CPB Netherlands Bureau for Economic Policy Analysis

CPB Discussion Paper | 348

Knowledge diffusion across regions and countries: evidence from patent citations

Wiljan van den Berge
Jonneke Bolhaar
Roel van Elk

Knowledge diffusion across regions and countries: evidence from patent citations

Working paper

Wiljan van den Berge, Jonneke Bolhaar and Roel van Elk*

April 3, 2017

Abstract

We study knowledge spillovers from European universities and other research organizations using data from patent citations at the EPO. Using matching techniques to construct a sample of control patents, we show that the probability to cite a university patent declines with distance. In particular, we find a sharp cut-off at around 25 kilometers. For longer distances the probability to cite a university patent is more or less constant. For other research organizations we find no evidence that distance plays a role. Country borders are shown to play an important role in restricting the diffusion of patents of both universities and other research organizations. These results are in line with recent literature for the U.S. and suggest that knowledge spillovers and tacit knowledge are important when using knowledge embodied in university patents.

Keywords: Knowledge spillovers, patents, universities, research organizations
JEL classification: O33 O34 I23

*CPB Netherlands Bureau for Economic Policy Analysis, P.O. Box 80510 2508 GM The Hague, The Netherlands. Van den Berge: W.van.den.Berge@cpb.nl, Bolhaar: J.A.Bolhaar@cpb.nl, Van Elk: R.A.van.Elk@cpb.nl. We thank Maarten Bosker and Daniel van Vuuren for comments on an earlier draft. Any remaining errors are our own.

1 Introduction

Knowledge spillovers are important for economic growth (Romer, 1986; Aghion and Howitt, 1992). Universities and other research organizations play a key role in producing knowledge.[1] Through academic papers, patents, seminars and licensing agreements to firms and other organizations, knowledge spreads and leads to productivity growth.

There is growing evidence that knowledge spillovers are localized (Jaffe et al., 1993; Audretsch and Feldman, 1996; Peri, 2005; Belenzon and Schankerman, 2013). Face to face contact is important to gain access to the tacit knowledge that is often necessary to apply new ideas in fruitful ways. Social networks are another important way through which knowledge spreads. Furthermore, the commercial application of new inventions through spin-off companies started by researchers close to universities could be an important driver of the localized spread of knowledge.

In this paper we estimate the extent to which knowledge spillovers from European universities and other research organizations are localized.[2] Spillovers are notoriously difficult to measure. We adopt a commonly used measure in the literature: citations to patents. Citations show a 'paper trail' of cases where knowledge created at a university is used in inventions by firms (Jaffe et al., 1993).[3] We use data on patent citations to university-owned patents from European firms to estimate the effect of distance to the university on the probability to cite a university patent. Patents are by themselves an important output of research institutes. They protect the generated knowledge and secure economic advantage. A drawback is that patents are focused on technology, hence an analysis on knowledge diffusion using patents is mostly restricted to the output from technical and medical institutes. However, patents are also a good proxy for the overall use of knowledge generated at an institute. Since patents are open and available, they can be used by anyone to build on the knowledge in them. This makes them very similar to other research outputs, such as scientific papers.[4]

Our analysis requires careful consideration of possible sources of selection bias. If we compare a patent that cites a university patent with a random patent that doesn't cite the university patent, it is likely that the citing patent will be closer to the university than the non-citing patent. However, this doesn't provide evidence of knowledge spillovers. There could be many reasons why a citing patent is closer to a university patent. For example, they could both be patents in the same technological field. This increases the probability of citation, but also increases the probability that the firm is closer to the university because of other reasons, such as the availability of graduates in relevant fields, the presence of natural resources or government policy. To overcome these possible sources of selection bias we follow the strategy introduced by Jaffe et al. (1993) and for example recently used by Belenzon and Schankerman (2013). They compare citing patents with a control group of patents that

[1] Of course universities also impact (regional) economic growth through other mechanisms, such as producing a higher educated population (see e.g. Leten et al. (2014); Andersson et al. (2004, 2009); Rosenthal and Strange (2008)), start-ups (see e.g. Abramovsky et al. (2007); Agrawal et al. (2014); Audretsch et al. (2005)) and other agglomeration externalities (see e.g. Liu (2015); Kantor and Whalley (2014)). See Drucker and Goldstein (2007) for a general overview of other approaches and mechanisms.

[2] In the analysis we distinguish between universities and other research organizations. For reasons of brevity we will refer to both as "universities" if the distinction is not important.

[3] Note that patent citations also capture cases where there is no pure spillover, such as when universities collaborate with firms or engage in licensing agreements.

[4] Belenzon and Schankerman (2013) show that localization patterns of patents and scientific papers in the US are similar.

do not cite, but that share the same technology and year as the citing patent. The idea is that since a university patent is publicly available, the ideas in it could be used by anyone, so the probability to cite the patent is *a priori* equal for each potential inventor working in the same technological field and in the same time period.

This identification strategy relies on a reliable measure of technology that sufficiently controls for heterogeneity. If there is an imperfect match between a citing patent and a non-citing patent, we haven't ruled out other confounding factors that might influence the probability to cite. Most of the literature relies on three-digit technology classes in USPTO patents (Jaffe et al., 1993; Belenzon and Schankerman, 2013). Some papers suggest that this is indeed too coarse and could lead to spurious evidence for local spillovers (Thompson and Fox-Kean, 2005a,b; Henderson et al., 2005). To ensure that we get the best possible match, we use International Patent Classification (IPC) codes on the detailed main group level and, if patents have multiple codes, we use multiple codes to precisely pinpoint the type of technology the patent contributes to.

Using this matched sample, we estimate the effect of distance on the probability to cite a university patent. We use a flexible step-wise distance specification that allows us to determine how the probability to cite is affected by distance for different intervals. We show that the probability to cite decreases with distance, but that after about 100 kilometers distance doesn't seem to play an important role anymore. There is a large drop-off in citation probability after 25 kilometers. We also show that country borders restrict knowledge spillovers. This means that national policies, but also a shared language, could be relevant factors in using knowledge produced at universities. These results remain robust after controlling for technology and within-country citations. For other research organizations we find, once we take into account country effects, no evidence that distance plays a role in the citation probability. An explanation could be that patents generated at universities contain more fundamental knowledge that requires more tacit knowledge to be applied than patents generated at other research organizations. However, due to uncertainty around the estimates for other research organizations, we can't reject the hypothesis that the effects of distance on citation probabilities are actually similar to those for universities. Country borders play a role for both universities and other research organizations. Finally, we find strong evidence that distance matters much more for early citations to university patents than for later citations. This is consistent with the diffusion of knowledge over time. A possible explanation for the patterns we find is the establishment of commerial spin-off companies, which are typically established close to a university.

Our results are in line with the literature. For the United States Jaffe et al. (1993) find that patents from American universities and firms are 2 to 6 times more likely to be cited in the same metropolitan area and twice as likely in the same state. Belenzon and Schankerman (2013) focus specifically on American universities and find that the probability to cite a university patent strongly declines with distance up to about 50 miles (80 kilometers). After 50 miles the citation probability is more or less constant. They also show that - separately from distance - state borders matter too. This suggests that local policies might play an important role in citation patterns. Finally, they show that the patterns are similar for citations to scientific publications in firm patents. Thompson (2006) uses a different strategy that relies on the distinction between inventor-added and examiner-added citations in USPTO patents. He finds similar localization patterns as the papers relying on a matching strategy. For Europe Maurseth and Verspagen (2002) look at knowledge flows between regions. They show that the number of citations to patents between regions declines as they are further away from each other, but increases as they share the same language, even when controlling for country fixed effects. This suggests that shared language (between e.g. Austria and Germany) affects the use of knowledge, apart from any country effects. Peri (2005) uses both European and North American data on patent citations at the USPTO to look at knowledge flows across regions, and finds that only 20% of knowledge flows out of an average region. Another 36% doesn't leave the next region and another 20% doesn't leave the country. Griffith et al. (2011) show that patent citations have become less local over time. Patents are now more often cited by organizations in other countries, which is probably related to the fall in communication and travel costs.[5]

We contribute to this literature in two ways. First, we provide the first evidence on how knowledge spillovers are constrained by distance for European universities and research organizations using patent citations. Second, as far as we're aware, this is the first paper to consider both universities and other (private and public) research organizations.

The remainder of the paper is organized as follows. Section 2 presents the data we use and the steps we took to prepare the data for analysis. Section 3 discusses our estimation strategy in more detail. Section 4 shows our main results, while section 5 discusses the robustness of our results. Section 6 concludes.

2 Data

We use the February 2015 edition of the OECD patent database, which is based on the Autumn 2014 version of the EPO Worldwide Statistical Patent Database (PATSTAT). The OECD patent database is freely available and consists of several separate databases extracted from PATSTAT. In our analysis we use the Citations database, the REGPAT database with regional information and the HAN database with harmonized firm and institution names.[6] We extract all patent applications to the European Patent Office (EPO) from 1978 to 2013. The main advantage of only using the EPO data is that there is no possibility of including multiple patent applications at different national patent agencies for the same invention.

To prepare the data for analysis, we use a three-step procedure.[7] In the first step we identify which patent applications cite a patent from a university or other research organization. This will be our treatment group. In the second step, we try to find one or more control patent applications for each patent application in the treatment group, using a matching procedure based on codes from the International Patent Classification (IPC) system. In the third step the distance is calculated between the two addresses using online mapping software.

[5]A recent case study also finds that national borders don't seem to restrict the diffusion of big data technology (Kiseleva et al., 2016).

[6]More information, including the data manual and information on how to obtain the data, is available at http://www.oecd.org/sti/inno/oecdpatentdatabases.htm.

[7]All syntax files to create the data are available from the authors.

2.1 Identifying universities and research organizations

We use an algorithm developed by van Looy et al. (2009) to identify universities and research organizations in the PATSTAT sample. Their algorithm, developed for Eurostat, assigns a sector to each patent applicant. It is based on an earlier version of the algorithm by Van Looy et al. (2006). The sector allocation from this algorithm is publicly available in the ECOOM-EUROSTAT-EPO PATSTAT Person Augmented Table (EEE-PPAT).[8] The algorithm assigns one of 5 sectors to every applicant. The sectors that are distinguished in this version are (1) individuals, (2) private business enterprises, (3) public and private non-profit organizations, (4) universities/higher education organizations, (5) hospitals. The algorithm uses the patentee names in an iterative procedure of rule based (i.e. using keywords) and case-based logic to assign a sector to each patentee name. The authors claim that after the last iteration, 99% of the patent volume is allocated correctly to a sector.

Although the earlier version of the algorithm (Van Looy et al. (2006)) separated public non-profit from private non-profit organizations, which would at first sight be preferable for our analysis, the later version of the algorithm (van Looy et al. (2009)) is used where these two categories are taken together. The reason the authors of the algorithm decided to combine the two categories was that it turned out to be very difficult to distinguish between the two. Background information on establishment, funding and governance was required to make the distinction. Given that the two categories are so close to each other that comparable research institutes may end up in different categories, we prefer to use the version of the algorithm from van Looy et al. (2009). The sector 'public and private non-profit' therefore contains all research institutes and government bodies that conduct research, that are not universities.[9] We will therefore use the sector 'public and private non-profit' as a proxy for all other research organizations outside academia (and will hereafter refer to them as 'other research organizations').

The advantage of using the sector allocation from EEE-PPAT is that it is rigorously checked which makes it very reliable. A disadvantage of this sector allocation is that it does not allow us to distinguish between private and public non-profit organisations. The earlier (2006) version of the algorithm did make that distinction, but it turned out to be very difficult to determine to which of the two categories an applicant belongs, solely based on the name of the applicant. Other sources on e.g. funding of the organisation are necessary to correctly determine whether it is a public or private non-profit organisation. As this would imply a large increase in the validation efforts that have to be made to get the allocation right, the authors decided to combine public and private non-profit as a single sector.[10]

2.2 Constructing a treatment and control group

The treatment group consists of patent applications by firms (referred to as the *citing* patents) that cite a patent from a university in their application (referred to as the *cited* patent). For every patent

[8] We like to thank Xiaoyan Song from KU Leuven for supplying the latest version of EEE-PPAT to us.

[9] Although only a small part of the government sector consists of research organizations, most (if not all) patents filed by the government sector will be from research organizations.

[10] We have tried to assign public and private labels to research organizations ourselves and frequently ran into similar problems. If we nevertheless use this sample and drop all research organizations categorized as private, the results remain similar.

in the treatment group, one or more control patents are selected. These control patents are patent applications by firms that were filed in the same year as the treatment patent and have the same technology (IPC) code, but do not cite the university patent that is cited by the treatment patent. The idea is that since a university patent is publicly available, the ideas in it could be used by anyone, so the probability to cite the patent is *a priori* equal for each potential inventor working in the same technological field and in the same time period. The analysis in section 3 tries to find out whether the probability that a university patent is cited is related to the (physical) distance between a firm and the university. This approach for constructing a (synthetic) control group follows the recent literature on spillovers and patent applications (see e.g. Belenzon and Schankerman (2013)).

Many patents have more than one IPC code attached to it by the EPO. We match on the exact combination of all IPC codes attached to a patent in the treatment group. A patent application in the treatment group can have multiple control patent applications in the control group. Patent applications are never used more than once in the control group. If multiple control patent applications are available for a treated patent application, all[11] available control patents are used and each of these control patents receives a weight equal to one dived by the number of control patents available for this particular treated patent application. The sum of the weights therefore always equals one.[12]

The IPC coding system is a hierarchical classification system. Every lower hierarchical level is a subdivision of the upper hierarchical level. It is therefore possible to use the codes at different levels of precision, by not using the lowest hierarchical level(s). An IPC code is built up as follows: the first digit (letters A-H) indicates the section, the next two digits (numbers) the class, the fourth digit (a letter again) the subclass. One to three additional digits (followed by an oblique stroke) indicate the main group of the patent application. The lowest level of aggregation (and hence the highest level of precision) is the subgroup-level, indicated by one to four digits after the oblique stroke.[13] In selecting the control group, the level of precision of the IPC code that is used to find a proper control for each patent in the treatment group is subject to a trade-off. Using a high level of precision, the matched patents are more alike, but it is harder to find an exact match on all the IPC codes assigned to the patent. The set of patents for which a match can be found might be a selective subsample (e.g. more frequently from technologies with high volumes of patents) and therefore not representative for the entire group of university patents. A lower level of precision makes it easier to find matches, but the matched patents may be less comparable, which could introduce measurement error. Resarch using US patent data shows that localization patterns seem to become less prevalent if patents are matched on a higher level of precision rather than a lower level of precision (Thompson and Fox-Kean, 2005b,a; Henderson et al., 2005).

[11]If two or more treated patent applications have exactly the same year-IPC codes-combination, and hence the same set of possible controls, the control patent applications are randomly allocated over the treated patent applications.

[12]This also ensures that across the sample the weighted probability to cite a university patent is 0.5, which helps in the interpretation of the results.

[13]Two examples to clarify the IPC code hierarchy:
code A61K 038/28 refers to section A 'Human necessities', class 61 'Medical or veterinary science; Hygiene', subclass K 'Preparations for medical, dental, or toilet purposes', main group level 38 'Medicinal preparations containing peptides' and subgroup 28 'Insulines'.
Code F16H 061/20 refers to section F 'Mechanical engeneering; Lighting; Heating; Weapons; Blasting', class 16 'Engeneering elements or units; General measures for producing and maintaining effective functioning of machines or installations; Thermal insulation in general', subclass H 'Gearing', main group 61 'Control functions within change-speed- or reversing-gearings for conveying rotary motion' and subgroup 20 'Preventing gear creeping'.

Table 1 highlights the trade-off. The first row shows the maximum number of IPC codes we have for one patent in our treatment group. If we match on the subclass level, we have at most 11 (15) different IPC codes, while if we match on the main group level we have at most 33 (34) different codes for our sample with universities (other research organizations). Since we ideally would like to use all the available information in a patent's technology classes to match, the requirements are much stricter using the main group level than when using the class or subclass level. If we want to use at least 95% of all observations from our sample for estimation, we need to match on upto 8 (9) IPC-codes if we want to use the main group level as level of precision. To increase the percentage of observations used, we quickly need to use much more IPC codes. If we want to increase it to 99%, of the sample, we need to use 14 (15) if we want to match on the main group level and 7 if we want to match on the subclass level. For our main analysis we use the main group IPC level to construct our control group, since this gives us a higher level of precision. This level is roughly comparable in terms of precision with the three-digit USPTO codes commonly used in the literature (Jaffe et al., 1993; Belenzon and Schankerman, 2013). We use upto 9 IPC-codes atached to a patent in the matching procedure to find a control patent. This implies that for more than 95% of the patent we match on the full set of IPC codes assigned to the patent.

Table 1: Trade-off between level of detail in matching and number of matches.

IPC level of precision	Subclass level		Main group level	
	Universities	Research organisations	Universities	Research organisations
of all patent citations that cite university/other research organisation				
Max. no. of IPC codes	11	15	33	34
Number of codes necessary to use at least 95%	5	5	8	9
Number of codes necessary to use at least 99%	7	7	14	15

Source: Own calculations using patent citation data from OECD.

2.3 Patent information

To obtain a patent, an application has to be filed at a patent office. The application includes a description of the invention and a list of the novel parts or characteristics of the invention, over which the applicant wants to claim the exclusive rights. Subsequently, an examiner of the patent office investigates the patent and its claims on novelty and decides whether the patent will be granted.

One of the most important differences between a patent application at the EPO (the applications we study) and applications at the US Patent and Trademark Office (USPTO) is that the USPTO legally requires the applicant to supply a list of all prior art he is aware of that might be relevant for patentability. At the EPO, in contrast, such a list is not required and the examiners are responsible for creating a list of prior art in the process of investigating the patentability. If information on prior art is supplied by the applicant, the examiner determines whether to include it in the search report or not. The examiners of the EPO try to minimize the number of citations and only cite the most important and earliest (if equally important) references. This procedural difference between the EPO and the USPTO results in a large difference in the number of citations. On average, a patent filed at the EPO between 1991 and 2001 cited 5.0 pieces of prior art, a patent filed at the USPTO cited on average 13.7 pieces or prior art (OECD, 2009). [14]

Table A1 in the Appendix provides descriptive statistics of the available information in PATSTAT on the source and origin of citations in our treatment group. 76% of the citations in the university sample and 88% of the citations in the other research organizations sample were added by the examiner during the search. These citations may include references already provided by the applicant. However, whether a reference was already provided by the applicant is not indicated in the PATSTAT database for citations that are used in the search report. References provided by the applicant that the examiner decided not to use in the report are listed separately. These account 20% of the citations for the university sample and 88% of the citations for the other research organizations sample in the PATSTAT database.

The examiner is also responsible for assigning International Patent Classification (IPC) codes to the patent. A patent can have multiple IPC codes assigned to it. The EPO does not create a hierarchy in the codes if multiple codes are assigned to a patent.

2.4 Calculating distances

To calculate the distance between a citing and cited patent, we start by geocoding the address of the citing and cited patent to obtain the longitude and latitude of each address.[15] The longitude and latitude of the citing and cited patent are subsequently used in Vincenty's formulas to obtain the distance between the two points on the globe (Vincenty, 1975).[16] Each patent has both an inventor and an applicant attached to it, both of them with an address. The inventor is an individual, the applicant usually the employer (company, government, etc) of the inventor. With large firms the applicant may be the headquarter of the firm, possibly far away from the actual workplace of the inventor. For our purposes, where we want to investigate the spillovers from universities and research organisations using the distance between inventors and universities with a relevant patent, it is therefore preferred to use the address of the inventor instead of the address of the applicant. This difference is less relevant for the university patents, as a university usually only has locations in a single city.

2.5 Estimation sample and descriptive statistics

The PATSTAT database contains 3626 patent citations that cite a patent from a university and 16730 patents citations that cite a patent from other research organizations. However, not all of these observations can be included in our estimation sample as treatment group. Table 2 shows how we get from the original PATSTAT database to our estimation sample. More than 70% of the original number of patent citations cannot be used. The larger part of this 70% is due to no address information being available for the citing patent and hence no usable control can be found (second row in Table 2). A small number of observations is excluded because they are self-citations. Of the patents for which the address information *is* available for both the cited and the citing patent, we lose observations again

[14] The EPO citation procedure has some consequences for the interpretation of our results. We discuss them in detail in section 3.

[15] We use the geocoding facility provided by www.gpsvisualizer.com with an account from Bing Maps.

[16] We use the Stata command by Austin Nichols, see https://ideas.repec.org/c/boc/bocode/s456815.html.

in the process of geocoding the address of both the citing and the cited patent (third row in Table 2). In these cases the geocoding software was unable to find the address of the cited or the citing patent. Finally, we are left with 20% of the original sample of universities and 15% of the original sample of other research organizations. A patent may have more than one inventor or applicant. This implies that we may have multiple distances for the same patent citation. As we are after estimating the effect of geographical distance on the probability to cite a patent, it is the shortest distance of this set of distances than the average distance that is relevant. Only one of the inventors/applicants needs to be close to the inventor/applicant of another patent to (possibly) create a spillover. When restricting ourselves to use only one distance between citing and cited patent (the shortest) for every patent citation, we get to our final number of observations in the treatment group (fourth row in Table 2).

Table 2: Selection of treatment group based on main group IPC level.

	Universities	Research organisations
All patent citations that cite university or other research organisation	3626	16730
With at least 1 usable* control available	1039	4300
Where distance between cited and citing can be calculated	760	2447
If include only shortest distance between inventors of citing and cited patent	292	1092

Notes: *=usable control has at least some address information for citing patent.
Source: Own calculations using patent citation data from OECD.

Table 3 gives descriptive statistics of the two samples used for estimation, the sample with only universities and the sample with other research organizations (excluding the universities). Both are based on matching on main group level IPC codes. The sample with only universities has 292 citations in the treatment group and 1682 in the control group. On average, every treated patent in this sample has 5.8 control patents. In the sample with other research organizations there are 3832 controls for 1092 citations in the treatment group, on average 3.5 controls per treated. The descriptive statistics are not weighted with the weights given to the patents in the control group (for a description of the weights see section 2.2). The distance between most citing and cited patents is quite large, 85% lies more than 200 kilometers apart. Since we expect that an effect, if present, will be at the very short distances, we choose to use a finer grid for the shorter distances than for the long distances. As sensitivity check we also used different distance categorizations, for example also using a finer grid for the long distances, but the coefficients in the estimation results were very much the same. Many patent applications have their origin in Germany (47% and 46% in the university and other research organizations sample respectively), followed by France (15% and 17% in the university and other research organizations sample respectively). Most patents are in the 'Human necessities' technology section, including agriculture and food, and the 'Physics' section.

Table 3: Descriptive statistics for estimation samples based on matching on main group level.

	Universities	Research organisations
Number of observations	1974	4924
of which in treatment group	292	1092
of which in control group	1682	3832
Distance		
$0-5$km	0.017	0.017
$5-25$km	0.016	0.033
$25-100$km	0.035	0.039
$100-200$km	0.073	0.065
$200-400$km	0.217	0.237
$400-800$km	0.399	0.450
$800-1200$km	0.187	0.111
≥ 1200 km	0.056	0.048
Citing and cited from the same country	0.180	0.267
Country of the citing patent		
Austria	0.023	0.021
Belgium	0.037	0.027
Switzerland	0.082	0.069
Denmark	0.013	0.006
Spain	0.017	0.009
Finland	0.019	0.006
France	0.150	0.169
United Kingdom	0.082	0.079
Greece	0.001	0.000
Ireland	0.004	0.001
Italy	0.019	0.068
Luxemburg	0.000	0.001
Netherlands	0.053	0.065
Norway	0.004	0.002
Portugal	0.001	0.000
Sweden	0.022	0.020
Germany	0.475	0.456
IPC section level		
A Human necessities	0.421	0.256
B Performing operations/transporting	0.044	0.123
C Chemistry/metallurgy	0.112	0.112
D Textiles/paper	0.004	0.005
E Fixed constructions	0.003	0.031
F Mech.engineering/lighting/heating/..	0.068	0.075
G Physics	0.206	0.304
H Electricity	0.183	0.155

Notes: All variables are indicator variables, hence the reported means can be interpreted as fractions. Statistics are not weighted with the weights given to the citations in the control group.
Source: Own calculations using patent citation data from OECD.

3 Estimation strategy

On the matched set of treatment and control patent citations, we follow Belenzon and Schankerman (2013) and estimate the following linear probability model

$$C_{i(u),j} = \alpha D_{i,j} + \beta X_{i,j} + \epsilon_{i,j} \quad (1)$$

where $C_{i(u),j}$ is an indicator with value 1 if patent j from a firm cites a patent i from a university u and 0 if the firm patent does not cite a university patent. $D_{i,j}$ is a vector with dummies specifying a flexible function for distance to the cited university and $X_{i,j}$ is a vector with control variables. In our main specification we include controls for whether the patent is cited within the same country, country fixed effects and interactions of country fixed effects and a dummy variable indicating whether the citation is within the same country. Standard errors are clustered at the level of the cited patent to take into account correlation in observations that cite the same patent. For example, some patents might be cited more often than others, because of their higher quality.

Due to our matching methodology we already take into account any differences between technologies and time periods. For example, in some technologies it is more common to cite patents than in others. The matching on technology also takes into account that production, and hence innovation, might be geographically concentrated for other reasons. There might be natural advantages, such as the availability of water or oil, or social advantages, such as the availability of a large market or a specialized university with many graduates in one field. One would expect that firms working in the same technological fields would take such advantages into account and hence locate together for reasons other than profiting from knowledge spillovers from a university.

As discussed in section 2.3, a potential concern with our data is that patent citations in Europe, contrary to the U.S., are predominantly added by examiners rather than applicants (76% and 88% in our estimation samples).Citations introduced by the examiner could introduce measurement error, because citations might not reflect an actual knowledge transfer. Examiners might have a broader view of the state of patents related to a given technology, which means they could more frequently add citations with a larger 'distance' to a patent. On the other hand, examiners are more objective evaluators and therefore only include citations to patents of which the knowledge is actually used. At any rate, it seems likely that patents cited by examiners are not more localized than the control patents (which, remember, are in the same technological field and from the same year as the citing patents). Hence, examiner citations introduce a bias towards finding no localization. If anything, this means that, if we still find an effect of distance on the probability that a university patent is cited, we underestimate the actual knowledge spillovers that took place (see Jaffe et al. (1993) for a similar argument). The empirical evidence regarding examiner versus inventor citations shows mixed results. Some papers find no differences regarding geographical distance between examiner and inventor citations (Alcácer and Gittelman, 2006), while others find that inventor citations are geographically closer than examiner citations (Thompson, 2006; Criscuolo and Verspagen, 2008).

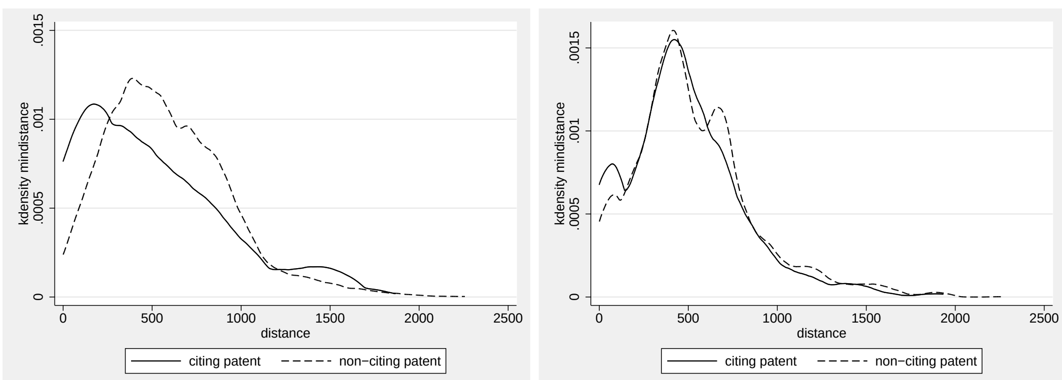
4 Results

In this section we present our main results. We start with non-parametric evidence showing the relationship between the probability to cite a university patent and distance. We continue with our econometric analysis where we take into account the role of country borders and within-country citations. Finally, we show how the effect of distance on citation probability changes over the 'age' of a patent.

4.1 Non-parametric evidence

Figure 1 shows kernel density plots for the distance from cited universities and other research organizations for the treatment group of citing patents and the control group of non-citing patents. The graphs show that the probability mass of the citing patents is higher at the lower distances, while the probability mass of the non-citing patents is higher at the medium to longer distances. At the very long distances (more than 1,000 kilometers) there are hardly any differences between the citing and non-citing patents. This provides some suggestive evidence that the citation probability declines with distance, but these figures could be driven by other factors, such as within-country citations. We will take this into account in our econometric analyses.

Figure 1: Kernel density plot of citing and non-citing patents and their distance to the cited organization.



(a) Universities

(b) Other research organizations

Source: Own calculations based on the OECD Patents Database.

4.2 Main econometric analysis

Table 4 shows the baseline results from estimating linear probability models for our main sample of university patents matched at the IPC main group level. The first specification without any control variables shows a clear declining pattern of citation probabilities with distance to the university. Moving from 0-5 to 5-25 kilometers leads to a small and statistically insignificant drop-off in the citation probability of 9.9%. Moving from 5-25 to 25-100 kilometers, however, leads to a substantial drop-off of 18.7% (28.6 − 9.9). The weighted mean citation rate is 50%.[17] Hence, the drop off in moving from 5-25 to 25-100 is 37.4% of the mean citation rate (18.7/50), which is substantial. There is again a sharp decline in the citation probability if we move from 100-200 to 200-400 kilometers. For longer distances the coefficients are similar, so distance doesn't seem to play an important role anymore.[18]

The specification in column (2) includes a control variable for whether the citation is in the same country. The coefficient on this dummy is quite large (about 25% of the mean citation rate) and shows that university patents are more likely to be cited within the same country. The coefficients on distances larger than 25 kilometers become substantially smaller. The additional drop in the citation probability from e.g. 100-200 to 200-400 kilometers is now only 9% instead of 15%. Since citation patterns might differ between countries due to for example national policies, it is important to control for country effects as well. In column (3) we include country fixed effects and a set of interactions between the country fixed effects and the dummy for whether a citation is within the same country. This leads to smaller coefficients overall. However, the sharp drop off at 25 kilometers remains and actually increases somewhat to around 41.0% of the mean citation rate ($(28.5-8.0)/50$). The coefficient on the citation within the same country dummy also becomes smaller (about 16% of the mean citation rate) and is only statistically significant at the 10% level. The country effects and interactions likely pick up much of the variation for the same country dummy. Adding technology controls in column (4) doesn't seem to have a substantial effect on the coefficients. The coefficient on the same country dummy remains similar, and again only significant at 10% level. We take the specification of column (4) as our main specification.

These results show two important findings. First, distance plays an important role in the probability that a university patent is cited. There is a sharp cutoff point at around 25 kilometers. Second, a substantial part of the effect of longer distances is mitigated by country borders. This could point to the importance of country-specific policies, such as specialization in some fields, or to the existence of language and institutional barriers making communication between researchers more difficult.

Table 5 shows the results from estimating the same models using all other research organizations. Overall, the pattern suggests that distance plays a smaller role for other research organizations than for universities. Once we include our full set of country controls in column (3), almost none of the distance parameters are significant. The estimated coefficients are small compared to the results for universities and quite similar over the whole range. The indicator variable for citations within the same country remains quite large (about 23% of the mean citation rate) and significant throughout.

The question remains whether the differences between universities and all other research organizations are significant. In Table 6 we show estimates where we use the full sample, but include interaction terms between all variables and a dummy for whether a cited patent originates from a research organization other than a university. The distance terms without interactions, which effectively give the effects of distance on the citation probability of university patents, show significant negative estimates in line with the results in Table 4. The interaction terms are insignificant for distances

[17]For each citing patent we have at least one non-citing patent in the control group. The patents in the control group receive equal weight such that their weight sums up to 1.

[18]We used a range of alternative distance categories to investigate the sensitivity of the results to the grouping of distances. Different categorization yields the same conclusion of a cut-off around 25 kilometers.

up to 400km after including all our controls. This means that there is no statistically significant difference between universities and all other research organizations as to how distance affects their citation probability for the shorter distances. So while we do find smaller point estimates for all other research organizations, the uncertainty around these estimates is too large to reject the hypothesis that the citation probability is similarly affected by distance for universities and all other research organizations.

These results suggest two conclusions. First, we find that physical distance plays an important role in determining the citation probability for universities. For other research organizations we find smaller effects of distance, but due to uncertainty around these estimates, we can't reject the hypothesis that the patterns are actually the same up to about 400 kilometers. Second, country borders seem to play an important role for both universities and other research organizations.

Table 4: Main results for universities.

	(1)	(2)	(3)	(4)
5-25km	-0.0974	-0.0974	-0.0798	-0.0591
	(0.0733)	(0.0733)	(0.0847)	(0.0856)
25-100km	-0.2849^{**}	-0.2470^{**}	-0.2869^{***}	-0.2827^{***}
	(0.1099)	(0.1023)	(0.1058)	(0.1056)
100-200km	-0.3006^{***}	-0.2581^{***}	-0.2316^{***}	-0.2229^{***}
	(0.0733)	(0.0720)	(0.0852)	(0.0848)
200-400km	-0.4390^{***}	-0.3451^{***}	-0.2731^{***}	-0.2624^{***}
	(0.0648)	(0.0781)	(0.0966)	(0.0975)
400-800km	-0.4897^{***}	-0.3811^{***}	-0.3239^{***}	-0.3136^{***}
	(0.0529)	(0.0676)	(0.0861)	(0.0875)
800-1200km	-0.5089^{***}	-0.3850^{***}	-0.3120^{***}	-0.3039^{***}
	(0.0553)	(0.0724)	(0.0894)	(0.0921)
≥ 1200	-0.3373^{***}	-0.2128^{**}	-0.2685^{**}	-0.2544^{**}
	(0.0826)	(0.1016)	(0.1144)	(0.1152)
citing and cited from same country		0.1245^{**}	0.0796^{*}	0.0867^{*}
		(0.0498)	(0.0448)	(0.0479)
Intercept	0.9174^{***}	0.7929^{***}	0.7228^{***}	0.7157^{***}
	(0.0430)	(0.0649)	(0.0849)	(0.0874)
Country controls	No	No	Yes	Yes
Same country X country	No	No	Yes	Yes
Technology controls	No	No	No	Yes
R^2	0.0668	0.0736	0.1131	0.1150
Observations	1974	1974	1974	1974

Notes: Main sample with patents matched at the main group IPC level. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

Table 5: Main results for all public research organizations.

	(1)	(2)	(3)	(4)
5-25km	-0.0560	-0.0527	-0.0230	-0.0218
	(0.0685)	(0.0683)	(0.0776)	(0.0778)
25-100km	-0.1339^{*}	-0.1265^{*}	-0.0911	-0.0883
	(0.0685)	(0.0684)	(0.0770)	(0.0775)
100-200km	-0.1740^{**}	-0.1381^{**}	-0.0862	-0.0848
	(0.0682)	(0.0682)	(0.0762)	(0.0764)
200-400km	-0.2145^{***}	-0.1407^{**}	-0.0844	-0.0833
	(0.0589)	(0.0589)	(0.0682)	(0.0684)
400-800km	-0.2139^{***}	-0.1057^{*}	-0.0354	-0.0344
	(0.0568)	(0.0581)	(0.0672)	(0.0675)
800-1200km	-0.2379^{***}	-0.1109^{*}	-0.0424	-0.0415
	(0.0619)	(0.0637)	(0.0738)	(0.0739)
$\geq$ 1200	-0.3386^{***}	-0.2140^{***}	-0.1411^{*}	-0.1391^{*}
	(0.0674)	(0.0681)	(0.0790)	(0.0794)
citing and cited from same country		0.1271^{***}	0.1098^{***}	0.1108^{***}
		(0.0229)	(0.0268)	(0.0271)
Intercept	0.7181^{***}	0.5910^{***}	0.5101^{***}	0.5100^{***}
	(0.0549)	(0.0571)	(0.0683)	(0.0689)
Country controls	No	No	Yes	Yes
Same country X country	No	No	Yes	Yes
Technology controls	No	No	No	Yes
R^2	0.0121	0.0205	0.0324	0.0325
Observations	4925	4925	4925	4925

Notes: Main sample with patents matched at the main group IPC level. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

Table 6: Results for single model with both universities and research organisations.

	(1)	(2)	(3)	(4)
5−25km	0.0305	0.0305	0.0577	0.0591
	(0.0497)	(0.0498)	(0.0688)	(0.0720)
25−100km	−0.2008**	−0.1723*	−0.2121**	−0.2110**
	(0.0952)	(0.0924)	(0.0992)	(0.0990)
100−200km	−0.2896***	−0.2492***	−0.1930**	−0.1914**
	(0.0726)	(0.0729)	(0.0813)	(0.0810)
200−400km	−0.4408***	−0.3560***	−0.2544***	−0.2538***
	(0.0616)	(0.0753)	(0.0918)	(0.0923)
400−800km	−0.4782***	−0.3783***	−0.2913***	−0.2910***
	(0.0524)	(0.0696)	(0.0855)	(0.0865)
800−1200km	−0.4981***	−0.3844***	−0.2776***	−0.2778***
	(0.0557)	(0.0738)	(0.0889)	(0.0918)
≥1200km	−0.3484***	−0.2346**	−0.2139*	−0.2099*
	(0.0806)	(0.1012)	(0.1189)	(0.1204)
5−25km ∗ researchorg	−0.1030	−0.0998	−0.1052	−0.1051
	(0.0837)	(0.0835)	(0.1029)	(0.1053)
25−100km ∗ researchorg	0.0424	0.0209	0.0860	0.0876
	(0.1162)	(0.1138)	(0.1251)	(0.1252)
100−200km ∗ researchorg	0.1072	0.1014	0.0920	0.0916
	(0.0992)	(0.0995)	(0.1112)	(0.1111)
200−400km ∗ researchorg	0.2068**	0.1940**	0.1404	0.1406
	(0.0844)	(0.0949)	(0.1137)	(0.1143)
400−800km ∗ researchorg	0.2450***	0.2507***	0.2297**	0.2303**
	(0.0764)	(0.0898)	(0.1081)	(0.1091)
800−1200km ∗ researchorg	0.2357***	0.2459**	0.2016*	0.2026*
	(0.0825)	(0.0969)	(0.1151)	(0.1175)
≥1200km ∗ researchorg	−0.0187	−0.0111	0.0341	0.0316
	(0.1044)	(0.1215)	(0.1422)	(0.1436)
Citing and cited from same country		0.1138**	0.0834	0.0831
		(0.0519)	(0.0513)	(0.0538)
Ctng&ctd from same country ∗ resorg		0.0100	0.0286	0.0296
		(0.0569)	(0.0579)	(0.0604)
Research organsation	−0.1827***	−0.1927**	−0.1587	−0.1630
	(0.0680)	(0.0863)	(0.1077)	(0.1102)
Intercept	0.9214***	0.8076***	0.6950***	0.6997***
	(0.0416)	(0.0656)	(0.0839)	(0.0866)
Country controls	No	No	Yes	Yes
Same country X country	No	No	Yes	Yes
Technology controls	No	No	No	Yes
R^2	0.0277	0.0352	0.0519	0.0523
Observations	6242	6242	6242	6242

Notes: Main sample with patents matched at the main group IPC level. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

4.3 Citation patterns over time

To investigate differences in the effect of distance over time, the sample is split into two subsamples. A subsample of patent citations for which the time between publication of the citing patent and publication date of the cited patent, the *citation lag*, is less than four years and a subsample where the citation lag is more than four years. Table 7 shows for the university sample how the effect of distance varies between the two subsamples. In the first column the main results from subsection 4.2 are shown for comparison. The results in the second column refer to the subsample with a citation lag of less than four years, the results in the third column to the sample with a citation lag of more than four years. About one-third of citations has a citation lag smaller than four years. The role of physical distance found in subsection 4.2 seems to be entirely driven by citations with a short citation lag. For this subsample the coefficients for longer distances are substantially higher than those from the full sample, and also strongly significant. For the subsample with long citation lags on the other hand the effect of physical distance becomes smaller and insignificant.[19]

Using an alternative cut-off point instead of four years for creation of the subsamples gives qualitatively similar results. Physical distance is important for citing patents shortly after they are filed. Table 8 shows how the estimated coefficients for the effect of distance change in the subsample with short citation lag when the cutoff point for the citation lag is varied. If the cutoff point is decreased to two years (column (1)), the number of observations drops substantially to only 133 but the estimated coefficients for the distance variables become even larger (they more than double compared to the main results). Due to the small number of observations the coefficient for 25-100km is no longer significant. Columns (2)-(4) show that the effect of physical distance becomes smaller when the cut-off point for the citation lag increases. This indicates that the effect of physical distance is very strong short after the publication date of the patent and that adding observations with somewhat longer citation lags to the subsample only dilutes the effect of physical distance from the patents with a short citation lag.

The patterns we observe are consistent with the diffusion of knowledge spillovers over time. The specific pattern of spillovers over time and distance could be explained by a story where spin-offs by university researchers develop a commercial application of their invention. Spin-off companies are typically started quite close to the university, and the information advantage the inventors have over others helps them to develop a commercial application. The spin-off company will apply for a patent for their application as well, and this generates the pattern where early citations are very close to the university, while later citations don't depend on distance as much. If commercial spin-offs are an important driver of the spillover effect observed from university patents, it is most likely that these spillovers occur relatively short after the patent is filed. In this case the pattern with distance should be strongest short after the patent is filed and diminish when time proceeds. This is indeed what we find. Unfortunately, our data don't allow us to examine this mechanism in detail. This would be a very interesting avenue for further research.

[19] In Table A2 in the Appendix we show the same estimates for all public research organizations. Consistent with our main results we find no effect of distance on citation probability for public research organizations but a large effect for whether a patent is cited within the same country.

Table 7: Difference over time in distance pattern for the university sample.

	(1) Main results	(2) $\leq$ 4 years	(3) $>$ 4 years
5-25km	-0.0591	-0.0172	-0.1415
	(0.0856)	(0.0837)	(0.1081)
25-100km	-0.2827^{***}	-0.2263^{*}	-0.3092^{**}
	(0.1056)	(0.1185)	(0.1412)
100-200km	-0.2229^{***}	-0.3834^{***}	-0.0939
	(0.0848)	(0.1150)	(0.1173)
200-400km	-0.2624^{***}	-0.5057^{***}	-0.0844
	(0.0975)	(0.1140)	(0.1327)
400-800km	-0.3136^{***}	-0.4545^{***}	-0.1855
	(0.0875)	(0.1235)	(0.1166)
800-1200km	-0.3039^{***}	-0.5429^{***}	-0.1330
	(0.0921)	(0.1313)	(0.1222)
$\geq$ 1200	-0.2544^{**}	-0.3379^{*}	-0.2497
	(0.1152)	(0.1738)	(0.1567)
Citing and cited from same country	0.0867^{*}	0.0124	0.0950
	(0.0479)	(0.0846)	(0.0679)
Intercept	0.7157^{***}	0.8687^{***}	0.5836^{***}
	(0.0874)	(0.1329)	(0.1101)
Country controls	Yes	Yes	Yes
Same country X country	Yes	Yes	Yes
Technology controls	Yes	Yes	Yes
R^2	0.1150	0.1968	0.1185
Observations	1974	694	1280

Notes: Main sample with patents matched at the main group IPC level. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

Table 8: Distance effect for patent citations within 2, 3, 4 and 5 years after publication for the university sample.

	(1) $\leq$ 2 years	(2) $\leq$ 3 years	(3) $\leq$ 4 years	(4) $\leq$ 5 years
5-25km	-0.1468	-0.0957	-0.0172	0.1117
	(0.1527)	(0.0908)	(0.0837)	(0.1113)
25-100km	-0.3790	-0.2291	-0.2263^{*}	-0.1602
	(0.2814)	(0.2291)	(0.1185)	(0.1511)
100-200km	-0.5953^{**}	-0.5209^{***}	-0.3834^{***}	-0.2760^{**}
	(0.2429)	(0.1605)	(0.1150)	(0.1365)
200-400km	-0.7065^{***}	-0.5906^{***}	-0.5057^{***}	-0.3218^{**}
	(0.2157)	(0.1354)	(0.1140)	(0.1328)
400-800km	-0.6284^{**}	-0.5092^{***}	-0.4545^{***}	-0.3729^{***}
	(0.2247)	(0.1288)	(0.1235)	(0.1330)
800-1200km	-0.8410^{***}	-0.6177^{***}	-0.5429^{***}	-0.4310^{***}
	(0.2857)	(0.1564)	(0.1313)	(0.1341)
$\geq$ 1200	-0.1286	-0.6194^{***}	-0.3379^{*}	-0.1776
	(0.3227)	(0.2176)	(0.1738)	(0.1742)
Citing and cited from same country	-0.0895	-0.0143	0.0124	0.0266
	(0.2408)	(0.1353)	(0.0846)	(0.0708)
Intercept	1.2329^{***}	0.9606^{***}	0.8687^{***}	0.7709^{***}
	(0.2550)	(0.1522)	(0.1329)	(0.1309)
Country controls	Yes	Yes	Yes	Yes
Same country X country	Yes	Yes	Yes	Yes
Technology controls	Yes	Yes	Yes	Yes
R^2	0.2764	0.2089	0.1968	0.1500
Observations	133	325	694	942

Notes: Main sample with patents matched at the main group IPC level. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

5 Sensitivity analyses

In this section we show several sensitivity analyses that highlight the robustness of our main results. First, we check whether our matching procedure affects our results. We match on multiple IPC codes at the main group level, so we lose a lot of patents that don't have a match at such a detailed level. We examine whether matching on the less detailed subclass level affects our results.

Second, we examine whether our effects remain robust when excluding the two largest countries, Germany and France, from our sample. Finally, we check whether the advent of ICT has affected the effect of physical distance on the citation probability by comparing localization patterns before and after 2000.

5.1 Matching

As discussed in section 2.2, the level of precision we use for matching is subject to a trade-off. On the one hand we get more precision if we match on the main group level, while on the other hand we lose a lot of treated observations because there are no suitable controls. Furthermore, the literature for the US shows that localization patterns are smaller if patents are matched on detailed six-digit USPTO codes rather than three digit codes (Thompson and Fox-Kean, 2005a,b; Henderson et al., 2005). It is therefore instructive to compare the results obtained using a detailed matching procedure with a more coarse procedure. Table 9 shows the estimation results if we match on IPC subclass level and compares this to our main results obtained with matching on IPC main group level.

Column (1) reproduces the main results for the university sample using the IPC main group level for matching. Column (2) presents the estimation results using the IPC subclass level. The number of observations increases, because the matching is less precise. The coefficients for the sample matched on the subclass level are in almost all cases larger than for the sample matched on the main group level. This is in line with the literature - weaker localization patterns when matching on more precise technology fields -, but the overall pattern is actually quite similar. There is still a sharp drop-off at 25 kilometers, and distance seems to play almost no additional role after 200 kilometers. Column (3) reproduces the main results for other research organizations and column (4) presents the estimation results using subclass level matching. Similar to the university sample we find almost consistently larger coefficients when matching on the subclass level. All coefficients are now statistically significant as well.[20] These findings suggest that there is some heterogeneity remaining in the less detailed subclass level, and that matching on the main group level is more appropriate.

5.2 Other robustness checks

We also perform several other sensitivity analyses on subsamples. Our sample size becomes quite small if we start using subsamples, so the results below may suffer from power issues.

As discussed in section 2, citations to patents in Germany and to a lesser extent France make up the bulk of the sample. These are also the two largest countries in the sample, and given the importance of country effects, it would be instructive to see to what extent the results are driven by

[20]Similar to the main results we also find that when mating on the subclass level the localization patterns for citations to university patents are much stronger than for other research organizations.

Table 9: Parameter estimates comparing matching on the IPC main group and IPC subclass level for both universities and other research organizations.

	(1)	(2)	(3)	(4)
	Universities		Research organizations	
	Main group	Subclass	Main group	Subclass
5−25km	−0.0584	−0.0849	−0.0270	−0.0643*
	(0.0875)	(0.0565)	(0.0781)	(0.0391)
25−100km	−0.2801***	−0.2475***	−0.0822	−0.1030***
	(0.1057)	(0.0638)	(0.0775)	(0.0398)
100−200km	−0.2124**	−0.3119***	−0.0852	−0.1195***
	(0.0826)	(0.0550)	(0.0762)	(0.0396)
200−400km	−0.2570***	−0.3504***	−0.0818	−0.1303***
	(0.0970)	(0.0617)	(0.0684)	(0.0344)
400−800km	−0.3066***	−0.3337***	−0.0333	−0.1224***
	(0.0870)	(0.0560)	(0.0674)	(0.0345)
800−1200km	−0.2973***	−0.3340***	−0.0403	−0.1282***
	(0.0914)	(0.0576)	(0.0739)	(0.0392)
≥1200km	−0.2478**	−0.3950***	−0.1328*	−0.1291***
	(0.1147)	(0.0727)	(0.0795)	(0.0497)
Citing and cited from same country	0.0876*	0.0638*	0.1123***	0.1172***
	(0.0478)	(0.0381)	(0.0270)	(0.0173)
Intercept	0.7094***	0.7799***	0.5084***	0.5557***
	(0.0870)	(0.0587)	(0.0688)	(0.0367)
Country controls	Yes	Yes	Yes	Yes
Same country X country	Yes	Yes	Yes	Yes
Technology controls	Yes	Yes	Yes	Yes
R^2	0.1146	0.1017	0.0331	0.0383
Observations	1974	8796	4924	27556

Notes: Columns (1) and (3) reproduce the main estimates from column (4) of Table 4 and Table 5, respectively. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

Table 10: Results of several sensitivity analyses for the university sample.

	(1) Main results	(2) No FR and DE	(3) Pre-2000	(4) Post-2000
5-25km	-0.0591	-0.1511	-0.0394	-0.1123
	(0.0856)	(0.1011)	(0.1139)	(0.0973)
25-100km	-0.2827^{***}	-0.2935^{**}	-0.3077	-0.3013^{**}
	(0.1056)	(0.1253)	(0.2222)	(0.1196)
100-200km	-0.2229^{***}	-0.2832^{**}	-0.3001^{**}	-0.2235^{**}
	(0.0848)	(0.1412)	(0.1449)	(0.0872)
200-400km	-0.2624^{***}	-0.2920^{*}	-0.2186	-0.3075^{***}
	(0.0975)	(0.1592)	(0.1487)	(0.1109)
400-800km	-0.3136^{***}	-0.3714^{***}	-0.3382^{***}	-0.3376^{***}
	(0.0875)	(0.1397)	(0.1251)	(0.0951)
800-1200km	-0.3039^{***}	-0.2749^{*}	-0.4912^{***}	-0.2925^{***}
	(0.0921)	(0.1466)	(0.1389)	(0.1049)
$\geq$ 1200	-0.2544^{**}	-0.1895	-0.4892^{*}	-0.3011^{**}
	(0.1152)	(0.1791)	(0.2659)	(0.1247)
citing and cited from same country	0.0867^{*}	0.5089	-0.2744^{*}	0.1008^{*}
	(0.0479)	(0.3165)	(0.1562)	(0.0532)
Intercept	0.7157^{***}	0.2940	0.7810^{***}	0.7331^{***}
	(0.0874)	(0.1816)	(0.1376)	(0.0959)
Country controls	Yes	Yes	Yes	Yes
Same country X country	Yes	Yes	Yes	Yes
Technology controls	Yes	Yes	Yes	Yes
R^2	0.1150	0.2315	0.2050	0.1155
Observations	1974	562	519	1455

Notes: Main sample with patents matched at the main group IPC level. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

these two countries. Column (2) in Table 10 shows the estimation results when we exclude France and Germany. The estimates are quite similar to the main results reproduced in column (1). The pattern at the shorter distances is similar: a substantial drop-off in citation probability at 25 kilometers. This confirms that short distances are important. At longer distances the coefficients are very similar, so the results are not driven by the large share of cited patents from France and Germany.

We might expect that with the widespread use of ICT physical distance plays a smaller role than before in the dissemination of knowledge. Griffith et al. (2011) provide evidence that localization patterns have declined over time. We test this hypothesis by estimating our model separately for pre-2000 and post-2000 patents, with the year 2000 somewhat arbitrarily marking the advent of ICT. Column (3) shows the estimates for all pre-2000 patents and column (4) shows the estimates for all post-2000 patents. The pattern is actually very similar to the main results, with a sharp drop-off in citation probability at 25 kilometers and a relatively small role for physical distance after that.

The exception seems to be the coefficients on the very long distances ($\geq$ 800km) for the pre-2000 patents. This suggests that the role of distance has declined somewhat since 2000, since patents at longer distances are more likely to be cited post-2000 than pre-2000. However, physical distance still plays an important role in the post-2000 era, especially on the shorter distances.

6 Discussion and conclusion

In this study we use data on patent citations to examine the extent to which knowledge spillovers from European universities and other research organizations are localized. We find strong evidence for localized knowledge spillovers, in particular for universities. Distance from the university affects the probability to cite a university patent. There is a sharp drop-off in citation probability at 25 kilometers. After taking into account the effects of country borders and country effects we find that longer distances don't seem to affect the citation probability. For other research organizations we find very little evidence of a role of physical distance in citation probabilities. This could be related to the more fundamental nature of university patents, which require more tacit knowledge to be absorbed. However, due to uncertainty around the estimates for other research organizations, we can't reject the hypothesis that their citation probability is actually similarly affected by distance as for universities. National borders seem to play a role in reducing the citation probability for both universities and other research organizations. Our results support the idea that regional clustering is important for the use of knowledge generated at universities.

Our approach to knowledge spillovers is limited, because we only look at patents from universities and other research organizations. These are just one way through which knowledge spillovers manifest themselves. Others include for example scientific publications, students and seminars.[21] Using patent data also necessarily limits us to technical and medical research. However our approach allows us to show how the mechanism of knowledge spillovers operates and to what extent it is constrained by physical distance.

A further line of research that could be interesting is to examine the underlying mechanisms for

[21] See e.g. Belenzon and Schankerman (2013) for a similar analysis for the US that also includes scientific publications. They find for both scientific publications and university patents that distance matters in the citation probabilities.

why distance matters for the spread of knowledge. Our estimates on the diffusion of knowledge over time are consistent with the creation of spin-off companies that try to market an invention at a university. Our data don't allow us to explore this mechanism in more detail, but it would be a very interesting avenue for further research. We show substantial differences in citation patterns between universities and other research organizations. While we offer some explanations, we can't precisely pinpoint the mechanism behind this.

References

Abramovsky, L., R. Harrison, and H. Simpson (2007). University research and the location of business r&d. *The Economic Journal 117*(519), 114–141.

Aghion, P. and P. Howitt (1992). A model of growth through creative destruction. *Econometrica 60*(2), 323–351.

Agrawal, A., I. Cockburn, A. Galasso, and A. Oettl (2014). Why are some regions more innovative than others? The role of small firms in the presence of large labs. *Journal of Urban Economics 81*, 149–165.

Alcácer, J. and M. Gittelman (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics 88*(4), pp. 774–779.

Andersson, R., J. M. Quigley, and M. Wilhelmson (2004). University decentralization as regional policy: the Swedish experiment. *Journal of Economic Geography 4*(4), 371–388.

Andersson, R., J. M. Quigley, and M. Wilhelmsson (2009). Urbanization, productivity, and innovation: Evidence from investment in higher education. *Journal of Urban Economics 66*(1), 2 – 15.

Audretsch, D. B. and M. P. Feldman (1996). R&d spillovers and the geography of innovation and production. *The American Economic Review 86*(3), 630–640.

Audretsch, D. B., E. E. Lehmann, and S. Warning (2005). University spillovers and new firm location. *Research Policy 34*(7), 1113 – 1122.

Belenzon, S. and M. Schankerman (2013). Spreading the word: Geography, policy, and knowledge spillovers. *Review of Economics and Statistics 95*(3), 884–903.

Criscuolo, P. and B. Verspagen (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in european patents. *Research Policy 37*(10), 1892 – 1908.

Drucker, J. and H. Goldstein (2007). Assessing the regional economic development impacts of universities: A review of current approaches. *International Regional Science Review 30*(1), 20–46.

Griffith, R., S. Lee, and J. Van Reenen (2011). Is distance dying at last? Falling home bias in fixed-effects models of patent citations. *Quantitative Economics 2*(2), 211–249.

Henderson, R., A. Jaffe, and M. Trajtenberg (2005). Patent citations and the geography of knowledge spillovers: A reassessment: Comment. *The American Economic Review 95*(1), pp. 461–464.

Jaffe, A. B., M. Trajtenberg, and R. Henderson (1993). Geographic localization of knowledge spillovers as eevidence by patent citations. *The Quarterly Journal of Economics 108*(3), 577–598.

Kantor, S. and A. Whalley (2014, March). Knowledge spillovers from research universities: Evidence from endowment value shocks. *The Review of Economics and Statistics 96*(1), 171–188.

Kiseleva, T., A. Palali, and B. Straathof (2016). Do national borders slow down knowledge diffusion within new technological fields? The case of big data in Europe. *CPB Discussion Paper 330*.

Leten, B., P. Landoni, and B. Van Looy (2014). Science or graduates: How do firms benefit from the proximity of universities?

Liu, S. (2015). Spillovers from universities: Evidence from the land-grant program. *Journal of Urban Economics 87*, 25–41.

Maurseth, P. B. and B. Verspagen (2002). Knowledge spillovers in Europe: A patent citations analysis. *Scandinavian Journal of Economics 104*(4), 531–545.

OECD (2009). OECD patent statistics manual. Technical report, OECD.

Peri, G. (2005). Determinants of knowledge flows and their effect on innovation. *Review of Economics and Statistics 87*(2), 308–322.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy 94*(5), 1002–1037.

Rosenthal, S. S. and W. C. Strange (2008). The attenuation of human capital spillovers. *Journal of Urban Economics 64*(2), 373 – 389.

Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: Evidence from inventor- and examiner-added citations. *The Review of Economics and Statistics 88*(2), 383–388.

Thompson, P. and M. Fox-Kean (2005a). Patent citations and the geography of knowledge spillovers: A reassessment. *The American Economic Review 95*(1), pp. 450–460.

Thompson, P. and M. Fox-Kean (2005b). Patent citations and the geography of knowledge spillovers: A reassessment: Reply. *The American Economic Review 95*(1), pp. 465–466.

Van Looy, B., M. du Plessis, and T. Magerman (2006). Data production methods for harmonized patent statistics: Patentee sector allocation. *Eurostat / K.U. Leuven Working Paper*.

van Looy, B., M. du Plessis, T. Magerman, and X. Song (2009). Data production methods for harmonized patent indicators: Assignee sector allocation. *Eurostat Working Paper and Studies*.

Vincenty, T. (1975). Direct ad inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review 22*, 88–93.

A Appendix

A.1 Characteristics of citations

Table A1: Origin of citations in the treatment group when matching on main group level.

	Universities	Research organisations
Number of observations in treatmentgroup	292	1092
Source of the citation		
Added during search (whether or not provided by applicant)	0.760	0.876
Introduced by applicant, but not used in search report	0.195	0.095
Introduced during examination	0.041	0.022
Provided during opposition proceedings	0.000	0.004
Observed by third parties	0.003	0.003
Search code for citations introduced during search (up to 3 codes can be allocated)		
A Defining general state of the art, not prejudicing novelty or inventive step	0.366	0.506
D Already cited in the patent application and decisive in terms of state of the art or necessary to understand the application	0.103	0.089
E Potentially conflicting documents	0.017	0.005
I Particularly relevant when taken alone (prejudicing inventive step)	0.130	0.001
P Intermediate docs, published between filing date and date priority claimed	0.007	0.038
X Particularly relevant when taken alone (prejudicing novelty)	0.223	0.203
Y Particularly relevant if combined with other documents	0.134	0.139
- none	0.243	0.142

Notes: All variables are indicator variables, hence the reported means can be interpreted as fractions.
Only citations added during search can have a search code. Fractions reported in the lower part of the table are hence fractions of citations added during search. Up to 3 search codes can be allocated and as a result fractions in the lower part of the table sum up to more than one. 11% (universities) to 12% (research organisations) has more than one search code, more than half of these have combination AD.
For granted patents, a search code X is often an indication that the claims have been modified during the granting process as the original claims were prejudiced.
Source: Own calculations using patent citation data from OECD.

A.2 Citation patterns over time for other research organizations

Table A2: Difference over time in distance pattern for the research organizations sample.

	(1) Main results	(2) $\leq$ 4 years	(3) $>$ 4 years
5-25km	-0.0218	0.0829	-0.0739
	(0.0778)	(0.1358)	(0.0920)
25-100km	-0.0883	-0.1226	-0.0883
	(0.0775)	(0.1407)	(0.0915)
100-200km	-0.0848	-0.0321	-0.1172
	(0.0764)	(0.1337)	(0.0906)
200-400km	-0.0833	0.0099	-0.1370^{*}
	(0.0684)	(0.1192)	(0.0822)
400-800km	-0.0344	0.0427	-0.0772
	(0.0675)	(0.1171)	(0.0809)
800-1200km	-0.0415	0.0433	-0.0890
	(0.0739)	(0.1278)	(0.0897)
$\geq$ 1200	-0.1391^{*}	-0.0859	-0.1802^{*}
	(0.0794)	(0.1411)	(0.0950)
Citing and cited from same country	0.1108^{***}	0.1509^{***}	0.0949^{***}
	(0.0271)	(0.0455)	(0.0347)
Intercept	0.5100^{***}	0.4384^{***}	0.5479^{***}
	(0.0689)	(0.1167)	(0.0844)
Country controls	Yes	Yes	Yes
Same country X country	Yes	Yes	Yes
Technology controls	Yes	Yes	Yes
R^2	0.0325	0.0398	0.0378
Observations	4925	1590	3335

Notes: Main sample with patents matched at the main group IPC level. Standard errors clustered at the level of the citing patent in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Own calculations using patent citation data from OECD.

Publisher:

CPB Netherlands Bureau for Economic Policy Analysis
P.O. Box 80510 | 2508 GM The Hague
T (088) 984 60 00

April 2017 | ISBN 978-90-5833-770-2