q. R., & Zheng, Y. (2023). Patent regime and the geography of cumulative innovation. *Research Policy*, *52*(7), 104809. https://doi.org/10.1016/j.respol.2023.104809

WIPO. (2023, June 15). *International Patent Classification (IPC)*. Https://Ipcpub.Wipo.Int/?Notion=scheme&version=20230101&symbol=none&menulang=en&lang=en&viewmode=f&fipcpc=no&showdeleted=yes&indexes=no&headings=yes&notes=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart.

Wu, H., Han, Z., & Zhou, Y. (2021). Optimal degree of openness in open innovation: A perspective from knowledge acquisition & knowledge leakage. *Technology in Society*, *67*, 101756. https://doi.org/10.1016/j.techsoc.2021.101756

Yang, P., Liu, X., Hu, Y., & Gao, Y. (2022). Entrepreneurial ecosystem and urban economic growth-from the knowledge-based view. *Journal of Digital Economy*, *1*(3), 239–251. https://doi.org/10.1016/j.jdec.2023.02.002

Zucker, L. G., Darby, M. R., & Armstrong, J. (1998). Geographically localized knowledge: spillovers or markets? *Economic Inquiry*, *36*(1), 65–86. https://doi.org/10.1111/j.1465-7295.1998.tb01696.x

Appendix A: SQL Transcript

This Appendix covers the transcript of the code that was used in SQL during this research. The content of this Appendix is structured as displayed in the Table of Contents on page X. First, a brief overview of the most important tables will be provided. This is done to illustrate how the quantity of ecosystem -and control patents changed throughout the process of data treatment. After that, for each separate sample, the transcript of the code is provided.

Brief summary of the Microsoft SQL Tables

Table	Explanation	Number of observations
‘dbo.one_g’	Citation links including self-citations	7280
‘dbo.one_n’	Citation links excluding self-citations	7031
‘dbo.one_q’	Citation links including coordinates, multiple locations per appln_id present.	5312
‘dbo.one_r’	Unique citations links. This represents the quantity of citation links of which the geo-coordinates are known.	2206
‘dbo.two_g’	All control patents per one HTCE patent.	2420
‘Sample1_15’	Final table HTCE patents and citing patens	769
‘dbo.two_h’	All unique control patents	1956
‘dbo.two_t’	All citation links of control patents and citing patents: - Excluding self-citations. - Excluding co-invented patents within brainport region.	11878
‘dbo.two_z’	Citation links including coordinates, multiple locations per appln_id present.	17269
‘dbo.two_y’	Unique citations links. This represents the quantity of citation links of which the geo-coordinates are known. Roughly 50% of ‘two_t’.	5783
‘Sample2_16’	Final table of control patents	857

Table 14. Number of observations for the HTCE -and control group.

Table	Explanation	Number of observations
‘dbo.three_g’	Citation links including self-citations	62772
‘dbo.three_n’	Citation links excluding self-citations	50702
‘dbo.three_q’	Citation links including coordinates, multiple locations per appln_id present.	46468
‘dbo.three_r’	Unique citations links. This represents the quantity of citation links of which the geo-coordinates are known.	16578
‘sample3_15’	Final table Brainport patents	2992
‘dbo.four_g’	A control patents per one brainport patent.	7454
‘dbo.four_h’	All unique control patents.	4688
‘dbo.four_t’	All citation links of control patents and citing patents: - Excluding self-citations. - Excluding co-invented patents within brainport region.	26509
‘dbo.four_z’	Citation links including coordinates, multiple locations per appln_id present.	35256
‘dbo.four_y’	Unique citations links. This represents the quantity of citation links of which the geo-coordinates are known. Roughly 60% of ‘four_t’.	11613
‘sample4_16’	Final table control patents	1900

Table 15. Number of observations for the Brainport Eindhoven -and control group.

Sample 1: HTCE patents

PART I: Collection of the HTCE patents and the citing patents.

HTCE patents

```
-- Step 1: Select HTCE patents from Rassenfosse database.
-- Conditions: coordinates and filing date.

select Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.*
into one_a
from Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean

where ((lat < 51.416598 AND lat > 51.403) AND (lng < 5.47028 AND lng > 5.447))
AND filing_date > '2003-01-01'
AND filing_date < '2011-01-01'

-- Step 2: Isolate appln_id of HTCE patents.
-- Remove duplicates

select dbo.one_a.appln_id as appln_id_HTCE
into one_b
from dbo.one_a

with CTE as
(Select appln_id_HTCE,
row_number() over (partition by appln_id_HTCE order by appln_id_HTCE) rn
from dbo.one_b)

delete from CTE where rn > 1

-- Step 3: Add pat_publn_id to appln_id of HTCE patents.

select dbo.one_b.appln_id_HTCE, patstat2020a.dbo.tls211_pat_publn.pat_publn_id as
pat_publn_id_HTCE
into one_c
from dbo.one_b
join patstat2020a.dbo.tls211_pat_publn on patstat2020a.dbo.tls211_pat_publn.appln_id =
dbo.one_b.appln_id_HTCE
```

Citing patents

```
-- Step 4: Find the citing patents.
-- Add the pat_publn_id of the citing patents.

select dbo.one_c.*, patstat2020a.dbo.tls212_citation.pat_publn_id as
pat_publn_id_citing
into one_d
from dbo.one_c
join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.cited_pat_publn_id = dbo.one_c.pat_publn_id_HTCE

-- Step 5: Add appln_id of the citing patents.

Select dbo.one_d.*, patstat2020a.dbo.tls211_pat_publn.appln_id as appln_id_citing
into one_e
from dbo.one_d
join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.pat_publn_id = dbo.one_d.pat_publn_id_citing
```

```
-- Step 6: Delete duplicate rows (if any).

with CTE as
(Select appln_id_HTCE, pat_publn_id_HTCE, pat_publn_id_citing, appln_id_citing,
row_number() over (partition by appln_id_HTCE, pat_publn_id_HTCE, pat_publn_id_citing,
appln_id_citing order by appln_id_HTCE, pat_publn_id_HTCE, pat_publn_id_citing,
appln_id_citing) rn
from dbo.one_e)

delete from CTE where rn > 1

-- 347 duplicates are removed
-- 10912 rows remain.

select dbo.one_e.* from dbo.one_e

-- Step 7: Create a table only with appln_id_HTCE, appln_id_citing. That's everything
what is needed.

select dbo.one_e.appln_id_HTCE, dbo.one_e.appln_id_citing
into one_f
from dbo.one_e

-- Step 8: Remove duplicates again.

with CTE as
(select appln_id_HTCE, appln_id_citing,
row_number() over (partition by appln_id_HTCE, appln_id_citing
order by appln_id_HTCE, appln_id_citing) rn
from dbo.one_f)

delete from CTE where rn > 1

-- 439 rows removed.

select dbo.one_f.* from dbo.one_f

-- 10473 unique citation links remain.

-- step 9: Add filing_date to citing patents.
-- Only include citations between 2003 - 2014.

select dbo.one_f.*, patstat2020a.dbo.tls201_appln.appln_filing_date
into one_g
from dbo.one_f
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
dbo.one_f.appln_id_citing
where appln_filing_date < '2015-01-01'
```

Self-citations

```
-- Step 10: Remove self-citations.
-- Add person_id to patents.

select dbo.one_g.appln_id_HTCE, patstat2020a.dbo.tls207_pers_appln.person_id as
person_id_HTCE, dbo.one_g.appln_id_citing
into one_h
from dbo.one_g
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.one_g.appln_id_HTCE
```

```
select dbo.one_h.*, patstat2020a.dbo.tls207_pers_appln.person_id as person_id_citing
into one_i
from dbo.one_h
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.one_h.appln_id_citing

-- Step 11: Add inventor sequence number.
-- This number indicates whether a person is an organization yes or not.

select distinct dbo.one_i.person_id_HTCE,
patstat2020a.dbo.tls227_pers_publn.invt_seq_nr
into one_j
from dbo.one_i
join patstat2020a.dbo.tls227_pers_publn on
patstat2020a.dbo.tls227_pers_publn.person_id = dbo.one_i.person_id_HTCE

select distinct dbo.one_i.*, dbo.one_j.invt_seq_nr
into one_k
from dbo.one_i
join dbo.one_j on dbo.one_j.person_id_HTCE = dbo.one_i.person_id_HTCE

-- Step 12: identify self-citations by organizations.
-- Filter on invt_seq_nr = 0. This are citations by organizations.

select dbo.one_k.*
from dbo.one_k
where ((person_id_HTCE = person_id_citing) AND (invt_seq_nr = 0))

-- Result: 268 self-citations.

-- Step 13: Create a table of the self-citations.

select dbo.one_k.*
into one_l
from dbo.one_k
where ((person_id_HTCE = person_id_citing) AND (invt_seq_nr = 0))

select dbo.one_l.appln_id_HTCE, dbo.one_l.appln_id_citing
into one_m
from dbo.one_l

with CTE as
(select appln_id_HTCE, appln_id_citing,
row_number() over (partition by appln_id_HTCE, appln_id_citing
order by appln_id_HTCE, appln_id_citing) rn
from dbo.one_m)

delete from CTE where rn > 1

-- 19 duplicates removed.

-- Step 14: Delete citing patents from the data.

select dbo.one_g.appln_id_HTCE, dbo.one_g.appln_id_citing
into one_n
from dbo.one_g

delete one_n
from one_n
inner join one_m
on one_n.appln_id_HTCE = one_m.appln_id_HTCE AND
```

```
      one_n.appln_id_citing = one_m.appln_id_citing;

-- Result: 249 self-citations removed.

-- Step 15: check results.

select dbo.one_n.* from dbo.one_n

-- 7031 patents.
```

Geo-coordinates

```
-- Step 16: Find the geo-coordinates of the citation links.
-- Link latitude and longitude to appln_id_HTCE

select dbo.one_n.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_HTCE, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as lng_HTCE
into one_o
from dbo.one_n
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.one_n.appln_id_HTCE

-- NOTE: one filing can more locations. Therefore, more results.

-- Step 17: Filter out the non-HTCE locations.

select dbo.one_o.*
into one_p
from dbo.one_o
where ((lat_HTCE < 51.416598 AND lat_HTCE > 51.403) AND (lng_HTCE < 5.47028 AND
lng_HTCE > 5.447))

-- Step 18: Find geo-coordinates of the citing patents.

select dbo.one_p.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_citing, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as lng_citing
into one_q
from dbo.one_p
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.one_p.appln_id_citing

-- NOTE: Quite a proportion of the citing patents are not known in the dataset by
Rassenfosse et al. (2019)
-- Therefore, the geo-cordinates of some of the citing patents are unknown.

-- Step 19: Remove duplicates.

with CTE as
(Select appln_id_HTCE, appln_id_citing, lat_HTCE, lng_HTCE, lat_citing, lng_citing,
row_number() over (partition by appln_id_HTCE, appln_id_citing, lat_HTCE, lng_HTCE,
lat_citing, lng_citing order by appln_id_HTCE, appln_id_citing, lat_HTCE, lng_HTCE,
lat_citing, lng_citing) rn
from dbo.one_q)

delete from CTE where rn > 1

select dbo.one_q.* from dbo.one_q


-- Result: 5312 rows.
```

Quantity of neglected citations

```
-- Step 20: Isolate appln_id_HTCE and appln_id_citing

select dbo.one_q.appln_id_HTCE, dbo.one_q.appln_id_citing
into one_r
from dbo.one_q

-- Step 21: Remove duplicates.

with CTE as
(select appln_id_HTCE, appln_id_citing,
row_number() over (partition by appln_id_HTCE, appln_id_citing
order by appln_id_HTCE, appln_id_citing) rn
from dbo.one_r)

delete from CTE where rn > 1

select dbo.one_r.* from dbo.one_r

-- Step 22: Analysis

-- Only 2206 citations are still present.
-- This means that for almost 70% of the citation links the geo-coordinates are
missing in the dataset by Rassenfosse et al. (2019).
-- Also, the dataset by Rassenfosse et al. probably has a small proportion of the HTCE
patents present.
```

PART II: Variables

All information related to distance.

```
-- Step 1: Save data, including haversine distance, as new table: sample1_1

select sample1.*
into sample1_1
from sample1

-- Step 2: Find the shortest distance per citation link.
-- Since there are multiple inventors, there are multiple distances per citation link.

with CTE as
(Select appln_id_HTCE, appln_id_citing, Haversine,
row_number() over (partition by appln_id_HTCE, appln_id_citing order by haversine asc)
rn
from dbo.sample1_1)

delete from CTE where rn > 1

select sample1_1.* from sample1_1

-- Result: 2206 citation links.

-- Step 3: Per HTCE patent, find the shortest citation. This will be used as a proxy
for distance in the model.

Select appln_id_HTCE, appln_id_citing, Haversine,
row_number() over (partition by appln_id_HTCE order by haversine asc) rn
into sample1_2
```

```
from dbo.sample1_1

select appln_id_HTCE, appln_id_citing, Haversine as shortest_citation
into sample1_3
from Sample1_2
where rn=1

-- Step 4: Import sample1_4
-- This is the same data as sample1_1, but then the distance is rounded off.

select sample1_4.* from sample1_4

-- Step 5: per HTCE patent, flag quanitiy of citations within a certain range.
-- range 30 km.

select appln_id_HTCE, haversine as dist_30
into sample1_dist30_1
from sample1_4
where haversine < 30

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_30
into sample1_dist30
from sample1_dist30_1
group by appln_id_HTCE

-- 50 km.

select appln_id_HTCE, haversine as dist_50
into sample1_dist50_1
from sample1_4
where haversine < 50

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_50
into sample1_dist50
from sample1_dist50_1
group by appln_id_HTCE

-- 100 km

select appln_id_HTCE, haversine as dist_100
into sample1_dist100_1
from sample1_4
where haversine < 100

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_100
into sample1_dist100
from sample1_dist100_1
group by appln_id_HTCE

-- 200 km

select appln_id_HTCE, haversine as dist_200
into sample1_dist200_1
from sample1_4
where haversine < 200

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_200
into sample1_dist200
from sample1_dist200_1
group by appln_id_HTCE

-- 500 km
```

```
select appln_id_HTCE, haversine as dist_500
into sample1_dist500_1
from sample1_4
where haversine < 500

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_500
into sample1_dist500
from sample1_dist500_1
group by appln_id_HTCE

-- 1000 km

select appln_id_HTCE, haversine as dist_1000
into sample1_dist1000_1
from sample1_4
where haversine < 1000

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_1000
into sample1_dist1000
from sample1_dist1000_1
group by appln_id_HTCE

-- 2500 km

select appln_id_HTCE, haversine as dist_2500
into sample1_dist2500_1
from sample1_4
where haversine < 2500

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_2500
into sample1_dist2500
from sample1_dist2500_1
group by appln_id_HTCE

-- 5000 km

select appln_id_HTCE, haversine as dist_5000
into sample1_dist5000_1
from sample1_4
where haversine < 5000

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_5000
into sample1_dist5000
from sample1_dist5000_1
group by appln_id_HTCE

-- 0 km

select appln_id_HTCE, haversine as dist_0
into sample1_dist0_1
from sample1_4
where haversine = 0

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_0
into sample1_dist0
from sample1_dist0_1
group by appln_id_HTCE

-- Step 6: Calculate average distance of all citation links per HTCE patent.
```

```
select appln_id_HTCE, avg(haversine) as avg_haversine, count(appln_id_HTCE) as
nb_citations
into sample1_5
from Sample1_4
group by appln_id_HTCE

-- Step 7: Put all distance related information into one table.

select sample1_5.*, sample1_3.shortest_citation, sample1_dist0.dist_0,
sample1_dist30.dist_30, sample1_dist50.dist_50, sample1_dist100.dist_100,
sample1_dist200.dist_200, sample1_dist500.dist_500, sample1_dist1000.dist_1000,
sample1_dist2500.dist_2500, sample1_dist5000.dist_5000
into sample1_6
from sample1_5
join sample1_3 on sample1_3.appln_id_HTCE = sample1_5.appln_id_HTCE
LEFT join sample1_dist0 on sample1_dist0.appln_id_HTCE = sample1_5.appln_id_HTCE
left join  sample1_dist30 on sample1_dist30.appln_id_HTCE = sample1_5.appln_id_HTCE
left join  sample1_dist50 on sample1_dist50.appln_id_HTCE = sample1_5.appln_id_HTCE
left join  sample1_dist100 on sample1_dist100.appln_id_HTCE = sample1_5.appln_id_HTCE
left join  sample1_dist200 on sample1_dist200.appln_id_HTCE = sample1_5.appln_id_HTCE
left join  sample1_dist500 on sample1_dist500.appln_id_HTCE = sample1_5.appln_id_HTCE
left join  sample1_dist1000 on sample1_dist1000.appln_id_HTCE =
sample1_5.appln_id_HTCE
left join  sample1_dist2500 on sample1_dist2500.appln_id_HTCE =
sample1_5.appln_id_HTCE
left join  sample1_dist5000 on sample1_dist5000.appln_id_HTCE =
sample1_5.appln_id_HTCE

select sample1_6.* from sample1_6
```

Control variables

```
-- Step 1: Find number of forward -and backward citations
-- Step 1a: Forward citations
-- This includes all citations after 2014 as well.

select one_f.* from one_f

select appln_id_HTCE, COUNT(appln_id_HTCE) as nb_forward_citations
into sample1_7_1
from one_f
group by appln_id_HTCE

-- Step 1b: Backward citations

select sample1_6.appln_id_HTCE, patstat2020a.dbo.tls211_pat_publn.pat_publn_id
into sample1_7_2
from sample1_6
join  patstat2020a.dbo.tls211_pat_publn on  patstat2020a.dbo.tls211_pat_publn.appln_id
= sample1_6.appln_id_HTCE

select sample1_7_2.appln_id_HTCE, sample1_7_2.pat_publn_id,
patstat2020a.dbo.tls212_citation.cited_pat_publn_id
into sample1_7_3
from sample1_7_2
join patstat2020a.dbo.tls212_citation on patstat2020a.dbo.tls212_citation.pat_publn_id
= sample1_7_2.pat_publn_id
where cited_pat_publn_id != 0

select appln_id_HTCE, COUNT(appln_id_HTCE) as nb_backward_citations
into sample1_7_4
```

```
from sample1_7_3
group by appln_id_HTCE

select sample1_6.appln_id_HTCE, sample1_7_1.nb_forward_citations,
sample1_7_4.nb_backward_citations
into sample1_7
from sample1_6
left join sample1_7_1 on sample1_7_1.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_7_4 on sample1_7_4.appln_id_HTCE = sample1_6.appln_id_HTCE

-- Step 2: IPC and year

-- most frequent IPC code and year can be found in ""two_e"

-- Find number of IPC codes per HTCE patent.

select sample1_6.appln_id_HTCE, patstat2020a.dbo.tls209_appln_ipc.ipc_class_symbol
into sample1_8_1
from sample1_6
left join patstat2020a.dbo.tls209_appln_ipc on
patstat2020a.dbo.tls209_appln_ipc.appln_id = sample1_6.appln_id_HTCE

select appln_id_HTCE, COUNT(appln_id_HTCE) as nb_IPC_codes
into sample1_8
from sample1_8_1
group by appln_id_HTCE

-- Step 3: nb_claims. Select patent with highest claims if patent is similar.

select sample1_6.appln_id_HTCE, patstat2020a.dbo.tls211_pat_publn.publn_claims
into sample1_9_1
from sample1_6
left join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.appln_id = sample1_6.appln_id_HTCE

with CTE as
(Select appln_id_HTCE, publn_claims,
row_number() over (partition by appln_id_HTCE order by publn_claims desc) rn
from dbo.sample1_9_1)

delete from CTE where rn > 1

select sample1_9_1.* from sample1_9_1

-- Step 4: Backward citations to literature

select sample1_7_2.* from sample1_7_2

select sample1_7_2.*, patstat2020a.dbo.tls212_citation.cited_npl_publn_id
into sample1_10_1
from sample1_7_2
left join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.pat_publn_id = sample1_7_2.pat_publn_id
where cited_npl_publn_id != '0'

select appln_id_HTCE, COUNT(appln_id_HTCE) as nb_cited_literature
into sample1_10
from sample1_10_1
group by appln_id_HTCE

-- Step 5: Find company names
```

```
-- In case of multiple companies per one HTCE patent, choose one company name at
random per one HTCE patent.

select sample1_6.appln_id_HTCE, patstat2020a.dbo.tls207_pers_appln.person_id
into sample1_11_1
from sample1_6
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= sample1_6.appln_id_HTCE

select sample1_11_1.appln_id_HTCE, patstat2020a.dbo.tls206_person.person_name,
patstat2020a.dbo.tls206_person.psn_level, patstat2020a.dbo.tls206_person.psn_sector
into sample1_11_2
from sample1_11_1
join patstat2020a.dbo.tls206_person on patstat2020a.dbo.tls206_person.person_id =
sample1_11_1.person_id
where psn_sector = 'company'

select sample1_11_2.* from sample1_11_2

WITH CTE AS (
    SELECT
        appln_id_HTCE,
        person_name,
        ROW_NUMBER() OVER(PARTITION BY appln_id_HTCE ORDER BY NEWID()) AS RowNum
    FROM
        dbo.sample1_11_2)

SELECT
    appln_id_HTCE,
    person_name INTO sample1_11
FROM
    CTE
WHERE
    RowNum = 1

select sample1_11.* from sample1_11

-- Step 6: Find quantity of foreign inventors/locations per HTCE patent.
-- Quantity of locations/inventors outside HTCE

select sample1_6.appln_id_HTCE,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
into sample1_12_1
from sample1_6
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
sample1_6.appln_id_HTCE
where not ((lat < 51.416598 AND lat > 51.403) AND (lng < 5.47028 AND lng > 5.447))

select sample1_12_1.appln_id_HTCE, count(appln_id_HTCE) as nb_outside_inventors
into sample1_12_2
from sample1_12_1
group by appln_id_HTCE

-- Number of foreign inventors

select sample1_6.appln_id_HTCE,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
```

```
into sample1_12_3
from sample1_6
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
sample1_6.appln_id_HTCE
where ctry_code != 'NL' and ctry_code != 'BE'

select sample1_12_3.appln_id_HTCE, count(appln_id_HTCE) as nb_foreign_inventors
into sample1_12_4
from sample1_12_3
group by appln_id_HTCE

-- Create overall table of inventors

select sample1_6.appln_id_HTCE, sample1_12_2.nb_outside_inventors,
sample1_12_4.nb_foreign_inventors
into sample1_12
from sample1_6
left join sample1_12_2 on sample1_12_2.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_12_4 on sample1_12_4.appln_id_HTCE = sample1_6.appln_id_HTCE

-- Step 7: Add nb_applicants, only companies

select sample1_11_2.* from sample1_11_2

select sample1_11_2.appln_id_HTCE, count(appln_id_HTCE) as nb_applicants_companies
into sample1_13
from sample1_11_2
group by appln_id_HTCE

-- Step 8: Put everything into one table.

select sample1_6.*, patstat2020a.dbo.tls201_appln.appln_auth,
patstat2020a.dbo.tls201_appln.nb_inventors, sample1_13.nb_applicants_companies,
sample1_11.person_name, sample1_12.nb_foreign_inventors,
sample1_12.nb_outside_inventors, two_e.IPC, sample1_8.nb_IPC_codes,
two_e.earliest_filing_year, sample1_7.nb_backward_citations,
sample1_7.nb_forward_citations,
sample1_10.nb_cited_literature, sample1_9_1.publn_claims
into sample1_14
from sample1_6
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
sample1_6.appln_id_HTCE
LEFT join sample1_13 on sample1_13.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_11 on sample1_11.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_12 on sample1_12.appln_id_HTCE = sample1_6.appln_id_HTCE
left join two_e on two_e.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_8 on sample1_8.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_7 on sample1_7.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_10 on sample1_10.appln_id_HTCE = sample1_6.appln_id_HTCE
left join sample1_9_1 on sample1_9_1.appln_id_HTCE = sample1_6.appln_id_HTCE
```

Sample 2: Control patents of the HTCE patents.

Part I: Collection of the control patents and the citing patents.

Find all possible control patents, including IPC and year.

```
-- Step 1: Find all possible control patents. From Belgium and Netherlands. Exclude
location of the Brainport region.

select distinct Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id
into two_a
from Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean
where ((name_0 = 'Belgium') OR (name_0 = 'Netherlands'))
AND not ((lat < 51.541469 AND lat > 51.318343) AND (lng < 5.875337 AND lng >
5.231379))
AND filing_date > '2003-01-01'
AND filing_date < '2011-01-01'

-- Although patents from the brainport are excluded. It could be that they're still
present.
-- All the appln_id's, having brainport coordinates, are removed from the set.
-- However, since a patent can have more than one location. It could be that a
brainport appln_id 'patent' is still present when that patent has more then one
location, including a location outside the brainport region.
-- Therefore, these patents that have a location within -and outside the brainport
region, have to be removed.

select dbo.two_a.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng
into two_a_1
from dbo.two_a
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id = dbo.two_a.appln_id

select dbo.two_a_1.*
into two_a_2
from dbo.two_a_1
where ((lat < 51.541469 AND lat > 51.318343) AND (lng < 5.875337 AND lng > 5.231379))

with CTE as
(Select appln_id,
row_number() over (partition by appln_id order by appln_id) rn
from dbo.two_a_2)

delete from CTE where rn > 1

select dbo.two_a_2.* from dbo.two_a_2

-- So, there are 809 Brainport patents still present in the possible pool of control
patents.

delete two_a
from two_a
inner join two_a_2
on two_a_2.appln_id = two_a.appln_id

select dbo.two_a.* from dbo.two_a

-- 809 patents removed. There are now 33965 possible control patents in the pool.
```

```
-- Step 2: Find IPC, year of pool.

 select dbo.two_a.appln_id as appln_id_controls,
left(patstat2020a.dbo.tls209_appln_ipc.ipc_class_symbol, 4) as IPC,
patstat2020a.dbo.tls201_appln.earliest_filing_year
 into two_b
 from dbo.two_a
 join patstat2020a.dbo.tls209_appln_ipc on patstat2020a.dbo.tls209_appln_ipc.appln_id
= dbo.two_a.appln_id
 join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
dbo.two_a.appln_id

 select dbo.two_b.* from dbo.two_b

   -- Step 3: For each appln_id, find the most frequent IPC code. Create table with
one unique appln_id and IPC code per row.
  WITH cte AS (
SELECT
    appln_id_controls
  , IPC
  , earliest_filing_year
  , ROW_NUMBER() OVER (PARTITION BY appln_id_controls ORDER BY COUNT(IPC) DESC) rn
FROM dbo.two_b
GROUP BY
  appln_id_controls,
  IPC,
  earliest_filing_year)

SELECT
    appln_id_controls,
       IPC,
       earliest_filing_year
into two_c
FROM cte WHERE rn = 1

select dbo.two_c.* from dbo.two_c

-- Result: 31,181 patents. Less than dataset at begin. But duplicates are removed.
```

Match HTCE patents with control patents at random

```
-- Step 4: Add IPC code, year to HTCE patents

select dbo.one_b.*, left(patstat2020a.dbo.tls209_appln_ipc.ipc_class_symbol, 4) as
IPC, patstat2020a.dbo.tls201_appln.earliest_filing_year
into two_d
from dbo.one_b
join patstat2020a.dbo.tls209_appln_ipc on patstat2020a.dbo.tls209_appln_ipc.appln_id =
dbo.one_b.appln_id_HTCE
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
dbo.one_b.appln_id_HTCE

-- Each patent has multiple IPC codes. Therefore, create a query that finds the most
frequent IPC code.
-- That IPC will be used to find control patent.

SELECT
    appln_id_HTCE
  , IPC
  , earliest_filing_year
  , ROW_NUMBER() OVER (PARTITION BY appln_id_HTCE ORDER BY COUNT(IPC) DESC) rn
```

```
FROM dbo.two_d
GROUP BY
  appln_id_HTCE,
  IPC,
  earliest_filing_year

  -- Now only use the IPC code that is used most frequent.

WITH cte AS (
SELECT
    appln_id_HTCE
  , IPC
  , earliest_filing_year
  , ROW_NUMBER() OVER (PARTITION BY appln_id_HTCE ORDER BY COUNT(IPC) DESC) rn
FROM dbo.two_d
GROUP BY
  appln_id_HTCE,
  IPC,
  earliest_filing_year)

SELECT
    appln_id_HTCE,
       IPC,
       earliest_filing_year
into two_e
FROM cte WHERE rn = 1

select dbo.one_b.* from dbo.one_b

-- Step 5: Join control patents to HTCE patents.

select dbo.two_e.*, dbo.two_c.appln_id_controls
into two_f
from dbo.two_e
join dbo.two_c on dbo.two_c.earliest_filing_year = dbo.two_e.earliest_filing_year
where dbo.two_c.IPC = dbo.two_e.IPC

-- Step 6: Do a check.
-- Check whether IPC, year are similar for a patent in both the output and in table
control_pool2
-- If they're similar, proceed. Then the matching of the two tables went fine.

select dbo.two_f.* from dbo.two_f

-- Check patents: 6949211 and 241491, IPC: H04W
-- 21422910 and 241492, IPC: H01Q

select dbo.two_c.* from dbo.two_c
where appln_id_controls = 241491

select dbo.two_e.* from dbo.two_e
where appln_id_HTCE = 6949211

select dbo.two_c.* from dbo.two_c
where appln_id_controls = 241492

select dbo.two_e.* from dbo.two_e
where appln_id_HTCE = 21422910

-- Check is OK. Proceed.

-- Step 7: find one control patent per HTCE patent. Select at random.
```

```
WITH CTE AS (
    SELECT
        appln_id_HTCE,
        IPC,
        earliest_filing_year,
        appln_id_controls,
        ROW_NUMBER() OVER(PARTITION BY appln_id_HTCE ORDER BY NEWID()) AS RowNum
    FROM
        dbo.two_f
)

SELECT
    appln_id_HTCE,
    IPC,
    earliest_filing_year,
    appln_id_controls INTO two_g
FROM
    CTE
WHERE
    RowNum = 1

select dbo.two_g.* from dbo.two_g

-- Results: There good, but some have the same control patents. Is that a bad thing?

-- Step 8:

-- It could be the case that some HTCE patents have the same control patents. Check
whether this holds for the data.

select dbo.two_g.appln_id_controls
into two_h
from two_g

with CTE as
(select appln_id_controls,
row_number() over (partition by appln_id_controls
order by appln_id_controls) rn
from dbo.two_h)

delete from CTE where rn > 1

-- Result: 464 rows removed. So, 464 control patents were double.
-- Is this an issue?

select dbo.two_g.* from dbo.two_g
select dbo.two_h.* from dbo.two_h
```

Citing patents

```
-- Step 1: Add pat_publn_id to appln_id of the control patents.

select dbo.two_h.appln_id_controls, patstat2020a.dbo.tls211_pat_publn.pat_publn_id as
pat_publn_id_controls
into two_i
from dbo.two_h
join patstat2020a.dbo.tls211_pat_publn on patstat2020a.dbo.tls211_pat_publn.appln_id =
dbo.two_h.appln_id_controls

-- Step 2: Find the citing patents.
-- Add the pat_publn_id of the citing patents.
```

```
select dbo.two_i.*, patstat2020a.dbo.tls212_citation.pat_publn_id as
pat_publn_id_citing
into two_j
from dbo.two_i
join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.cited_pat_publn_id = dbo.two_i.pat_publn_id_controls

-- Step 3: Add appln_id of the citing patents.

Select dbo.two_j.*, patstat2020a.dbo.tls211_pat_publn.appln_id as appln_id_citing
into two_k
from dbo.two_j
join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.pat_publn_id = dbo.two_j.pat_publn_id_citing

-- Step 4: Delete duplicate rows (if any).

with CTE as
(Select appln_id_controls, pat_publn_id_controls, pat_publn_id_citing,
appln_id_citing,
row_number() over (partition by appln_id_controls, pat_publn_id_controls,
pat_publn_id_citing, appln_id_citing order by appln_id_controls,
pat_publn_id_controls, pat_publn_id_citing, appln_id_citing) rn
from dbo.two_k)

delete from CTE where rn > 1

-- 386 duplicates are removed
-- 23740 rows remain.

-- Step 5: Create a table only with appln_id_controls, appln_id_citing. That's
everything what is needed.

select dbo.two_k.appln_id_controls, dbo.two_k.appln_id_citing
into two_l
from dbo.two_k

-- Step 6: Remove duplicates again.

with CTE as
(select appln_id_controls, appln_id_citing,
row_number() over (partition by appln_id_controls, appln_id_citing
order by appln_id_controls, appln_id_citing) rn
from dbo.two_l)

delete from CTE where rn > 1

-- 4452 rows removed.

select dbo.two_l.* from dbo.two_l

-- 19288 unique citation links remain.

-- step 7: Add filing_date to citing patents.
-- Only include citations between 2003 - 2014.

select dbo.two_l.*, patstat2020a.dbo.tls201_appln.appln_filing_date
into two_m
from dbo.two_l
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
dbo.two_l.appln_id_citing
```

```
where appln_filing_date < '2015-01-01'

-- 13140 citation links remain.
```

Self-citations

```
-- Step 8: Remove self-citations.
-- Add person_id to patents.

select dbo.two_m.appln_id_controls, patstat2020a.dbo.tls207_pers_appln.person_id as
person_id_controls, dbo.two_m.appln_id_citing
into two_n
from dbo.two_m
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.two_m.appln_id_controls

select dbo.two_n.*, patstat2020a.dbo.tls207_pers_appln.person_id as person_id_citing
into two_o
from dbo.two_n
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.two_n.appln_id_citing

-- Step 9: Add inventor sequence number.
-- This number indicates whether a person is an organization yes or not.

select distinct dbo.two_o.person_id_controls,
patstat2020a.dbo.tls227_pers_publn.invt_seq_nr
into two_p
from dbo.two_o
join patstat2020a.dbo.tls227_pers_publn on
patstat2020a.dbo.tls227_pers_publn.person_id = dbo.two_o.person_id_controls

select distinct dbo.two_o.*, dbo.two_p.invt_seq_nr
into two_q
from dbo.two_o
join dbo.two_p on dbo.two_p.person_id_controls = dbo.two_o.person_id_controls

-- Step 10: identify self-citations by organizations.
-- Filter on invt_seq_nr = 0. This are citations by organizations.

select dbo.two_q.*
from dbo.two_q
where ((person_id_controls = person_id_citing) AND (invt_seq_nr = 0))

-- Result: 1027 self-citations

-- Step 11: Create a table of self-citations

select dbo.two_q.*
into two_r
from dbo.two_q
where ((person_id_controls = person_id_citing) AND (invt_seq_nr = 0))

select dbo.two_r.appln_id_controls, dbo.two_r.appln_id_citing
into two_s
from dbo.two_r

with CTE as
(select appln_id_controls, appln_id_citing,
row_number() over (partition by appln_id_controls, appln_id_citing
order by appln_id_controls, appln_id_citing) rn
```

```
from dbo.two_s)

delete from CTE where rn > 1

-- 38 duplicates removed. 989 self-citations remain.

-- Step 12: Delete citing patents from the data.

select dbo.two_m.appln_id_controls, dbo.two_m.appln_id_citing
into two_t
from dbo.two_m

delete two_t
from two_t
inner join two_s
on two_t.appln_id_controls = two_s.appln_id_controls AND
       two_t.appln_id_citing = two_s.appln_id_citing;

-- Result: 1166 self-citations removed.

-- Step 13: check results.

select dbo.two_t.* from dbo.two_t

-- 12151 patents
```

Geo-coordinates

```
-- Step 14: Find the geo-coordinates of the citation links.
-- Link latitude and longitude to appln_id_controls

select dbo.two_t.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_controls, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as
lng_controls
into two_u
from dbo.two_t
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.two_t.appln_id_controls

-- NOTE: one filing can more locations. Therefore, more results.

-- Step 15: Find geocoordinates of the citation links.

select dbo.two_t.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_controls, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as
lng_controls
into two_w
from dbo.two_t
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.two_t.appln_id_controls

select dbo.two_w.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_citing, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as lng_citing
into two_x
from dbo.two_w
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.two_w.appln_id_citing
```

```
-- NOTE: Quite a proportion of the citing patents are not known in the dataset by
Rassenfosse et al. (2019)
-- Therefore, the geo-cordinates of some of the citing patents are unknown.

-- Step 16: Remove duplicates.

with CTE as
(Select appln_id_controls, appln_id_citing, lat_controls, lng_controls, lat_citing,
lng_citing,
row_number() over (partition by appln_id_controls, appln_id_citing, lat_controls,
lng_controls, lat_citing, lng_citing order by appln_id_controls, appln_id_citing,
lat_controls, lng_controls, lat_citing, lng_citing) rn
from dbo.two_x)

delete from CTE where rn > 1

select dbo.two_x.* from dbo.two_x

-- Result: 36.867 rows.
```

Filter on BE/NE

```
-- Step 1:
-- Only citation links that origin from BE/NL. Otherwise the shortest citation link
could be outside of BE/NL, which is no longer representative to the untreated set.

select distinct dbo.two_x.*,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
into two_z
from dbo.two_x
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
(Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.two_x.appln_id_controls AND
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat = dbo.two_x.lat_controls)
where ctry_code = 'NL' OR ctry_code = 'BE'

select two_x.* from two_x
```

Neglected citation links

```
-- Step 17: Isolate appln_id_controls and appln_id_citing

select dbo.two_z.appln_id_controls, dbo.two_z.appln_id_citing
into two_y
from dbo.two_z

-- Step 18: Remove duplicates.

with CTE as
(select appln_id_controls, appln_id_citing,
row_number() over (partition by appln_id_controls, appln_id_citing
order by appln_id_controls, appln_id_citing) rn
from dbo.two_y)

delete from CTE where rn > 1

select dbo.two_y.* from dbo.two_y

-- Step 29: Analysis

-- Only 5783 citations are still present.
```

```
-- This means that for almost 50% of the citation links the geo-coordinates are
missing in the dataset by Rassenfosse et al. (2019)
```

Part II: Variables

All information related to distance.

```
-- Step 1: Save data, including haversine distance, as new table: sample2_1

select sample2.*
into sample2_1
from sample2

-- Step 2: Find the shortest distance per citation link.
-- Since there are multiple inventors, there are multiple distances per citation link.

with CTE as
(Select appln_id_controls, appln_id_citing, Haversine,
row_number() over (partition by appln_id_controls, appln_id_citing order by haversine
asc) rn
from dbo.sample2_1)

delete from CTE where rn > 1

select sample2_1.* from sample2_1

-- Result: 5783 citation links.

-- Step 3: Per control patent, find the shortest citation. This will be used as a
proxy for distance in the model.

Select appln_id_controls, appln_id_citing, Haversine,
row_number() over (partition by appln_id_controls order by haversine asc) rn
into sample2_2
from dbo.sample2_1

select appln_id_controls, appln_id_citing, Haversine as shortest_citation
into sample2_3
from Sample2_2
where rn=1

select sample2_3.* from sample2_3

-- Step 4: Import sample2_4
-- This is the same data as sample2_1, but then the distance is rounded off. Again,
find shortest distance for each citation link.

select sample2_4.* from sample2_4

with CTE as
(Select appln_id_controls, appln_id_citing, Haversine,
row_number() over (partition by appln_id_controls, appln_id_citing order by haversine
asc) rn
from dbo.sample2_4)

delete from CTE where rn > 1

select sample2_4.* from sample2_4

-- Step 5: per control patent, flag quanitiy of citations within a certain range.
-- range 30 km.
```

```
select appln_id_controls, haversine as dist_30
into sample2_dist30_1
from sample2_4
where haversine < 30

select appln_id_controls, COUNT(appln_id_controls) as dist_30
into sample2_dist30
from sample2_dist30_1
group by appln_id_controls

-- 50 km.

select appln_id_controls, haversine as dist_50
into sample2_dist50_1
from sample2_4
where haversine < 50

select appln_id_controls, COUNT(appln_id_controls) as dist_50
into sample2_dist50
from sample2_dist50_1
group by appln_id_controls

-- 100 km

select appln_id_controls, haversine as dist_100
into sample2_dist100_1
from sample2_4
where haversine < 100

select appln_id_controls, COUNT(appln_id_controls) as dist_100
into sample2_dist100
from sample2_dist100_1
group by appln_id_controls

-- 200 km

select appln_id_controls, haversine as dist_200
into sample2_dist200_1
from sample2_4
where haversine < 200

select appln_id_controls, COUNT(appln_id_controls) as dist_200
into sample2_dist200
from sample2_dist200_1
group by appln_id_controls

-- 500 km

select appln_id_controls, haversine as dist_500
into sample2_dist500_1
from sample2_4
where haversine < 500

select appln_id_controls, COUNT(appln_id_controls) as dist_500
into sample2_dist500
from sample2_dist500_1
group by appln_id_controls

-- 1000 km

select appln_id_controls, haversine as dist_1000
```

```
into sample2_dist1000_1
from sample2_4
where haversine < 1000

select appln_id_controls, COUNT(appln_id_controls) as dist_1000
into sample2_dist1000
from sample2_dist1000_1
group by appln_id_controls

-- 2500 km

select appln_id_controls, haversine as dist_2500
into sample2_dist2500_1
from sample2_4
where haversine < 2500

select appln_id_controls, COUNT(appln_id_controls) as dist_2500
into sample2_dist2500
from sample2_dist2500_1
group by appln_id_controls

-- 5000 km

select appln_id_controls, haversine as dist_5000
into sample2_dist5000_1
from sample2_4
where haversine < 5000

select appln_id_controls, COUNT(appln_id_controls) as dist_5000
into sample2_dist5000
from sample2_dist5000_1
group by appln_id_controls

-- 0 km

select appln_id_controls, haversine as dist_0
into sample2_dist0_1
from sample2_4
where haversine = 0

select appln_id_controls, COUNT(appln_id_controls) as dist_0
into sample2_dist0
from sample2_dist0_1
group by appln_id_controls

-- Step 6: Calculate average distance of all citation links per control patent.

select appln_id_controls, avg(haversine) as avg_haversine, count(appln_id_controls) as
nb_citations
into sample2_5
from Sample2_4
group by appln_id_controls

-- Step 7: Put all distance related information into one table.

select sample2_5.*, sample2_3.shortest_citation, sample2_dist0.dist_0,
sample2_dist30.dist_30, sample2_dist50.dist_50, sample2_dist100.dist_100,
sample2_dist200.dist_200, sample2_dist500.dist_500, sample2_dist1000.dist_1000,
sample2_dist2500.dist_2500, sample2_dist5000.dist_5000
into sample2_6
from sample2_5
join sample2_3 on sample2_3.appln_id_controls = sample2_5.appln_id_controls
```

```
LEFT join sample2_dist0 on sample2_dist0.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist30 on sample2_dist30.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist50 on sample2_dist50.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist100 on sample2_dist100.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist200 on sample2_dist200.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist500 on sample2_dist500.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist1000 on sample2_dist1000.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist2500 on sample2_dist2500.appln_id_controls =
sample2_5.appln_id_controls
left join  sample2_dist5000 on sample2_dist5000.appln_id_controls =
sample2_5.appln_id_controls

select sample2_6.* from sample2_6

-- Result: 973 unique control patents.
```

Control variables

```
-- Step 1: Find number of forward -and backward citations
-- Step 1a: Forward citations

select dbo.two_l.* from dbo.two_l
-- This table includes all forward citations of the control patents.
-- So, not only the citations before 2014.
-- Just everything.

select appln_id_controls, COUNT(appln_id_controls) as nb_forward_citations
into sample2_7_1
from two_l
group by appln_id_controls

-- Step 1b: Backward citations

select sample2_6.appln_id_controls, patstat2020a.dbo.tls211_pat_publn.pat_publn_id
into sample2_7_2
from sample2_6
join  patstat2020a.dbo.tls211_pat_publn on  patstat2020a.dbo.tls211_pat_publn.appln_id
= sample2_6.appln_id_controls

select sample2_7_2.appln_id_controls, sample2_7_2.pat_publn_id,
patstat2020a.dbo.tls212_citation.cited_pat_publn_id
into sample2_7_3
from sample2_7_2
join patstat2020a.dbo.tls212_citation on patstat2020a.dbo.tls212_citation.pat_publn_id
= sample2_7_2.pat_publn_id
where cited_pat_publn_id != 0

select appln_id_controls, COUNT(appln_id_controls) as nb_backward_citations
into sample2_7_4
from sample2_7_3
group by appln_id_controls

select sample2_6.appln_id_controls, sample2_7_1.nb_forward_citations,
sample2_7_4.nb_backward_citations
```

```
into sample2_7
from sample2_6
left join sample2_7_1 on sample2_7_1.appln_id_controls = sample2_6.appln_id_controls
left join sample2_7_4 on sample2_7_4.appln_id_controls = sample2_6.appln_id_controls

select sample2_7.* from sample2_7

-- Step 2: IPC and year

select dbo.two_c.* from dbo.two_c

-- most frequent IPC code and year can be found in 'two_c'.

-- Find number of IPC codes per control patent.

select sample2_6.appln_id_controls, patstat2020a.dbo.tls209_appln_ipc.ipc_class_symbol
into sample2_8_1
from sample2_6
left join patstat2020a.dbo.tls209_appln_ipc on
patstat2020a.dbo.tls209_appln_ipc.appln_id = sample2_6.appln_id_controls

select appln_id_controls, COUNT(appln_id_controls) as nb_IPC_codes
into sample2_8
from sample2_8_1
group by appln_id_controls

-- Step 3: nb_claims. Select patent with highest claims if patent is similar.

select sample2_6.appln_id_controls, patstat2020a.dbo.tls211_pat_publn.publn_claims
into sample2_9_1
from sample2_6
left join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.appln_id = sample2_6.appln_id_controls

with CTE as
(Select appln_id_controls, publn_claims,
row_number() over (partition by appln_id_controls order by publn_claims desc) rn
from dbo.sample2_9_1)

delete from CTE where rn > 1

select sample2_9_1.* from sample2_9_1

-- Step 4: Backward citations to literature

select sample2_7_2.* from sample2_7_2

select sample2_7_2.*, patstat2020a.dbo.tls212_citation.cited_npl_publn_id
into sample2_10_1
from sample2_7_2
left join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.pat_publn_id = sample2_7_2.pat_publn_id
where cited_npl_publn_id != '0'

select appln_id_controls, COUNT(appln_id_controls) as nb_cited_literature
into sample2_10
from sample2_10_1
group by appln_id_controls

-- Step 5: Find company names
-- In case of multiple companies per one HTCE patent, choose one company name at
random per one HTCE patent.
```

```
select sample2_6.appln_id_controls, patstat2020a.dbo.tls207_pers_appln.person_id
into sample2_11_1
from sample2_6
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= sample2_6.appln_id_controls

select sample2_11_1.appln_id_controls, patstat2020a.dbo.tls206_person.person_name,
patstat2020a.dbo.tls206_person.psn_level, patstat2020a.dbo.tls206_person.psn_sector
into sample2_11_2
from sample2_11_1
join patstat2020a.dbo.tls206_person on patstat2020a.dbo.tls206_person.person_id =
sample2_11_1.person_id
where psn_sector = 'company'

select sample2_11_2.* from sample2_11_2

WITH CTE AS (
    SELECT
        appln_id_controls,
        person_name,
        ROW_NUMBER() OVER(PARTITION BY appln_id_controls ORDER BY NEWID()) AS RowNum
    FROM
        dbo.sample2_11_2)

SELECT
    appln_id_controls,
    person_name INTO sample2_11
FROM
    CTE
WHERE
    RowNum = 1

select sample2_11.* from sample2_11

-- Step 6: Number of foreign inventors

select sample2_6.appln_id_controls,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
into sample2_12_3
from sample2_6
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
sample2_6.appln_id_controls
where ctry_code != 'NL' and ctry_code != 'BE'

select sample2_12_3.appln_id_controls, count(appln_id_controls) as
nb_foreign_inventors
into sample2_12
from sample2_12_3
group by appln_id_controls

-- Step 7: Add nb_applicants, only companies

select sample2_11_2.* from sample2_11_2

select sample2_11_2.appln_id_controls, count(appln_id_controls) as
nb_applicants_companies
into sample2_13
from sample2_11_2
```

```
group by appln_id_controls

-- Step 8: Add a variable that indicates how much HTCE patents have the same control
patent.
-- So, a weight can be included on those patents.

select appln_id_controls, count(appln_id_controls) as frq_controls
into sample2_14
from two_g
group by appln_id_controls

-- Step 9: Put everything into one table.

select sample2_6.*, patstat2020a.dbo.tls201_appln.appln_auth,
patstat2020a.dbo.tls201_appln.nb_inventors, sample2_13.nb_applicants_companies,
sample2_11.person_name,
sample2_12.nb_foreign_inventors, two_c.IPC, sample2_8.nb_IPC_codes,
two_c.earliest_filing_year, sample2_7.nb_backward_citations,
sample2_7.nb_forward_citations,
sample2_10.nb_cited_literature, sample2_9_1.publn_claims, sample2_14.frq_controls
into sample2_15
from sample2_6
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
sample2_6.appln_id_controls
LEFT join sample2_13 on sample2_13.appln_id_controls = sample2_6.appln_id_controls
left join sample2_11 on sample2_11.appln_id_controls = sample2_6.appln_id_controls
left join sample2_12 on sample2_12.appln_id_controls = sample2_6.appln_id_controls
left join two_c on two_c.appln_id_controls = sample2_6.appln_id_controls
left join sample2_8 on sample2_8.appln_id_controls = sample2_6.appln_id_controls
left join sample2_7 on sample2_7.appln_id_controls = sample2_6.appln_id_controls
left join sample2_10 on sample2_10.appln_id_controls = sample2_6.appln_id_controls
left join sample2_9_1 on sample2_9_1.appln_id_controls = sample2_6.appln_id_controls
left join sample2_14 on sample2_14.appln_id_controls = sample2_6.appln_id_controls
```

Match sample1 & sample2 on IPC and year, find number of matches

```
-- Step 1: Find all possible control, from controls dataset, patents per HTCE patent.

select sample1_14.appln_id_HTCE, sample1_14.IPC, sample1_14.earliest_filing_year,
sample2_15.appln_id_controls
into joint1_1
from sample1_14
join sample2_15 on sample2_15.IPC = sample1_14.IPC
where sample2_15.earliest_filing_year = sample1_14.earliest_filing_year

-- Step 2: For each HTCE, find possible number of control patents.

select joint1_1.* from joint1_1

select appln_id_HTCE, COUNT(appln_id_HTCE) as nb_matching_control_patents
into joint1_2
from joint1_1
group by appln_id_HTCE

-- Result: Out of the 955, 769 HTCE patents do have a matching control patent.

-- Step 3: For each control patent, find the number of matches to the HTCE patents.

select appln_id_controls, COUNT(appln_id_controls) as nb_matching_HTCE_patents
into joint1_3
from joint1_1
```

```
group by appln_id_controls

-- Result: 857 control patents have a match to a HTCE patent, this number is higher
than the HTCE patents.
-- So, there are more control patents than treated patents in the dataset.

-- Step 4: Join 'nb_matching_control_patents' and 'nb_matching_HTCE_patents', to final
tables.

select sample1_14.*, joint1_2.nb_matching_control_patents
into sample1_15
from sample1_14
join joint1_2 on joint1_2.appln_id_HTCE = sample1_14.appln_id_HTCE

select sample2_15.*, joint1_3.nb_matching_HTCE_patents
into sample2_16
from sample2_15
join joint1_3 on joint1_3.appln_id_controls = sample2_15.appln_id_controls

-- Step 5: Check tables

select sample1_15.* from sample1_15
select sample2_16.* from sample2_16
```

Assess weights to control patents

```
-- For each couple of IPC and year, find the number of matching HTCE -and control
patents.
-- In that way, a weight can be assessed to the controls, since they're more control
patents in the dataset.

-- Step 1: For each couple of IPC and year, find the number of matching HTCE patents.

select appln_id_HTCE, IPC, earliest_filing_year
into weight1_1
from sample1_15

select IPC, earliest_filing_year, COUNT(appln_id_HTCE) as fq_combination_HTCE
into weight1_2
from weight1_1
group by IPC, earliest_filing_year

-- Step 2: For each couple of IPC and year, find the number of matching control
patents.

select IPC, earliest_filing_year, COUNT(appln_id_controls) as fq_combination_controls
into weight1_3
from sample2_16
group by IPC, earliest_filing_year

-- Step 3: Combine values into one table

select weight1_2.*, weight1_3.fq_combination_controls
into weight1_4
from weight1_2
join weight1_3 on weight1_3.IPC = weight1_2.IPC
where weight1_3.earliest_filing_year = weight1_2.earliest_filing_year

select weight1_4.* from weight1_4

-- Step 4: Export 'weight1_4' into excel.
```

```
-- Divide both frequencies with one another.

-- Import new data as table 'weight1_5'

select weight1_5.* from weight1_5

-- Step 5: connect variable 'weight_control' to sample2_16.
-- The treated patents get value one.

select sample2_16.*, weight1_5.weight_control
from sample2_16
join weight1_5 on weight1_5.IPC = sample2_16.IPC
where weight1_5.earliest_filing_year = sample2_16.earliest_filing_year
```

Sample 3: Brainport patents

Part I: Collection of the Brainport patents and the citing patents.

***The query is exactly the same as in Sample 1. Only the coordinates are different. ***

Brainport patents

```
-- Step 1: Find brainport patents.

select Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.*
into three_a
from Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean

where ((lat < 51.541469 AND lat > 51.318343) AND (lng < 5.875337 AND lng > 5.231379))
AND filing_date > '2003-01-01'
AND filing_date < '2011-01-01'


-- Step 2: Isolate appln_id of Brainport patents.
-- Remove duplicates

select dbo.three_a.appln_id as appln_id_brainport
into three_b
from dbo.three_a

with CTE as
(Select appln_id_brainport,
row_number() over (partition by appln_id_brainport order by appln_id_brainport) rn
from dbo.three_b)

delete from CTE where rn > 1

select dbo.three_b.* from dbo.three_b

-- Results: 7885 patents

-- Step 3: Add pat_publn_id to appln_id of brainport patents.

select dbo.three_b.appln_id_brainport, patstat2020a.dbo.tls211_pat_publn.pat_publn_id
as pat_publn_id_brainport
into three_c
from dbo.three_b
join patstat2020a.dbo.tls211_pat_publn on patstat2020a.dbo.tls211_pat_publn.appln_id =
dbo.three_b.appln_id_brainport
```

Citing patents

```
-- Step 4: Find the citing patents.
-- Add the pat_publn_id of the citing patents.

select dbo.three_c.*, patstat2020a.dbo.tls212_citation.pat_publn_id as
pat_publn_id_citing
into three_d
from dbo.three_c
join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.cited_pat_publn_id =
dbo.three_c.pat_publn_id_brainport

-- Step 5: Add appln_id of the citing patents.

Select dbo.three_d.*, patstat2020a.dbo.tls211_pat_publn.appln_id as appln_id_citing
into three_e
from dbo.three_d
join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.pat_publn_id = dbo.three_d.pat_publn_id_citing
```

```
-- Step 6: Delete duplicate rows (if any).

with CTE as
(Select appln_id_brainport, pat_publn_id_brainport, pat_publn_id_citing,
appln_id_citing,
row_number() over (partition by appln_id_brainport, pat_publn_id_brainport,
pat_publn_id_citing, appln_id_citing order by appln_id_brainport,
pat_publn_id_brainport, pat_publn_id_citing, appln_id_citing) rn
from dbo.three_e)

delete from CTE where rn > 1

-- 1544 duplicates are removed

select dbo.three_e.* from dbo.three_e

-- Step 7: Create a table only with appln_id_brainport, appln_id_citing. That's
everything what is needed.

select dbo.three_e.appln_id_brainport, dbo.three_e.appln_id_citing
into three_f
from dbo.three_e

-- Step 8: Remove duplicates again.

with CTE as
(select appln_id_brainport, appln_id_citing,
row_number() over (partition by appln_id_brainport, appln_id_citing
order by appln_id_brainport, appln_id_citing) rn
from dbo.three_f)

delete from CTE where rn > 1

-- 25901 rows removed.

select dbo.three_f.* from dbo.three_f

-- 85583 unique citation links remain.

-- step 9: Add filing_date to citing patents.
-- Only include citations between 2003 - 2014.

select dbo.three_f.*, patstat2020a.dbo.tls201_appln.appln_filing_date
into three_g
from dbo.three_f
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
dbo.three_f.appln_id_citing
where appln_filing_date < '2015-01-01'

-- 62772 rows remain.
```

Self-citations

```
-- Step 10: Remove self-citations.
-- Add person_id to patents.
```

```
select dbo.three_g.appln_id_brainport, patstat2020a.dbo.tls207_pers_appln.person_id as
person_id_brainport, dbo.three_g.appln_id_citing
into three_h
from dbo.three_g
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.three_g.appln_id_brainport

select dbo.three_h.*, patstat2020a.dbo.tls207_pers_appln.person_id as person_id_citing
into three_i
from dbo.three_h
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.three_h.appln_id_citing

-- Step 11: Add inventor sequence number.
-- This number indicates whether a person is an organization yes or not.

select distinct dbo.three_i.person_id_brainport,
patstat2020a.dbo.tls227_pers_publn.invt_seq_nr
into three_j
from dbo.three_i
join patstat2020a.dbo.tls227_pers_publn on
patstat2020a.dbo.tls227_pers_publn.person_id = dbo.three_i.person_id_brainport

select distinct dbo.three_i.*, dbo.three_j.invt_seq_nr
into three_k
from dbo.three_i
join dbo.three_j on dbo.three_j.person_id_brainport = dbo.three_i.person_id_brainport

-- Step 12: identify self-citations by organizations.
-- Filter on invt_seq_nr = 0. This are citations by organizations.

select dbo.three_k.*
from dbo.three_k
where ((person_id_brainport = person_id_citing) AND (invt_seq_nr = 0))

-- Result: 12180 self-citations.

-- Step 13: Create a table of the self-citations.

select dbo.three_k.*
into three_l
from dbo.three_k
where ((person_id_brainport = person_id_citing) AND (invt_seq_nr = 0))

select dbo.three_l.appln_id_brainport, dbo.three_l.appln_id_citing
into three_m
from dbo.three_l

with CTE as
(select appln_id_brainport, appln_id_citing,
row_number() over (partition by appln_id_brainport, appln_id_citing
order by appln_id_brainport, appln_id_citing) rn
from dbo.three_m)

delete from CTE where rn > 1

-- 110 duplicates removed.

-- Step 14: Delete citing patents from the data.

select dbo.three_g.appln_id_brainport, dbo.three_g.appln_id_citing
into three_n
```

```
from dbo.three_g

delete three_n
from three_n
inner join three_m
on three_n.appln_id_brainport = three_m.appln_id_brainport AND
       three_n.appln_id_citing = three_m.appln_id_citing;

-- Result: 12070 self-citations removed.

-- Step 15: check results.

select dbo.three_n.* from dbo.three_n

-- 50702 patents.
```

Geo-coordinates

```
-- Step 16: Find the geo-coordinates of the citation links.
-- Link latitude and longitude to appln_id_brainport

select dbo.three_n.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_brainport, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as
lng_brainport
into three_o
from dbo.three_n
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.three_n.appln_id_brainport

-- NOTE: one filing can more locations. Therefore, more results.

-- Step 17: Filter out the non-HTCE locations.

select dbo.three_o.*
into three_p
from dbo.three_o
where ((lat_brainport < 51.541469 AND lat_brainport > 51.318343) AND (lng_brainport <
5.875337 AND lng_brainport > 5.231379))

-- Step 18: Find geo-coordinates of the citing patents.

select dbo.three_p.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_citing, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as lng_citing
into three_q
from dbo.three_p
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.three_p.appln_id_citing

-- NOTE: Quite a proportion of the citing patents are not known in the dataset by
Rassenfosse et al. (2019)
-- Therefore, the geo-cordinates of some of the citing patents are unknown.

-- Step 19: Remove duplicates.

with CTE as
(Select appln_id_brainport, appln_id_citing, lat_brainport, lng_brainport, lat_citing,
lng_citing,
```

```
row_number() over (partition by appln_id_brainport, appln_id_citing, lat_brainport,
lng_brainport, lat_citing, lng_citing order by appln_id_brainport, appln_id_citing,
lat_brainport, lng_brainport, lat_citing, lng_citing) rn
from dbo.three_q)

delete from CTE where rn > 1

select dbo.three_q.* from dbo.three_q

-- Result: 46468 rows.
```

Quantity of neglected citations

```
-- Step 20: Isolate appln_id_HTCE and appln_id_citing

select dbo.three_q.appln_id_brainport, dbo.three_q.appln_id_citing
into three_r
from dbo.three_q

-- Step 21: Remove duplicates.

with CTE as
(select appln_id_brainport, appln_id_citing,
row_number() over (partition by appln_id_brainport, appln_id_citing
order by appln_id_brainport, appln_id_citing) rn
from dbo.three_r)

delete from CTE where rn > 1

select dbo.three_r.* from dbo.three_r

-- Step 22: Analysis

-- Only 16578 citations are still present.
-- This means that for almost 70% of the citation links the geo-coordinates are
missing in the dataset by Rassenfosse et al. (2019).
-- Also, the dataset by Rassenfosse et al. probably has a small proportion of the HTCE
patents present.
```

Part II: Variables

All information related to distance.

```
-- Step 1: Save data, including haversine distance, as new table: sample3_1

select sample3.*
into sample3_1
from sample3

-- Step 2: Find the shortest distance per citation link.
-- Since there are multiple inventors, there are multiple distances per citation link.

with CTE as
(Select appln_id_brainport, appln_id_citing, Haversine,
row_number() over (partition by appln_id_brainport, appln_id_citing order by haversine
asc) rn
from dbo.sample3_1)

delete from CTE where rn > 1
```

```
select sample3_1.* from sample3_1

-- Step 3: Per brainport patent, find the shortest citation. This will be used as a
proxy for distance in the model.

Select appln_id_brainport, appln_id_citing, Haversine,
row_number() over (partition by appln_id_brainport order by haversine asc) rn
into sample3_2
from dbo.sample3_1

select appln_id_brainport, appln_id_citing, Haversine as shortest_citation
into sample3_3
from Sample3_2
where rn=1

-- Step 4: Import sample1_4
-- This is the same data as sample3, but then the distance is rounded off.
-- Find shortest distance per citation link.

select sample3_4.* from sample3_4

with CTE as
(Select appln_id_brainport, appln_id_citing, Haversine,
row_number() over (partition by appln_id_brainport, appln_id_citing order by haversine
asc) rn
from dbo.sample3_4)

delete from CTE where rn > 1

select sample3_4.* from sample3_4

-- Step 5: Per Brainport patent, flag quanitiy of citations within a certain range.
-- range 30 km.

select appln_id_brainport, haversine as dist_30
into sample3_dist30_1
from sample3_4
where haversine < 30

select appln_id_brainport, COUNT(appln_id_brainport) as dist_30
into sample3_dist30
from sample3_dist30_1
group by appln_id_brainport

-- 50 km.

select appln_id_brainport, haversine as dist_50
into sample3_dist50_1
from sample3_4
where haversine < 50

select appln_id_brainport, COUNT(appln_id_brainport) as dist_50
into sample3_dist50
from sample3_dist50_1
group by appln_id_brainport

-- 100 km

select appln_id_brainport, haversine as dist_100
into sample3_dist100_1
from sample3_4
where haversine < 100
```

```
select appln_id_brainport, COUNT(appln_id_brainport) as dist_100
into sample3_dist100
from sample3_dist100_1
group by appln_id_brainport

-- 200 km

select appln_id_brainport, haversine as dist_200
into sample3_dist200_1
from sample3_4
where haversine < 200

select appln_id_brainport, COUNT(appln_id_brainport) as dist_200
into sample3_dist200
from sample3_dist200_1
group by appln_id_brainport

-- 500 km

select appln_id_brainport, haversine as dist_500
into sample3_dist500_1
from sample3_4
where haversine < 500

select appln_id_brainport, COUNT(appln_id_brainport) as dist_500
into sample3_dist500
from sample3_dist500_1
group by appln_id_brainport

-- 1000 km

select appln_id_brainport, haversine as dist_1000
into sample3_dist1000_1
from sample3_4
where haversine < 1000

select appln_id_brainport, COUNT(appln_id_brainport) as dist_1000
into sample3_dist1000
from sample3_dist1000_1
group by appln_id_brainport

-- 2500 km

select appln_id_brainport, haversine as dist_2500
into sample3_dist2500_1
from sample3_4
where haversine < 2500

select appln_id_brainport, COUNT(appln_id_brainport) as dist_2500
into sample3_dist2500
from sample3_dist2500_1
group by appln_id_brainport

-- 5000 km

select appln_id_brainport, haversine as dist_5000
into sample3_dist5000_1
from sample3_4
where haversine < 5000

select appln_id_brainport, COUNT(appln_id_brainport) as dist_5000
```

```
into sample3_dist5000
from sample3_dist5000_1
group by appln_id_brainport

-- 0 km

select appln_id_brainport, haversine as dist_0
into sample3_dist0_1
from sample3_4
where haversine = 0

select appln_id_brainport, COUNT(appln_id_brainport) as dist_0
into sample3_dist0
from sample3_dist0_1
group by appln_id_brainport

-- Step 6: Calculate average distance of all citation links per brainport patent.

select appln_id_brainport, avg(haversine) as avg_haversine, count(appln_id_brainport)
as nb_citations
into sample3_5
from Sample3_4
group by appln_id_brainport

-- Step 7: Put all distance related information into one table.

select sample3_5.*, sample3_3.shortest_citation, sample3_dist0.dist_0,
sample3_dist30.dist_30, sample3_dist50.dist_50, sample3_dist100.dist_100,
sample3_dist200.dist_200, sample3_dist500.dist_500, sample3_dist1000.dist_1000,
sample3_dist2500.dist_2500, sample3_dist5000.dist_5000
into sample3_6
from sample3_5
join sample3_3 on sample3_3.appln_id_brainport = sample3_5.appln_id_brainport
LEFT join sample3_dist0 on sample3_dist0.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist30 on sample3_dist30.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist50 on sample3_dist50.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist100 on sample3_dist100.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist200 on sample3_dist200.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist500 on sample3_dist500.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist1000 on sample3_dist1000.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist2500 on sample3_dist2500.appln_id_brainport =
sample3_5.appln_id_brainport
left join  sample3_dist5000 on sample3_dist5000.appln_id_brainport =
sample3_5.appln_id_brainport

select sample3_6.* from sample3_6
```

Control variables

```
-- Step 1: Find number of forward -and backward citations
-- Step 1a: Forward citations, all time

select three_f.* from three_f
```

```
select appln_id_brainport, COUNT(appln_id_brainport) as nb_forward_citations
into sample3_7_1
from three_f
group by appln_id_brainport

-- Step 1b: Backward citations

select sample3_6.appln_id_brainport, patstat2020a.dbo.tls211_pat_publn.pat_publn_id
into sample3_7_2
from sample3_6
join  patstat2020a.dbo.tls211_pat_publn on  patstat2020a.dbo.tls211_pat_publn.appln_id
= sample3_6.appln_id_brainport

select sample3_7_2.appln_id_brainport, sample3_7_2.pat_publn_id,
patstat2020a.dbo.tls212_citation.cited_pat_publn_id
into sample3_7_3
from sample3_7_2
join patstat2020a.dbo.tls212_citation on patstat2020a.dbo.tls212_citation.pat_publn_id
= sample3_7_2.pat_publn_id
where cited_pat_publn_id != 0

select appln_id_brainport, COUNT(appln_id_brainport) as nb_backward_citations
into sample3_7_4
from sample3_7_3
group by appln_id_brainport

select sample3_6.appln_id_brainport, sample3_7_1.nb_forward_citations,
sample3_7_4.nb_backward_citations
into sample3_7
from sample3_6
left join sample3_7_1 on sample3_7_1.appln_id_brainport = sample3_6.appln_id_brainport
left join sample3_7_4 on sample3_7_4.appln_id_brainport = sample3_6.appln_id_brainport

-- Step 2: IPC and year

-- most frequent IPC code and year can be found in "four_e"

select four_e.* from four_e

-- Find number of IPC codes per brainport patent.

select sample3_6.appln_id_brainport,
patstat2020a.dbo.tls209_appln_ipc.ipc_class_symbol
into sample3_8_1
from sample3_6
left join patstat2020a.dbo.tls209_appln_ipc on
patstat2020a.dbo.tls209_appln_ipc.appln_id = sample3_6.appln_id_brainport

select appln_id_brainport, COUNT(appln_id_brainport) as nb_IPC_codes
into sample3_8
from sample3_8_1
group by appln_id_brainport

-- Step 3: nb_claims. Select patent with highest claims if patent is similar.

select sample3_6.appln_id_brainport, patstat2020a.dbo.tls211_pat_publn.publn_claims
into sample3_9_1
from sample3_6
left join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.appln_id = sample3_6.appln_id_brainport

with CTE as
```

```
(Select appln_id_brainport, publn_claims,
row_number() over (partition by appln_id_brainport order by publn_claims desc) rn
from dbo.sample3_9_1)

delete from CTE where rn > 1

select sample3_9_1.* from sample3_9_1

-- Step 4: Backward citations to literature

select sample3_7_2.* from sample3_7_2

select sample3_7_2.*, patstat2020a.dbo.tls212_citation.cited_npl_publn_id
into sample3_10_1
from sample3_7_2
left join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.pat_publn_id = sample3_7_2.pat_publn_id
where cited_npl_publn_id != '0'

select appln_id_brainport, COUNT(appln_id_brainport) as nb_cited_literature
into sample3_10
from sample3_10_1
group by appln_id_brainport

-- Step 5: Find company names
-- In case of multiple companies per one brainport patent, choose one company name at
random per one brainport patent.

select sample3_6.appln_id_brainport, patstat2020a.dbo.tls207_pers_appln.person_id
into sample3_11_1
from sample3_6
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= sample3_6.appln_id_brainport

select sample3_11_1.appln_id_brainport, patstat2020a.dbo.tls206_person.person_name,
patstat2020a.dbo.tls206_person.psn_level, patstat2020a.dbo.tls206_person.psn_sector
into sample3_11_2
from sample3_11_1
join patstat2020a.dbo.tls206_person on patstat2020a.dbo.tls206_person.person_id =
sample3_11_1.person_id
where psn_sector = 'company'

select sample3_11_2.* from sample3_11_2

WITH CTE AS (
    SELECT
        appln_id_brainport,
        person_name,
        ROW_NUMBER() OVER(PARTITION BY appln_id_brainport ORDER BY NEWID()) AS RowNum
    FROM
        dbo.sample3_11_2)

SELECT
    appln_id_brainport,
    person_name INTO sample3_11
FROM
    CTE
WHERE
    RowNum = 1

select sample3_11.* from sample3_11
```

```
-- Step 6: Find quantity of foreign inventors/locations per brainport patent.
-- Quantity of locations/inventors outside brainport

select sample3_6.appln_id_brainport,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
into sample3_12_1
from sample3_6
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
sample3_6.appln_id_brainport
where not ((lat < 51.541469 AND lat > 51.318343) AND (lng < 5.875337 AND lng >
5.231379))

select sample3_12_1.appln_id_brainport, count(appln_id_brainport) as
nb_outside_inventors
into sample3_12_2
from sample3_12_1
group by appln_id_brainport

-- Number of foreign inventors

select sample3_6.appln_id_brainport,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
into sample3_12_3
from sample3_6
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
sample3_6.appln_id_brainport
where ctry_code != 'NL' and ctry_code != 'BE'

select sample3_12_3.appln_id_brainport, count(appln_id_brainport) as
nb_foreign_inventors
into sample3_12_4
from sample3_12_3
group by appln_id_brainport

-- Create overall table of inventors

select sample3_6.appln_id_brainport, sample3_12_2.nb_outside_inventors,
sample3_12_4.nb_foreign_inventors
into sample3_12
from sample3_6
left join sample3_12_2 on sample3_12_2.appln_id_brainport =
sample3_6.appln_id_brainport
left join sample3_12_4 on sample3_12_4.appln_id_brainport =
sample3_6.appln_id_brainport

-- Step 7: Add nb_applicants, only companies

select sample3_11_2.* from sample3_11_2

select sample3_11_2.appln_id_brainport, count(appln_id_brainport) as
nb_applicants_companies
into sample3_13
from sample3_11_2
group by appln_id_brainport

-- Step 8: Put everything into one table.
```

```
select sample3_6.*, patstat2020a.dbo.tls201_appln.appln_auth,
patstat2020a.dbo.tls201_appln.nb_inventors, sample3_13.nb_applicants_companies,
sample3_11.person_name, sample3_12.nb_foreign_inventors,
sample3_12.nb_outside_inventors, four_e.IPC, sample3_8.nb_IPC_codes,
four_e.earliest_filing_year, sample3_7.nb_backward_citations,
sample3_7.nb_forward_citations,
sample3_10.nb_cited_literature, sample3_9_1.publn_claims
into sample3_14
from sample3_6
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
sample3_6.appln_id_brainport
LEFT join sample3_13 on sample3_13.appln_id_brainport = sample3_6.appln_id_brainport
left join sample3_11 on sample3_11.appln_id_brainport = sample3_6.appln_id_brainport
left join sample3_12 on sample3_12.appln_id_brainport = sample3_6.appln_id_brainport
left join four_e on four_e.appln_id_brainport = sample3_6.appln_id_brainport
left join sample3_8 on sample3_8.appln_id_brainport = sample3_6.appln_id_brainport
left join sample3_7 on sample3_7.appln_id_brainport = sample3_6.appln_id_brainport
left join sample3_10 on sample3_10.appln_id_brainport = sample3_6.appln_id_brainport
left join sample3_9_1 on sample3_9_1.appln_id_brainport = sample3_6.appln_id_brainport

select sample3_14.* from sample3_14
```

Sample 4: Control patents of the Brainport patents.

Part I: Collection of control patents and citing patents.

Match HTCE patents with control patents at random

```
-- For the pool of control patents: use 'dbo.two_c' from Sample 2.
-- This are the possible control patents from NL and BE, except brainport region.

-- Step 1: Add IPC code, year to brainport patents

select dbo.three_b.*, left(patstat2020a.dbo.tls209_appln_ipc.ipc_class_symbol, 4) as
IPC, patstat2020a.dbo.tls201_appln.earliest_filing_year
into four_d
from dbo.three_b
join patstat2020a.dbo.tls209_appln_ipc on patstat2020a.dbo.tls209_appln_ipc.appln_id =
dbo.three_b.appln_id_brainport
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
dbo.three_b.appln_id_brainport

-- Each patent has multiple IPC codes. Therefore, create a query that finds the most
frequent IPC code.
-- That IPC will be used to find control patent.

SELECT
    appln_id_brainport
  , IPC
  , earliest_filing_year
  , ROW_NUMBER() OVER (PARTITION BY appln_id_brainport ORDER BY COUNT(IPC) DESC) rn
FROM dbo.four_d
GROUP BY
  appln_id_brainport,
  IPC,
  earliest_filing_year

  -- Now only use the IPC code that is used most frequent.

WITH cte AS (
SELECT
    appln_id_brainport
  , IPC
  , earliest_filing_year
  , ROW_NUMBER() OVER (PARTITION BY appln_id_brainport ORDER BY COUNT(IPC) DESC) rn
FROM dbo.four_d
GROUP BY
  appln_id_brainport,
  IPC,
  earliest_filing_year)

SELECT
    appln_id_brainport,
       IPC,
       earliest_filing_year
into four_e
FROM cte WHERE rn = 1

select dbo.three_b.* from dbo.three_b

-- Step 5: Join control patents to brainport patents.

select dbo.four_e.*, dbo.two_c.appln_id_controls
into four_f
from dbo.four_e
join dbo.two_c on dbo.two_c.earliest_filing_year = dbo.four_e.earliest_filing_year
where dbo.two_c.IPC = dbo.four_e.IPC
```

```
-- Step 6: Do a check.
-- Check whether IPC, year are similar for a patent in both the output and in table
control_pool2
-- If they're similar, proceed. Then the matching of the two tables went fine.

select dbo.four_f.* from dbo.four_f

-- Check patents: 8931 and 2720, IPC: H01L
-- 27356 and 2690, IPC: H01L

select dbo.two_c.* from dbo.two_c
where appln_id_controls = 2720

select dbo.four_e.* from dbo.four_e
where appln_id_brainport = 8931

select dbo.two_c.* from dbo.two_c
where appln_id_controls = 2690

select dbo.four_e.* from dbo.four_e
where appln_id_brainport = 27356

-- Check is OK. Proceed.

-- Step 7: find one control patent per HTCE patent. Select at random.

WITH CTE AS (
    SELECT
        appln_id_brainport,
        IPC,
        earliest_filing_year,
        appln_id_controls,
        ROW_NUMBER() OVER(PARTITION BY appln_id_brainport ORDER BY NEWID()) AS RowNum
    FROM
        dbo.four_f
)

SELECT
    appln_id_brainport,
    IPC,
    earliest_filing_year,
    appln_id_controls INTO four_g
FROM
    CTE
WHERE
    RowNum = 1

select dbo.four_g.* from dbo.four_g

-- Results: There good, but some have the same control patents. Is that a bad thing?

-- Step 8:

-- It could be the case that some HTCE patents have the same control patents. Check
whether this holds for the data.

select dbo.four_g.appln_id_controls
into four_h
from four_g

with CTE as
(select appln_id_controls,
```

```
row_number() over (partition by appln_id_controls
order by appln_id_controls) rn
from dbo.four_h)

delete from CTE where rn > 1

-- Result: 2766 rows removed. So, 2766 control patents were double.
-- Is this an issue?

select dbo.four_g.* from dbo.four_g
select dbo.four_h.* from dbo.four_h
```

Citing patents

```
-- Step 1: Add pat_publn_id to appln_id of the control patents.

select dbo.four_h.appln_id_controls, patstat2020a.dbo.tls211_pat_publn.pat_publn_id as
pat_publn_id_controls
into four_i
from dbo.four_h
join patstat2020a.dbo.tls211_pat_publn on patstat2020a.dbo.tls211_pat_publn.appln_id =
dbo.four_h.appln_id_controls

-- Step 2: Find the citing patents.
-- Add the pat_publn_id of the citing patents.

select dbo.four_i.*, patstat2020a.dbo.tls212_citation.pat_publn_id as
pat_publn_id_citing
into four_j
from dbo.four_i
join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.cited_pat_publn_id = dbo.four_i.pat_publn_id_controls

-- Step 3: Add appln_id of the citing patents.

Select dbo.four_j.*, patstat2020a.dbo.tls211_pat_publn.appln_id as appln_id_citing
into four_k
from dbo.four_j
join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.pat_publn_id = dbo.four_j.pat_publn_id_citing

-- Step 4: Delete duplicate rows (if any).

with CTE as
(Select appln_id_controls, pat_publn_id_controls, pat_publn_id_citing,
appln_id_citing,
row_number() over (partition by appln_id_controls, pat_publn_id_controls,
pat_publn_id_citing, appln_id_citing order by appln_id_controls,
pat_publn_id_controls, pat_publn_id_citing, appln_id_citing) rn
from dbo.four_k)

delete from CTE where rn > 1

-- 927 duplicates are removed
-- 50558 rows remain.

select dbo.four_k.* from dbo.four_k

-- Step 5: Create a table only with appln_id_controls, appln_id_citing. That's
everything what is needed.
```

```
select dbo.four_k.appln_id_controls, dbo.four_k.appln_id_citing
into four_l
from dbo.four_k

-- Step 6: Remove duplicates again.

with CTE as
(select appln_id_controls, appln_id_citing,
row_number() over (partition by appln_id_controls, appln_id_citing
order by appln_id_controls, appln_id_citing) rn
from dbo.four_l)

delete from CTE where rn > 1

-- 9669 rows removed.

select dbo.four_l.* from dbo.four_l

-- 40889 unique citation links remain.

-- step 7: Add filing_date to citing patents.
-- Only include citations between 2003 - 2014.

select dbo.four_l.*, patstat2020a.dbo.tls201_appln.appln_filing_date
into four_m
from dbo.four_l
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
dbo.four_l.appln_id_citing
where appln_filing_date < '2015-01-01'

select dbo.four_m.* from dbo.four_m

-- 28750 citation links remain.
```

Self-citations

```
-- Step 8: Remove self-citations.
-- Add person_id to patents.

select dbo.four_m.appln_id_controls, patstat2020a.dbo.tls207_pers_appln.person_id as
person_id_controls, dbo.four_m.appln_id_citing
into four_n
from dbo.four_m
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.four_m.appln_id_controls

select dbo.four_n.*, patstat2020a.dbo.tls207_pers_appln.person_id as person_id_citing
into four_o
from dbo.four_n
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= dbo.four_n.appln_id_citing

-- Step 9: Add inventor sequence number.
-- This number indicates whether a person is an organization yes or not.

select distinct dbo.four_o.person_id_controls,
patstat2020a.dbo.tls227_pers_publn.invt_seq_nr
into four_p
from dbo.four_o
join patstat2020a.dbo.tls227_pers_publn on
patstat2020a.dbo.tls227_pers_publn.person_id = dbo.four_o.person_id_controls
```

```
select distinct dbo.four_o.*, dbo.four_p.invt_seq_nr
into four_q
from dbo.four_o
join dbo.four_p on dbo.four_p.person_id_controls = dbo.four_o.person_id_controls

-- Step 10: identify self-citations by organizations.
-- Filter on invt_seq_nr = 0. This are citations by organizations.

select dbo.four_q.*
from dbo.four_q
where ((person_id_controls = person_id_citing) AND (invt_seq_nr = 0))

-- Result: 2322 self-citations

-- Step 11: Create a table of self-citations

select dbo.four_q.*
into four_r
from dbo.four_q
where ((person_id_controls = person_id_citing) AND (invt_seq_nr = 0))

select dbo.four_r.appln_id_controls, dbo.four_r.appln_id_citing
into four_s
from dbo.four_r

with CTE as
(select appln_id_controls, appln_id_citing,
row_number() over (partition by appln_id_controls, appln_id_citing
order by appln_id_controls, appln_id_citing) rn
from dbo.four_s)

delete from CTE where rn > 1

-- 81 duplicates removed. 2241 self-citations remain.

-- Step 12: Delete citing patents from the data.

select dbo.four_m.appln_id_controls, dbo.four_m.appln_id_citing
into four_t
from dbo.four_m

delete four_t
from four_t
inner join four_s
on four_t.appln_id_controls = four_s.appln_id_controls AND
       four_t.appln_id_citing = four_s.appln_id_citing;

-- Result: 2241 self-citations removed.

-- Step 13: check results.

select dbo.four_t.* from dbo.four_t

-- 26509 patents
```

Geo-coordinates

```
-- Step 14: Find the geo-coordinates of the citation links.
-- Link latitude and longitude to appln_id_controls
```

```
select dbo.four_t.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_controls, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as
lng_controls
into four_u
from dbo.four_t
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.four_t.appln_id_controls

-- NOTE: one filing can more locations. Therefore, more results.

-- Step 15: Find geocoordinates of the citation links.

select dbo.four_t.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_controls, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as
lng_controls
into four_w
from dbo.four_t
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.four_t.appln_id_controls

select dbo.four_w.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat as
lat_citing, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng as lng_citing
into four_x
from dbo.four_w
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.four_w.appln_id_citing

-- NOTE: Quite a proportion of the citing patents are not known in the dataset by
Rassenfosse et al. (2019)
-- Therefore, the geo-cordinates of some of the citing patents are unknown.

-- Step 16: Remove duplicates.

with CTE as
(Select appln_id_controls, appln_id_citing, lat_controls, lng_controls, lat_citing,
lng_citing,
row_number() over (partition by appln_id_controls, appln_id_citing, lat_controls,
lng_controls, lat_citing, lng_citing order by appln_id_controls, appln_id_citing,
lat_controls, lng_controls, lat_citing, lng_citing) rn
from dbo.four_x)

delete from CTE where rn > 1

select dbo.four_x.* from dbo.four_x

-- Result: 72047 rows.
```

Filter on BE/NE

```
-- Step 1:
-- Only citation links that origin from BE/NE.

select distinct dbo.four_x.*,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
into four_z
from dbo.four_x
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
(Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
```

```
dbo.four_x.appln_id_controls AND
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat = dbo.four_x.lat_controls)
where ctry_code = 'NL' OR ctry_code = 'BE'

select four_z.* from four_z
```

Neglected citation links

```
-- Step 17: Isolate appln_id_controls and appln_id_citing

select dbo.four_z.appln_id_controls, dbo.four_z.appln_id_citing
into four_y
from dbo.four_z

-- Step 18: Remove duplicates.

with CTE as
(select appln_id_controls, appln_id_citing,
row_number() over (partition by appln_id_controls, appln_id_citing
order by appln_id_controls, appln_id_citing) rn
from dbo.four_y)

delete from CTE where rn > 1

select dbo.four_y.* from dbo.four_y

-- Step 29: Analysis

-- Only 11,613 citations are still present.
-- This means that for almost 60% of the citation links the geo-coordinates are
missing in the dataset by Rassenfosse et al. (2019).
```

Part II: Variables

All information related to distance.

```
-- Step 1: Save data, including haversine distance, as new table: sample4_1

select sample4.*
into sample4_1
from sample4

-- Step 2: Find the shortest distance per citation link.
-- Since there are multiple inventors, there are multiple distances per citation link.

with CTE as
(Select appln_id_controls, appln_id_citing, Haversine,
row_number() over (partition by appln_id_controls, appln_id_citing order by haversine
asc) rn
from dbo.sample4_1)

delete from CTE where rn > 1

select sample4_1.* from sample4_1

-- Result: 11613 citation links.

-- Step 3: Per control patent, find the shortest citation. This will be used as a
proxy for distance in the model.

Select appln_id_controls, appln_id_citing, Haversine,
```

```
row_number() over (partition by appln_id_controls order by haversine asc) rn
into sample4_2
from dbo.sample4_1

select appln_id_controls, appln_id_citing, Haversine as shortest_citation
into sample4_3
from Sample4_2
where rn=1

select sample4_3.* from sample4_3

-- Step 4: Import sample2_4
-- This is the same data as sample2_1, but then the distance is rounded off.
-- Find shortest distance per citation link.

select sample4_4.* from sample4_4

with CTE as
(Select appln_id_controls, appln_id_citing, Haversine,
row_number() over (partition by appln_id_controls, appln_id_citing order by haversine
asc) rn
from dbo.sample4_4)

delete from CTE where rn > 1

select sample4_4.* from sample4_4

-- Step 5: per control patent, flag quanitiy of citations within a certain range.
-- range 30 km.

select appln_id_controls, haversine as dist_30
into sample4_dist30_1
from sample4_4
where haversine < 30

select appln_id_controls, COUNT(appln_id_controls) as dist_30
into sample4_dist30
from sample4_dist30_1
group by appln_id_controls

-- 50 km.

select appln_id_controls, haversine as dist_50
into sample4_dist50_1
from sample4_4
where haversine < 50

select appln_id_controls, COUNT(appln_id_controls) as dist_50
into sample4_dist50
from sample4_dist50_1
group by appln_id_controls

-- 100 km

select appln_id_controls, haversine as dist_100
into sample4_dist100_1
from sample4_4
where haversine < 100

select appln_id_controls, COUNT(appln_id_controls) as dist_100
into sample4_dist100
from sample4_dist100_1
```

```
group by appln_id_controls

-- 200 km

select appln_id_controls, haversine as dist_200
into sample4_dist200_1
from sample4_4
where haversine < 200

select appln_id_controls, COUNT(appln_id_controls) as dist_200
into sample4_dist200
from sample4_dist200_1
group by appln_id_controls

-- 500 km

select appln_id_controls, haversine as dist_500
into sample4_dist500_1
from sample4_4
where haversine < 500

select appln_id_controls, COUNT(appln_id_controls) as dist_500
into sample4_dist500
from sample4_dist500_1
group by appln_id_controls

-- 1000 km

select appln_id_controls, haversine as dist_1000
into sample4_dist1000_1
from sample4_4
where haversine < 1000

select appln_id_controls, COUNT(appln_id_controls) as dist_1000
into sample4_dist1000
from sample4_dist1000_1
group by appln_id_controls

-- 2500 km

select appln_id_controls, haversine as dist_2500
into sample4_dist2500_1
from sample4_4
where haversine < 2500

select appln_id_controls, COUNT(appln_id_controls) as dist_2500
into sample4_dist2500
from sample4_dist2500_1
group by appln_id_controls

-- 5000 km

select appln_id_controls, haversine as dist_5000
into sample4_dist5000_1
from sample4_4
where haversine < 5000

select appln_id_controls, COUNT(appln_id_controls) as dist_5000
into sample4_dist5000
from sample4_dist5000_1
group by appln_id_controls
```

```
-- 0 km

select appln_id_controls, haversine as dist_0
into sample4_dist0_1
from sample4_4
where haversine = 0

select appln_id_controls, COUNT(appln_id_controls) as dist_0
into sample4_dist0
from sample4_dist0_1
group by appln_id_controls

-- Step 6: Calculate average distance of all citation links per HTCE patent.

select appln_id_controls, avg(haversine) as avg_haversine, count(appln_id_citing) as
nb_citations
into sample4_5
from Sample4_4
group by appln_id_controls

-- Step 7: Put all distance related information into one table.

select sample4_5.*, sample4_3.shortest_citation, sample4_dist0.dist_0,
sample4_dist30.dist_30, sample4_dist50.dist_50, sample4_dist100.dist_100,
sample4_dist200.dist_200, sample4_dist500.dist_500, sample4_dist1000.dist_1000,
sample4_dist2500.dist_2500, sample4_dist5000.dist_5000
into sample4_6
from sample4_5
join sample4_3 on sample4_3.appln_id_controls = sample4_5.appln_id_controls
LEFT join sample4_dist0 on sample4_dist0.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist30 on sample4_dist30.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist50 on sample4_dist50.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist100 on sample4_dist100.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist200 on sample4_dist200.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist500 on sample4_dist500.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist1000 on sample4_dist1000.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist2500 on sample4_dist2500.appln_id_controls =
sample4_5.appln_id_controls
left join  sample4_dist5000 on sample4_dist5000.appln_id_controls =
sample4_5.appln_id_controls

select sample4_6.* from sample4_6

-- Result: 2151 unique control patents.
```

Control variables

```
-- Step 1: Find number of forward -and backward citations
-- Step 1a: Forward citations

select dbo.four_l.* from dbo.four_l

-- This table includes all forward citations of the control patents.
-- So, not only the citations before 2014.
```

```
-- Just everything.

select appln_id_controls, COUNT(appln_id_controls) as nb_forward_citations
into sample4_7_1
from four_1
group by appln_id_controls

-- Step 1b: Backward citations

select sample4_6.appln_id_controls, patstat2020a.dbo.tls211_pat_publn.pat_publn_id
into sample4_7_2
from sample4_6
join  patstat2020a.dbo.tls211_pat_publn on  patstat2020a.dbo.tls211_pat_publn.appln_id
= sample4_6.appln_id_controls

select sample4_7_2.appln_id_controls, sample4_7_2.pat_publn_id,
patstat2020a.dbo.tls212_citation.cited_pat_publn_id
into sample4_7_3
from sample4_7_2
join patstat2020a.dbo.tls212_citation on patstat2020a.dbo.tls212_citation.pat_publn_id
= sample4_7_2.pat_publn_id
where cited_pat_publn_id != 0

select appln_id_controls, COUNT(appln_id_controls) as nb_backward_citations
into sample4_7_4
from sample4_7_3
group by appln_id_controls

select sample4_6.appln_id_controls, sample4_7_1.nb_forward_citations,
sample4_7_4.nb_backward_citations
into sample4_7
from sample4_6
left join sample4_7_1 on sample4_7_1.appln_id_controls = sample4_6.appln_id_controls
left join sample4_7_4 on sample4_7_4.appln_id_controls = sample4_6.appln_id_controls

select sample4_7.* from sample4_7

-- Step 2: IPC and year

select dbo.two_c.* from dbo.two_c

-- most frequent IPC code and year can be found in 'two_c'.

-- Find number of IPC codes per control patent.

select sample4_6.appln_id_controls, patstat2020a.dbo.tls209_appln_ipc.ipc_class_symbol
into sample4_8_1
from sample4_6
left join patstat2020a.dbo.tls209_appln_ipc on
patstat2020a.dbo.tls209_appln_ipc.appln_id = sample4_6.appln_id_controls

select appln_id_controls, COUNT(appln_id_controls) as nb_IPC_codes
into sample4_8
from sample4_8_1
group by appln_id_controls

-- Step 3: nb_claims. Select patent with highest claims if patent is similar.

select sample4_6.appln_id_controls, patstat2020a.dbo.tls211_pat_publn.publn_claims
into sample4_9_1
from sample4_6
```

```
left join patstat2020a.dbo.tls211_pat_publn on
patstat2020a.dbo.tls211_pat_publn.appln_id = sample4_6.appln_id_controls

with CTE as
(Select appln_id_controls, publn_claims,
row_number() over (partition by appln_id_controls order by publn_claims desc) rn
from dbo.sample4_9_1)

delete from CTE where rn > 1

select sample4_9_1.* from sample4_9_1

-- Step 4: Backward citations to literature

select sample4_7_2.* from sample4_7_2

select sample4_7_2.*, patstat2020a.dbo.tls212_citation.cited_npl_publn_id
into sample4_10_1
from sample4_7_2
left join patstat2020a.dbo.tls212_citation on
patstat2020a.dbo.tls212_citation.pat_publn_id = sample4_7_2.pat_publn_id
where cited_npl_publn_id != '0'

select appln_id_controls, COUNT(appln_id_controls) as nb_cited_literature
into sample4_10
from sample4_10_1
group by appln_id_controls

-- Step 5: Find company names
-- In case of multiple companies per one control patent, choose one company name at
random per one control patent.

select sample4_6.appln_id_controls, patstat2020a.dbo.tls207_pers_appln.person_id
into sample4_11_1
from sample4_6
join patstat2020a.dbo.tls207_pers_appln on patstat2020a.dbo.tls207_pers_appln.appln_id
= sample4_6.appln_id_controls

select sample4_11_1.appln_id_controls, patstat2020a.dbo.tls206_person.person_name,
patstat2020a.dbo.tls206_person.psn_level, patstat2020a.dbo.tls206_person.psn_sector
into sample4_11_2
from sample4_11_1
join patstat2020a.dbo.tls206_person on patstat2020a.dbo.tls206_person.person_id =
sample4_11_1.person_id
where psn_sector = 'company'

select sample4_11_2.* from sample4_11_2

WITH CTE AS (
    SELECT
        appln_id_controls,
        person_name,
        ROW_NUMBER() OVER(PARTITION BY appln_id_controls ORDER BY NEWID()) AS RowNum
    FROM
        dbo.sample4_11_2)

SELECT
    appln_id_controls,
    person_name INTO sample4_11
FROM
    CTE
WHERE
```

```
    RowNum = 1

select sample2_11.* from sample2_11

-- Step 6: Number of foreign inventors

select sample4_6.appln_id_controls,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.ctry_code
into sample4_12_3
from sample4_6
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
sample4_6.appln_id_controls
where ctry_code != 'NL' and ctry_code != 'BE'

select sample4_12_3.appln_id_controls, count(appln_id_controls) as
nb_foreign_inventors
into sample4_12
from sample4_12_3
group by appln_id_controls

-- Step 7: Add nb_applicants, only companies

select sample4_11_2.* from sample4_11_2

select sample4_11_2.appln_id_controls, count(appln_id_controls) as
nb_applicants_companies
into sample4_13
from sample4_11_2
group by appln_id_controls

-- Step 8: Add a variable that indicates how much HTCE patents have the same control
patent.
-- So, a weight can be included on those patents.

select appln_id_controls, count(appln_id_controls) as frq_controls
into sample4_14
from four_g
group by appln_id_controls

-- Step 9: Put everything into one table.

select sample4_6.*, patstat2020a.dbo.tls201_appln.appln_auth,
patstat2020a.dbo.tls201_appln.nb_inventors, sample4_13.nb_applicants_companies,
sample4_11.person_name,
sample4_12.nb_foreign_inventors, two_c.IPC, sample4_8.nb_IPC_codes,
two_c.earliest_filing_year, sample4_7.nb_backward_citations,
sample4_7.nb_forward_citations,
sample4_10.nb_cited_literature, sample4_9_1.publn_claims, sample4_14.frq_controls
into sample4_15
from sample4_6
join patstat2020a.dbo.tls201_appln on patstat2020a.dbo.tls201_appln.appln_id =
sample4_6.appln_id_controls
LEFT join sample4_13 on sample4_13.appln_id_controls = sample4_6.appln_id_controls
left join sample4_11 on sample4_11.appln_id_controls = sample4_6.appln_id_controls
left join sample4_12 on sample4_12.appln_id_controls = sample4_6.appln_id_controls
left join two_c on two_c.appln_id_controls = sample4_6.appln_id_controls
left join sample4_8 on sample4_8.appln_id_controls = sample4_6.appln_id_controls
left join sample4_7 on sample4_7.appln_id_controls = sample4_6.appln_id_controls
left join sample4_10 on sample4_10.appln_id_controls = sample4_6.appln_id_controls
```

```
left join sample4_9_1 on sample4_9_1.appln_id_controls = sample4_6.appln_id_controls
left join sample4_14 on sample4_14.appln_id_controls = sample4_6.appln_id_controls

select sample4_15.* from sample4_15
```

Match sample1 & sample2 on IPC and year

```
-- Step 1: Find all possible control, from controls dataset, patents per brainport
patent.

select sample3_14.* from sample3_14
select sample4_15.* from sample4_15

select sample3_14.appln_id_brainport, sample3_14.IPC, sample3_14.earliest_filing_year,
sample4_15.appln_id_controls
into joint2_1
from sample3_14
join sample4_15 on sample4_15.IPC = sample3_14.IPC
where sample4_15.earliest_filing_year = sample3_14.earliest_filing_year

-- Step 2: For each brainport, find possible number of control patents.

select joint2_1.* from joint2_1

select appln_id_brainport, COUNT(appln_id_brainport) as nb_matching_control_patents
into joint2_2
from joint2_1
group by appln_id_brainport

-- Result: Out of the 3484, 2992 brainport patents do have a matching control patent.

-- Step 3: For each control patent, find the number of matches to the brainport
patents.

select appln_id_controls, COUNT(appln_id_controls) as nb_matching_brainport_patents
into joint2_3
from joint2_1
group by appln_id_controls

-- Result: 1900 control patents have a match to a brainport patent, this number is
lower than the brainport patents.
-- So, there are less control patents than treated patents in the dataset.

-- Step 4: Join 'nb_matching_control_patents' and 'nb_matching_brainport_patents', to
final tables.

select sample3_14.*, joint2_2.nb_matching_control_patents
into sample3_15
from sample3_14
join joint2_2 on joint2_2.appln_id_brainport = sample3_14.appln_id_brainport

select sample4_15.*, joint2_3.nb_matching_brainport_patents
into sample4_16
from sample4_15
join joint2_3 on joint2_3.appln_id_controls = sample4_15.appln_id_controls

-- Step 5: Create one table in Excel of both brainport -and control dataset.

select sample3_15.* from sample3_15
select sample4_16.* from sample4_16
```

Find weights for control patents

```
-- For each couple of IPC and year, find the number of matching brainport -and control
patents.
-- In that way, a weight can be assessed to the controls.
-- In this case, there are more brainport patents than control patents.

-- Step 1: For each couple of IPC and year, find the number of matching brainport
patents.

select appln_id_brainport, IPC, earliest_filing_year
into weight2_1
from sample3_15

select IPC, earliest_filing_year, COUNT(appln_id_brainport) as
fq_combination_brainport
into weight2_2
from weight2_1
group by IPC, earliest_filing_year

-- Step 2: For each couple of IPC and year, find the number of matching control
patents.

select IPC, earliest_filing_year, COUNT(appln_id_controls) as fq_combination_controls
into weight2_3
from sample4_16
group by IPC, earliest_filing_year

-- Step 3: Combine values into one table

select weight2_2.*, weight2_3.fq_combination_controls
into weight2_4
from weight2_2
join weight2_3 on weight2_3.IPC = weight2_2.IPC
where weight2_3.earliest_filing_year = weight2_2.earliest_filing_year

select weight2_4.* from weight2_4

-- Step 4: Export 'weight2_4' into excel.
-- Divide both frequencies with one another.

-- Import new data as table 'weight2_5'

select weight2_5.* from weight2_5

-- Step 5: connect variable 'weight_control' to sample4_16.
-- The treated patents get value one.

select sample4_16.*, weight2_5.weight_control
from sample4_16
join weight2_5 on weight2_5.IPC = sample4_16.IPC
where weight2_5.earliest_filing_year = sample4_16.earliest_filing_year
```

Extra

Finding province sample 2

```
-- Step 1: Create tables including province.

select dbo.two_z.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.name_2,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.name_1
into two_province
from dbo.two_z
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.two_z.appln_id_controls
where (lat_controls = Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat) and
(lng_controls = Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng)


-- In excel, the haversine distance was calculated and loaded in back to SQL under
table “sample2_province_2”.


-- Step 2: Find the shortest distance per citation link.
-- Since there are multiple inventors, there are multiple distances per citation link.

with CTE as
(Select appln_id_controls, appln_id_citing, ctry_code, name_2, name_1, haversine,
haversine_round,
row_number() over (partition by appln_id_controls, appln_id_citing order by haversine
asc) rn
from dbo.sample2_province_2)

delete from CTE where rn > 1

select dbo.sample2_province_2.* from dbo.sample2_province_2

-- Step 3: Per unique cited patent, find the corresponding province.

select distinct dbo.sample2_province_2.appln_id_controls,
dbo.sample2_province_2.ctry_code, dbo.sample2_province_2.name_2,
dbo.sample2_province_2.name_1
into sample2_province_3
from dbo.sample2_province_2

-- Since some control patents do have multiple locations, per cited patent, select one
location at random.

WITH CTE AS (
    SELECT
        appln_id_controls, ctry_code, name_2, name_1,
        ROW_NUMBER() OVER(PARTITION BY appln_id_controls ORDER BY NEWID()) AS RowNum
    FROM
        dbo.sample2_province_3)

SELECT
     appln_id_controls, ctry_code, name_2, name_1
        INTO sample2_province_4
FROM
    CTE
WHERE
    RowNum = 1

select sample2_province_4.* from sample2_province_4

-- Step 5: Add province to final table
```

```
select sample2_16.* from sample2_16

select sample2_16.*, dbo.sample2_province_4.ctry_code, dbo.sample2_province_4.name_2,
dbo.sample2_province_4.name_1
from sample2_16
left join sample2_province_4 on sample2_province_4.appln_id_controls =
sample2_16.appln_id_controls
```

Find distances larger than sample 1 and 2

```
-- Sample1

-- > 2500 km

select appln_id_HTCE, haversine as dist_larger_2500
into sample1_dist_larger_2500_1
from sample1_4
where haversine > 2500

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_larger_2500
into sample1_dist_larger_2500_2
from sample1_dist_larger_2500_1
group by appln_id_HTCE

-- > 5000 km

select appln_id_HTCE, haversine as dist_larger_5000
into sample1_dist_larger_5000_1
from sample1_4
where haversine > 5000

select appln_id_HTCE, COUNT(appln_id_HTCE) as dist_larger_5000
into sample1_dist_larger_5000_2
from sample1_dist_larger_5000_1
group by appln_id_HTCE

-- Sample 2

-- > 2500 km

select appln_id_controls, haversine as dist_larger_2500
into sample2_dist_larger_2500_1
from sample2_4
where haversine > 2500

select appln_id_controls, COUNT(appln_id_controls) as dist_larger_2500
into sample2_dist_larger_2500_2
from sample2_dist_larger_2500_1
group by appln_id_controls

-- > 5000 km

select appln_id_controls, haversine as dist_larger_5000
into sample2_dist_larger_5000_1
from sample2_4
where haversine > 5000

select appln_id_controls, COUNT(appln_id_controls) as dist_larger_5000
into sample2_dist_larger_5000_2
from sample2_dist_larger_5000_1
group by appln_id_controls
```

```
-- Combine data to tables

-- Sample 1

select sample1_15.* from sample1_15

select sample1_15.*, sample1_dist_larger_2500_2.dist_larger_2500,
sample1_dist_larger_5000_2.dist_larger_5000
from sample1_15
left join sample1_dist_larger_2500_2 on sample1_dist_larger_2500_2.appln_id_HTCE =
sample1_15.appln_id_HTCE
left join sample1_dist_larger_5000_2 on sample1_dist_larger_5000_2.appln_id_HTCE =
sample1_15.appln_id_HTCE

-- Sample 2

select sample2_16.* from sample2_16

select sample2_16.*, sample2_dist_larger_2500_2.dist_larger_2500,
sample2_dist_larger_5000_2.dist_larger_5000
from sample2_16
left join sample2_dist_larger_2500_2 on sample2_dist_larger_2500_2.appln_id_controls =
sample2_16.appln_id_controls
left join sample2_dist_larger_5000_2 on sample2_dist_larger_5000_2.appln_id_controls =
sample2_16.appln_id_controls
```

Find distances larger than sample 3 and 4

```
-- Sample3

-- > 2500 km

select appln_id_brainport, haversine as dist_larger_2500
into sample3_dist_larger_2500_1
from sample3_4
where haversine > 2500

select appln_id_brainport, COUNT(appln_id_brainport) as dist_larger_2500
into sample3_dist_larger_2500_2
from sample3_dist_larger_2500_1
group by appln_id_brainport

-- > 5000 km

select appln_id_brainport, haversine as dist_larger_5000
into sample3_dist_larger_5000_1
from sample3_4
where haversine > 5000

select appln_id_brainport, COUNT(appln_id_brainport) as dist_larger_5000
into sample3_dist_larger_5000_2
from sample3_dist_larger_5000_1
group by appln_id_brainport

-- Sample 4

-- > 2500 km

select appln_id_controls, haversine as dist_larger_2500
into sample4_dist_larger_2500_1
from sample4_4
where haversine > 2500
```

```
select appln_id_controls, COUNT(appln_id_controls) as dist_larger_2500
into sample4_dist_larger_2500_2
from sample4_dist_larger_2500_1
group by appln_id_controls

-- > 5000 km

select appln_id_controls, haversine as dist_larger_5000
into sample4_dist_larger_5000_1
from sample4_4
where haversine > 5000

select appln_id_controls, COUNT(appln_id_controls) as dist_larger_5000
into sample4_dist_larger_5000_2
from sample4_dist_larger_5000_1
group by appln_id_controls

-- Combine data to tables

-- Sample 3

select sample3_15.* from sample3_15

select sample3_15.*, sample3_dist_larger_2500_2.dist_larger_2500,
sample3_dist_larger_5000_2.dist_larger_5000
from sample3_15
left join sample3_dist_larger_2500_2 on sample3_dist_larger_2500_2.appln_id_brainport
= sample3_15.appln_id_brainport
left join sample3_dist_larger_5000_2 on sample3_dist_larger_5000_2.appln_id_brainport
= sample3_15.appln_id_brainport

-- Sample 4

select sample4_16.* from sample4_16

select sample4_16.*, sample4_dist_larger_2500_2.dist_larger_2500,
sample4_dist_larger_5000_2.dist_larger_5000
from sample4_16
left join sample4_dist_larger_2500_2 on sample4_dist_larger_2500_2.appln_id_controls =
sample4_16.appln_id_controls
left join sample4_dist_larger_5000_2 on sample4_dist_larger_5000_2.appln_id_controls =
sample4_16.appln_id_controls
```

Finding province sample 4

```
-- Step 1: Create tables including province.

select dbo.four_z.*, Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.name_2,
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.name_1
into four_province
from dbo.four_z
join Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean on
Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.appln_id =
dbo.four_z.appln_id_controls
where (lat_controls = Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lat) and
(lng_controls = Rassenfosse_Kozak_Seliger_geocoding.dbo.geoc_app_clean.lng)

select dbo.two_province.* from dbo.two_province

select sample4_province_2.* from sample4_province_2
```

```
-- Step 2: Find the shortest distance per citation link.
-- Since there are multiple inventors, there are multiple distances per citation link.

with CTE as
(Select appln_id_controls, appln_id_citing, ctry_code, name_2, name_1, haversine,
haversine_round,
row_number() over (partition by appln_id_controls, appln_id_citing order by haversine
asc) rn
from dbo.sample4_province_2)

delete from CTE where rn > 1

select dbo.sample4_province_2.* from dbo.sample4_province_2

-- Step 3: Per unique cited patent, find the corresponding province.

select distinct dbo.sample4_province_2.appln_id_controls,
dbo.sample4_province_2.ctry_code, dbo.sample4_province_2.name_2,
dbo.sample4_province_2.name_1
into sample4_province_3
from dbo.sample4_province_2

-- Since some control patents do have multiple locations, per cited patent, select one
location at random.

WITH CTE AS (
    SELECT
        appln_id_controls, ctry_code, name_2, name_1,
        ROW_NUMBER() OVER(PARTITION BY appln_id_controls ORDER BY NEWID()) AS RowNum
    FROM
        dbo.sample4_province_3)

SELECT
     appln_id_controls, ctry_code, name_2, name_1
       INTO sample4_province_4
FROM
    CTE
WHERE
    RowNum = 1

select sample4_province_4.* from sample4_province_4

-- Step 5: Add province to final table

select sample4_16.* from sample4_16

select sample4_16.*, dbo.sample4_province_4.ctry_code, dbo.sample4_province_4.name_2,
dbo.sample4_province_4.name_1
from sample4_16
left join sample4_province_4 on sample4_province_4.appln_id_controls =
sample4_16.appln_id_controls
```

Appendix B: Do file Stata

```
* DO FILE; HTCE group

* 1. import data
import excel "/Users/jespervangriensven/Documents/TU:e/MASTER/Scriptie /Data/Stata ready/joint1 v2 .xlsx", sheet("Sheet1") firstrow

* It appears that some patents are present before 2003. Delete them.
drop if earliest_filing_year < 2003

* 2. generate goups
egen firm = group(name)
egen year = group(earliest_filing_year)
egen prov = group(province)
egen code = group(IPC)

* 3. generate dummy variables for distance variables

* close range distance variables
gen dum_0 = (dist_0>0)
gen dum_30 = (dist_30>0)
gen dum_50 = (dist_50>0)
gen dum_100 = (dist_100>0)
gen dum_200 = (dist_200>0)
gen dum_500 = (dist_500>0)
gen dum_1000 = (dist_1000>0)
gen dum_2500 = (dist_2500>0)
gen dum_5000 = (dist_5000>0)

* long range distance variables
gen dum_larger_2500 = (dist_larger_2500>0)
gen dum_larger_5000 = (dist_larger_5000>0)

* dummy patents with more than one citation link

gen firm_1 = (nb_citations>1)

* dummy for number of firms as applicant on a patent

gen applicant = (nb_applicants_companies>1)
```

Descriptive statistics

```
* DESCRIPTIVE STATISTICS


* 1. Frequency companies

tabulate name if treated == 1
tabulate name if treated == 0

* 2. Frequncy IPC codes

ssc desc groups
ssc inst groups

groups IPC, order(h) select(10)

* 3. Frequency earliest_filing_year

hist earliest_filing_year


* 4. Mean and t-test distance variables

ssc install estout, replace

estpost ttest avg_haversine, by(treated)
esttab

estpost ttest shortest_citation, by(treated)
esttab

estpost ttest dum_0, by(treated)
esttab

estpost ttest dum_30, by(treated)
esttab

estpost ttest dum_50, by(treated)
esttab

estpost ttest dum_100, by(treated)
esttab
```

```
estpost ttest dum_200, by(treated)
esttab

estpost ttest dum_500, by(treated)
esttab

estpost ttest dum_1000, by(treated)
esttab

estpost ttest dum_2500, by(treated)
esttab

estpost ttest dum_5000, by(treated)
esttab

estpost ttest dum_larger_2500, by(treated)
esttab

estpost ttest dum_larger_5000, by(treated)
esttab

* 5. Density plots

* < 2500 km

twoway (kdensity shortest_citation if treated==0 [aweight = weight], lcolor(blue)) (kdensity shortest_citation if treated==1 [aweight = weight], lcolor(red)) if shortest_citation < 2500 & shortest_citation != 0

* < 200 km

twoway (kdensity shortest_citation if treated==0 [aweight = weight], lcolor(blue)) (kdensity shortest_citation if treated==1 [aweight = weight], lcolor(red)) if shortest_citation < 200 & shortest_citation != 0

* < 200 km, including 0 km

twoway (kdensity shortest_citation if treated==0 [aweight = weight], lcolor(blue)) (kdensity shortest_citation if treated==1 [aweight = weight], lcolor(red)) if shortest_citation < 200
```

Linear regressions

* FIRST group of linear regressions, to answer RQ

eststo: regress avg_haversine i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress shortest_citation i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress dum_50 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress dum_100 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress dum_200 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress dum_500 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

esttab using lr_HTCE_1.rtf, r2 drop(*.firm *.code *.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_forward_citations nb_backward_citations publn_claims)

eststo clear

* SECOND group of linear regressions, additional findings

eststo: regress dum_100 i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress dum_200 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight]

eststo: regress dum_200 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0 & firm_1 == 1

eststo: regress dum_larger_2500 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0 & firm_1 == 1

eststo: regress dum_larger_2500 i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress applicant i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress shortest_citation i.firm i.code i.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims applicant treated [aweight= weight] if shortest_citation != 0

eststo: regress nb_inventors i.firm i.code i.year nb_citations nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

eststo: regress nb_foreign_inventors i.firm i.code i.year nb_citations nb_ipc_codes nb_backward_citations nb_forward_citations publn_claims treated [aweight= weight] if shortest_citation != 0

esttab using lr_HTCE_2.rtf, r2 drop(*.firm *.code *.year nb_citations nb_inventors nb_foreign_inventors nb_ipc_codes nb_forward_citations nb_backward_citations publn_claims)

Declaration concerning the TU/e Code of Scientific Conduct for the Master's thesis

I have read the TU/e Code of Scientific Conduct[i].

I hereby declare that my Master's thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct

Date

06-07-2023

……..……..

Name

Jesper Hendricus Petrus van Griensven

……..……..

ID-number

1266586

……..……..

Signature

……..……..

Submit the signed declaration to the student administration of your department.

[i] See: https://www.tue.nl/en/our-university/about-the-university/organization/integrity/scientific-integrity/
The Netherlands Code of Conduct for Scientific Integrity, endorsed by 6 umbrella organizations, including the VSNU, can be found here also. More information about scientific integrity is published on the websites of TU/e and VSNU