MERIT-Infonomics Research Memorandum series

Does it matter where patent citations come from? Inventor versus examiner citations in European patents

Paola Criscuolo & Bart Verspagen

2005-018

MERIT – Maastricht Economic Research Institute on Innovation and Technology

PO Box 616
6200 MD Maastricht
The Netherlands
T: +31 43 3883875
F: +31 43 3884905

http://www.merit.unimaas.nl
e-mail:secr-merit@merit.unimaas.nl

International Institute of Infonomics

c/o Maastricht University
PO Box 616
6200 MD Maastricht
The Netherlands
T: +31 43 388 3875
F: +31 45 388 4905

http://www.infonomics.nl
e-mail: secr@infonomics.nl

Does it matter where patent citations come from? Inventor versus examiner citations in European patents

Paola Criscuolo[*] & Bart Verspagen[**]

[*] Tanaka Business School, Imperial College, London
[**] ECIS (Eindhoven University of Technology) and TIK (University of Oslo)

Abstract
This paper investigates whether the distinction between patent citations added by the inventor or the examiner is relevant for the issue of geographical concentration of knowledge flows (as embodied in citations). The distinction between inventor and examiner citations enables us to work with a more refined citation indicator of knowledge flows. We use information in the search reports of patent examiners at the European Patent Office to construct our dataset of regional patenting in Europe, and apply various econometric models to investigate our research question. The findings point to a significant localization effect of inventor citations, after controlling for various other factors, and hence suggest that knowledge flows are indeed geographically concentrated. This holds true also for a sub-sample of patents owned by 169 large multinational enterprises (MNEs). The results for the sample of MNEs suggest that multinational firms seek out specific regional knowledge specializations (and hence at least partly reinforce geographical concentration), but are also able to transfer knowledge "easier" over larger distances

JEL Classification Numbers: O30, O34, F230.
Keywords: Patent citations, local knowledge spillovers, multinational companies.

Acknowledgement: the authors would like to thank Aldo Geuna for helpful discussions and Gustavo Crespi and Joe Hilbe for advice on econometrics. The usual disclaimer applies.

1. Introduction

Patent documents contain citations to other patents and references to articles in order to comply with the legal requirement to supply a complete description of the state of the art upon which the invention described in the patent builds. Thus, citations limit the scope of the inventor's claim for novelty and they represent a link to pre-existing knowledge upon which the invention is built. This latter notion has been used to justify the use of patent citations as indicators of knowledge spillovers. When an inventor cites another patent or a scientific article, this may indicate that the knowledge contained in the cited document has been useful in the development of the citing patent, and therefore the citation might be a proxy for knowledge flows between two inventors.

A large body of empirical studies has exploited this use of patent citations to assess the local nature of technological spillovers (e.g. Jaffe *et al.*, 1998, Jaffe and Trajtenberg, 1996, Jaffe and Trajtenberg, 1999, Jaffe *et al.*, 1993, 2002). Here, the question is whether or not knowledge spillovers between firms, or from (semi-) public knowledge institutes to firms, depend on geographical distance, i.e., whether patent citations are, ceteris paribus, more frequent between two patents that originate from research projects undertaken by inventors that are located closely together. These studies find that both in the US and Europe, such a relationship indeed exists. Thus, knowledge spillovers tend to be more intense between parties that are located close to each other in space.

One of the criticisms of the use of patent citations as indicators of spillovers is that citations are a very noisy indicator of knowledge spillovers (Jaffe *et al.* 1998), i.e., they might be interpreted in different ways than pointing to an actual flow of knowledge from the cited to the citing patent. A crucial factor in this issue is that citations may be added by the applicant (or his/her patent lawyer), as well as by the patent examiner who judges the degree of novelty of the patent. Obviously, when citations are added by the applicant there is more of a case for taking citations as indicators of spillovers, because there is some chance that the inventor actually knew about the cited patent. This may not be the case when the examiner adds the citation: the inventor may never have known about the cited patent.

Most citation studies are not able to identify precisely those citations chosen by the inventor. Moreover, the role of examiner vs. inventor[1] citations differs somewhat between patent systems. In any case, when the inventor proposes citations, the final decision on which documents to cite in the patent publication lies with the patent examiners, and hence patent documents report the inventor citations as chosen by the examiner. The examiner might decide to accept the ones proposed by the applicant and/or add new references, where the latter leads to the bias already identified above, i.e., that patent citations might not reflect an actual source of knowledge spillovers.

A recent number of studies have investigated this issue in citations appearing in US patents (i.e., patents issued by the US Patent and Trademark Office, USPTO) exploiting the fact that, since 2001, the USPTO provides information on the source of the citations (Alcacer and Gittleman, 2004, Sampat, 2004, Thompson, 2004). In this study we explore the origin of patent citations in European Patent Office (EPO) data. We are able to discriminate between the citations listed by the examiner, on the one hand, and the ones proposed by the inventor and accepted by the patent examiner, on the other, exploiting the information contained in the search report.

[1] We will use the term 'inventor citation' to indicate a citation that was added in the original patent application, i.e., irrespective of whether the actual inventor, a patent lawyer or someone else otherwise involved with the application added the citation.

The main objective of this paper is to test whether the references added by the patent examiner are systematically and significantly different from the ones listed by the inventor. In particular, in light of the strong attention to regional spillovers using citations as indicators, this study tries to investigate whether inventor citations and examiner citations are similar with regard to their geographical nature. We draw on a large dataset (all EPO patents originating from a set of 18 European countries), and apply regression analysis to investigate our research question.

2. Theoretical background

It is often assumed that due to the nature of knowledge as only a partial public good, the costs of transferring it depend on distance. Knowledge can in principle be shared without diminishing its value (i.e., knowledge is a non-rival economic good), but there are costs involved in doing so. Face-to-face communication is an efficient way of knowledge transfer, and this is obviously easier at short distances than across the globe. Even with modern information and communication technologies, geographical proximity may be an important factor in transferring knowledge (Morgan, 2004).

Often, the tacit nature of knowledge is given as a reason why knowledge is more easily transferred face-to-face, and hence over small distances. Knowledge resides implicitly in the minds of people, and codification into written materials only partially reflects the full knowledge involved. Hence knowledge flows more intensively between people who have opportunities to physically meet on a regular basis.

Jaffe *et al.* (1993) have used this (often rather informal) reasoning as a starting point of their empirical analysis of the geographical concentration of patent citations. Citations are taken as 'paper trails' of knowledge spillovers from the cited inventor to the citing inventor. They find, in an analysis of U.S. patenting at the Metropolitan level, that citations are indeed more intense at the local level, even after taking account of the pre-existing production structure (i.e., if activities of a similar kind tend to be located near to each other, and patents of a similar kind have a higher probability to cite each other, citations will equally tend to be clustered even without localized spillovers). While they take this as an indicator of spillovers, i.e., unintended flows not directly related to any market transactions, Breschi and Lissoni (2001) have argued that citations are often related to research relationships that are somehow institutionalized, either through the market, or through some form of cooperation.

In the latter case, localized patent citations may be an indicator of the localized nature of knowledge flows in a broader sense than just spillovers (i.e., the flows may not be externalities), but the question as to why these flows are so localized remains the same. Face-to-face contact between researchers, and institutionalized contacts between organizations may just as well serve to explain why knowledge interaction in general, as opposed to spillovers in particular, is easier between firms and organizations that are located close to each other.

An additional issue is related to the notion of a patent as a way of codifying knowledge. Patent law requires the invention to be disclosed, and hence to become available for everyone across the globe. This makes a case – albeit a weak one – for patent citations as indicators of the flow of tacit knowledge. But actually to use the codified information in the patent as a source of inspiration for one's own work may involve becoming familiar with the work of the cited inventor in a different way, for example through conversations, presentations or through the reading of other research output that is less easily available. Hence there may still be a

localization effect in patent citations, despite the fact that patents are an ultimate source of codified knowledge.

A major problem in the use of patent citations as measures of spillovers has been the fact that "the likelihood of knowledge spillover (...) is significantly greater (...) than the likelihood without a citation (... but ...) a large fraction of citations, perhaps something like half, do not correspond to any apparent spillover (...) citations are a noisy signal of the presence of spillovers" (Jaffe *et al.*, 2002, p. 400). A large part of the noise involved in the relationship between citations and spillovers results from the legal framework in which citations are added. Citations are the most important way of limiting a patent's claims, and acknowledging claims made in other (earlier) patents. The fact that a cited patent has implications for the claims in the citing patent does not necessarily imply that spillovers have been going on between the inventors.

As was mentioned in the introduction, one way of diminishing the noise is to filter out patent citations that have been added by the examiner from those added by the inventor (applicant). Examiner citations are less likely to be related to spillovers, because one may suspect that the citing inventor was not aware of the cited patent. This seems to be one of the implications of the survey evidence reported in Jaffe *et al.* (2002). Hence we investigate whether or not there is a difference between inventor citations and examiner citations in terms of their geographical concentration. Because of differences between the patent systems in the US and Europe (on which we will elaborate below), we expect that our European-based evidence will be complementary to the existing studies, which solely use US patent data.

As an additional and related research question, we explore the role of multinational companies. Due to their global nature, these companies may be thought of as integrating elements from local knowledge pools. As such, they may bridge knowledge transfers over larger distances than smaller firms. On the other hand, they may be interested in specific local and strongly specialized strongholds in a region's knowledge base, and by locating in this region as a result of this, these firms may indeed reinforce the extent of local such specialization patterns. Thus, on the one hand we may expect to find a relatively higher tendency for patents owned by these multinational companies to cite patents invented in distant locations because of their ability to carry out international patent searches and because of their presence in several markets. On the other hand, though, we may expect patents cited by large multinationals still to be local, because of their ability to seek out specific local sources of knowledge and to co-locate with them. We will investigate which one of these two trends dominates by focusing on a specific sub sample of patents for multinational firms.

3. Data collection and descriptive statistics

Our primary data sources are the EPO database on patent applications (Bulletin CD) and patent citations to other patents within the EPO and to other non-EPO patents (REFI database) over the period 1985-2000. We also use information contained in the patentability search report that the EPO examiner completes during his screening of technically relevant literature. Contrary to other patent office search reports, the one compiled by EPO examiners contains various categories of citation which grade the cited document according to its relevance. As shown in Table 1, the category 'D' refers to those citations added by the examiner that were already mentioned in the patent application for which the search is carried out, i.e., were proposed by the applicant. This is our source for inventor citations. Thus, we only observe those citations added by the applicant that the examiner believed relevant with respect to the patentability of the invention.

********** INSERT TABLE 1 ABOUT HERE **********

We complement this with the information contained in the OECD citations database on patent applications filed under the Patent Cooperation Treaty (PCT) and on equivalent patents (Webb *et al.*, 2004). When patent applications are examined under the PCT, they undergo an international search that is carried out by one of the International Search Agencies (ISA), of which the EPO is one. If the EPO was the designated ISA, the cited documents together with the categories of citations are not recorded in the REFI database, but they are in the OECD database.[2] In the database where we combine these two sources, 2.5% of citing patents have none of their citations classified in any category, and 8.4% of citing patents have at least one of their citations without category of citation. Because this, in principle, corresponds to an omission, we decided to eliminate the citing patents with at least one citation not classified, which results in dropping 15.4% of the total citation pairs. For each EPO patent the OECD database provides also a list of all patents filed in other patent offices protecting the same innovation (equivalent patents). We use this data to replace citations to national patents with their EPO equivalent in order to increase the sample of within EPO citations for which we have detailed information on inventor's address, technological classes and priority dates.

Table 1 shows, in the last column, the distribution of citations over the citation categories. Note that cited patents can be classified with up to three categories (e.g., "ADL").[3] The largest share (62%) of citations is used to describe the state-of-the-art (*A*), followed by particularly relevant documents (*X*, 21% and *Y*, 16%). 9% of all citations are inventor citations (*D*). All other categories of citations are smaller than 5% of the total. An interesting result is that the predominance of *A* citations is even stronger in the sample of inventor citations in the search report: 72% of all inventor citations has a category *A* attached, vs. 62% for the total sample. Also interesting is the smaller fraction of *X* citations among the sample of inventor citations (11% vs. 21% for the total sample), indicating that inventors have a lesser inclination to cite patents 'particularly relevant if taken alone'. This seems to indicate the (expected) tendency for inventors to not cite patents that may compromise novelty of their own patent. On the contrary, the *Y* category, which similarly points to patents compromising novelty, but only in combination with other patents, occurs as frequent in the sample of inventor citations as in the total sample (both at 16%).

The 9% inventor citations in our database are a small percentage if compared with the fraction found using USPTO patent citations (in the sample of US patents used by Alcacer and Gittelman, 2004, applicant-citations represent 60% of all citations). This finding can be explained by the different legal requirements concerning the description of the state of the art in the two patent offices. While in the USPTO the inventor and his/her attorney are obliged to provide a list of those references describing the state of the art which are considered relevant to the patentability of the invention – the so called 'duty of candour' – the EPO has no similar requirement (Akers, 2000, Meyer, 2000, Michel and Bettles, 2001). As a result, in the EPO, examiners rather than inventors or applicants, add the large majority of patent citations. The obvious implication is that in the EPO system, more often than in the case of USPTO, inventors

[2] When "an application is published according to Art. 158 EPC we store the PCT publication information in REFI as a 'link to' without categories" (private communication with the EPO).

[3] It is almost the case that a citation is classed as one of X, Y, and A. But there is not restriction on additional categories it may be classed as.

may not be aware of patents they cite. As pointed out by Michel and Bettles (2001), applicants to the USPTO "rather than running the risk of filing an incomplete list of references, (…) tend to quote each and every reference even if it is only remotely related to what is to be patented. Since most US examiners apparently do not bother to limit the applicants' initial citations to those references which are really relevant in respect to patentability, this initial list tends to appear in unmodified form on the front page of most US patents" (p. 197).

Further descriptive statistics are given in Table 2. The bottom panel of this table reports statistics for a sub sample that consists of citation pairs where both the citing and cited patents are filed at the EPO. This sub sample is of particular relevance because it will be the source for our econometric analysis below. The reason why we focus on this sub sample is that we have auxiliary information (such as the IPC class, information on applicant/inventor, etc.) only for EPO patents.

The table shows that our 'within EPO' sample is slightly different from the total sample. Obviously, the number of citations per patent is lower, in this case, more than half when considering only within EPO citations. But also the fraction of patents that have only citations added by the examiner is different. The within EPO sample shows a smaller fraction of patents with only citations added by the examiner. The fraction of patents with citations only added by the inventor in the 'within EPO' sample is almost equal to that in the total sample.

********** INSERT TABLE 2 ABOUT HERE **********

4. Descriptive findings on the geographical citation patterns

As a first approach to our main research question, i.e., whether or not inventor and examiner citations have different geographical profiles, we proceed to analyse the geographical source of inventor and examiner citations at the country and regional (i.e., sub-country) level. We ask whether the inventor citations are more likely to originate in the same country (region) as the cited patent than examiner citations. An affirmative answer to this question would indicate that inventor citations may indeed be a better indicator of localized knowledge spillovers than examiner citations.

We attribute each citation to a particular set of countries using a fractional count method and create a dummy variable that equals to 1 if none of the inventors in the citing-cited pair are resident in the same country. Thus the assignment of patents to a country or a region is based on the inventor address (rather than the applicant address). Table 3 provides some basic statistics on this dummy across 30 technological sub-fields as defined by the *Observatoire des Sciences et des Techniques* (OST) and the *Fraunhofer Institute* (FhG-ISI) (see OST, 2002 appendix A5a-1 p. 346). As expected, across all technological classes inventor citations are more co-localized than the examiner citations (that is, the values in the table for inventor citations are smaller). Technology fields in which we find a particularly strong dominance of localized inventor citations are information technology, semiconductors, nuclear technology, motors-pumps-turbines, thermal processes, mechanical components, building and public works (these are the technology fields for which the numbers in the last column of Table 3 are below 30%). Inventor citations are relatively weakly localized (values in Table 3 above 40%) in organic chemistry and food & agricultural products.

********** INSERT TABLE 3 ABOUT HERE *******

We repeat this analysis at a finer level of geographical aggregation. To this end, we define the same indicator, but now at the level of regions within countries. We have a regional breakdown of patents for European countries.[4] The regional breakdown that we use is largely based on the NUTS classification scheme that Eurostat uses. This is based on administrative regions rather than economically coherent regions. Although we would have liked to use the latter, such a classification scheme is not available for the European Union. We use a mix between NUTS 2- and 3-digit level, and in cases where the NUTS region corresponds to a (large) city or very small area, we combine this with the surrounding or adjacent region in order to arrive at more homogenous spatial units (except for Brussels and Berlin). We use the same definition of co-location as in Table 3, i.e., the dummy variable is coded as 1 if none of the citing and cited regions overlap.

Because we only have a regional breakdown for a number of European countries, and hence have to exclude other countries, including large ones such as the U.S. and Japan, we now have a smaller number of observations (245,974 against 914,652 citations pairs in Table 3). The results for this are documented in Table 4. Obviously, because of the stricter geographical definition, we now find higher percentages than in Table 3. Still, the inventor citations appear as more co-located than the examiner citations, in all technological fields. The correlation coefficient between the last columns of Table 3 and 4 is 0.44. With regard to the individual technology fields that we identified above as particularly high or low in terms of localization of inventor citations, we now find some differences. Semiconductors and nuclear technology are still highly localized, but information technology, motors-pumps-turbines, thermal processes, mechanical components, building and public works are now closer to the mean. Organic chemistry is now closer to the average, but food & agricultural products are still more weakly localized, as are medical equipment, agricultural machinery and food processing.

********** INSERT TABLE 4 ABOUT HERE **********

Concluding, our descriptive evidence indeed indicates that inventor citations are more indicative of localized knowledge interaction that examiner citations, with variations by technology field, but this needs to be put to a test in a multivariate analysis.

5. Econometric approach

We proceed to investigate the differences between inventor and examiner citations in a broader and more formal context. To this end, we apply a formal econometric model, in which the citation type (examiner or inventor) is the dependent variable. This is a binary variable that takes the value 1 (0) if the citation was added by the examiner (inventor). The explanatory variables used in the regressions are listed in Table 5.

********** INSERT TABLE 5 ABOUT HERE **********

Among the independent variables, we have three variables measuring geographical proximity. The first of these is a standardised measure of regional distance in kilometres

[4] A full list of the 135 regions we use is provided in the appendix. Our countries include the EU-16 plus Norway and Switzerland.

(*DistanceKM*) between the region of the citing and cited patents. Appendix I explains how this variable was calculated. We expect that *DistanceKM* to be positively correlated with examiner citations (i.e., an odds ratio larger than one).

In addition to this, we have the two dummy variables that have been used in Tables 3 and 4. One is coded as 0 if the citing and cited patents originate from the same country (*Diff_Ctrys*), and the other is similarly defined at the regional level (*Diff_Regions*). Based on the hypothesis of local interaction, we expect these geographical variables to have a odds ratio greater than 1, i.e. that examiners are more likely to add citations to patents originating from distant location than inventors.

Our next variable is the C*itation lag* (in years), which is the time period elapsed between the priority dates of the citing and cited patents. This controls for a potential difference in time scope between inventors and examiners. We have no strong theoretical expectations on the value of the odds ratios for this variable, but we could hypothesize that examiners, because of their detailed knowledge of patent literature in the specific field they cover, have a 'longer memory' and thus they would have a tendency to add older patents in the search reports.

Technological relatedness is another variable that we wish to control for, and this is why we include a dummy variable that is coded as 0 if the citing and cited patents are classified in the same 4-digit IPC class (*Diff_Tech*). We include this variable in order to be able to account for the potential effect of co-location of similar types of R&D activities. Jaffe *et al.* (1993) have argued that it may be the case that R&D in a certain field tends to be co-located in space (e.g., research on semiconductors may be concentrated in Silicon Valley). Because patent citations are by definition to technologically related patents, this would lead to a geographical concentration of patent citations without necessarily pointing to any additional effect related to stronger knowledge flows at the local level. Our *Diff_Tech* variable, to the extent that its 4-digit IPC level indeed captures the relevant technological linkages, accounts for this. If inventors are more likely to cite local patents for reasons of technological relatedness, we expect this will turn up in the coefficient of the *Diff_Tech* variable. If, on the other hand, we find that the geographical variables are significant in addition to the *Diff_Tech* variable, this is evidence for a localization effect in addition to that of the geographical concentration of R&D activities of a specific kind.

We also include a dummy variable indicating whether or not the citing patent was actually granted (*Citing_Granted*). Similarly *Cited_Granted* is a dummy variable that takes the value of 1 if the cited patent application was successful. We have no specific expectations on the sign of these variables. A next set of variables is related to the citation categories that were explained in Table 1 above. We construct three mutually exclusive dummy variables capturing the classes (other than *D*, which defines our dependent variable) that are most frequent (*A*, *Y* and *X*). The remaining categories account for a minor fraction of the patents in our sample (see Table 1), and hence we drop citations classified under one of these categories. This implied excluding from the analysis only 3096 citation pairs. The categories *X* and *Y* pose a serious threat to the novelty of the patent, and hence, as already observed above, we expect that inventors will be less likely to add citations in these categories.

Finally, for a subset of patents owned by a group of 169 multinational enterprises (MNEs), we have information on their subsidiaries' names under which the total company group patents at the EPO (see Verspagen and Schoenmakers, 2004 for an explanation of how this dataset was constructed). Thus, we are able to account for self-citations for this sub sample. Accordingly, we define a dummy variable that is 0 if the cited and citing patents are owned by the same MNE (*Diff_MNE*). We can calculate this variable only for the sub sample of patents in which both the

cited and citing patent are owned by one of the 169 MNEs. We expect the odds ratio of this variable to be greater than 1, i.e. that self-citations are more likely to be generated by the inventors.

Tables 6 and 7 provide, respectively, descriptive statistics and the correlation matrix for the variables used in the regressions.

********** INSERT TABLE 6 ABOUT HERE **********

********** INSERT TABLE 7 ABOUT HERE **********

Our baseline estimation method is the logit model. But as was already indicated, our dependent variable is skewed, i.e., it contains relatively more 1s than 0s. Also, because citation may be influenced by personal characteristics of the applicant or examiner, as well as the specific technology involved in the patent, we might expect that the error term in our econometric equation is correlated between citation pairs that involve the same citing patent. In order to take account of these special features of the data, we apply a range of specific logit models that address this in various ways. In order to deal with the correlated error terms, we follow Alcacer and Gittelman (2004) and first apply a random effects panel model, in which the random effects refer to the citing patent, and the 'time' dimension is represented by the various citations in a given citing patent.

We also apply a model with clustered errors on citing patents to take account of this (Moulton, 1990). This assumes that the observations (citations) are drawn from a population with a grouped structure, and that the errors are correlated within the groups. The clustered error structure solves for a downward bias that would result in a model that wrongly assumes no clustered errors.

The skewed nature of the data is addressed by using two special logit models, in which the actual logit function that is used in the specification is asymmetric. The first of these is the so-called skewed logit model (scobit), the other is the complementary log-log model (cloglog). The cloglog model fits an asymmetric sigmoid function to the probability between zero and one, unlike the probit and logit models, which are both symmetric around ½.[5] The probability function of the cloglog model approaches zero fairly slowly, but approaches one quite sharply, i.e. the sigmoid function is more elongated in comparison to the logit or probit models (Agresti, 2002). The scobit, or skewed logit model also departs from the logit and probit model in that it is another asymmetrical extension of the logit model. In particular, it generalizes the logit model by introducing an additional skewness parameter in the form of the power of the logit function (Nagler, 1994).[6] Hence the logit model is nested in the scobit model, which allows carrying out a log-likelihood ratio test to compare scobit to logit.[7]

[5] This model has been used extensively to model grouped survival data (Greenland, 1994). The model can be written as $\Pr(Y = 1 \mid x) = 1-\exp(-\exp(\alpha + \beta x))$, or as $\log(-\log(1-p(x))) = \alpha + \beta x$, where $p(x) = \Pr(Y = 1 \mid x)$.

[6] Thus, the scobit model can be written as: $\Pr(Y = 1 \mid x) = 1-1/\{1+\exp(\beta x)\}^{\alpha}$

[7] The ratio (Log-likelihood (scobit)/log-likelihood (logit)) is distributed as a χ^2 with one degree of freedom.

6. Estimation results

The results of the various models are presented in terms of odds ratios in Tables 8 – 11. In Table 8, which reports results for the total sample of within EPO patents, all regressions confirm that citations that are added by examiners tend be further apart (in terms of geographical distance between the citing an cited inventor), or, in other words, that inventor citations are more geographically concentrated. This is shown by the odds ratios for the variable *DistanceKM*, which is always larger than one. The table also confirms that examiners are more likely to add the 'dangerous' citation types *X* than the 'common' citation type *A*, which is the reference category. But contrary to our expectations, examiners are less likely to add citations type *Y* compared to citations type *A*.

********** INSERT TABLE 8 ABOUT HERE **********

Examiners have a higher tendency than inventors to cite patents over longer citations lags and to cite within the same technology class, however the odds-ratios of these variables (*Citation lag* and *Diff_Tech)* are very close to one (especially *Citation lag*), pointing to only small differences between inventor and examiner citations in this respect. Finally examiners are less likely to add citations in successful patent applications and to cite granted patents than inventors.

With regard to the magnitude of the estimated odds ratios, it is notable that the scobit model provides the largest deviations from one. This is particularly visible in case of the *Cited_granted* and the citation type variable *ClassX*, but also present in the other ones (including *DistanceKM*).

********** INSERT TABLE 9 ABOUT HERE **********

The results on geographical concentration are confirmed in Table 9, which reports on the smaller sample of MNEs' patent citations. Here we can include the self-citation dummy. This turns up with a larger than one odds ratio, as expected, i.e., examiners (inventors) tend to cite more patents owned by other (the own) companies. In fact, in our sample of MNE patents for which we have self-citations, more than 60% of all citations made in these patents are self-citations. Within this sample, the geographical concentration of inventor citations is still present and significant, although the effect is somewhat less strong (i.e., lower odds ratios) than in Table 8, when compared within the same estimation method.

Thus, even in this sample of MNEs, inventor citations are geographically concentrated, although at a lower intensity. This lower odds ratio may be due both to the inclusion of the self-citation dummy, and to the multinational nature of the companies in the sample. Self-citations will, by definition, have a larger tendency to be located close in space, and hence in a regression without this variable included, one may expect a larger coefficient on the *DistanceKM* variable. This is confirmed in column 3 of Table 9 where we report estimates of a model that does not include the *Diff_MNE* variable, but for the same sample of MNE patents as in the rest of the table.

On the other hand, the large MNEs in this sample may be expected to tap into a more global knowledge base, and hence be less susceptible to distance. Both effects are consistent with our results in Tables 8 and 9, but the conclusion from Table 9 is that none of these effects can account for the full geographical effect found in Table 8. In other words, even when taking

account of self-citations, and in the sample of 'globalized' firms, inventor citations seem to be more localized than examiner citations.

In the MNEs sample, the effect of the technology class and citation lag reverses. Also, the effect of the citations types *X* and *Y* seems smaller. This may be a result of the fact that patent citations by large MNEs are based on more professionalized patent searches, due to the resources these firms have available to undertake such search processes, leading to both broader technological scope and longer citation coverage.

From the various econometric specifications, we choose the logit model with random effects as the one that performs best, as indicated, e.g., by the two information criteria (AIC and BIC) that we document in the table. The random effect logit model scores better than the scobit and complementary log-log models, which suggests that accounting for heterogeneity in the citing patents by means of random effects is statistically more important than accounting for skewness of the dependent variable. The importance of the individual variance component (within citing patent variation) indeed seems to be quite important in our sample (see value and significancy of ρ in Tables 8 and 9).[8] Even though the scobit model take account of heterogeneity by means of clustered errors, the random effects approach still dominates in a statistical sense.

We therefore use the random effect logit model, in Table 10, to further investigate the effect of variations in the geography variable definitions, and the inclusion or exclusion of certain parts of the sample. We also use this model to perform a number of estimations for sub samples of technology classes. Summary statistics on these regressions are documented in Table 11.

********** INSERT TABLE 10 ABOUT HERE **********

In Table 10, the first column is repeated from Table 8 for comparison. The next two columns substitute the *DistanceKM* variable by our two dummy variables of Tables 3 and 4. In the case of the region dummy (second column), this implies a much stricter definition of the localization effect. Whereas the use of *DistanceKM* allows for a smooth decay of the probability of an inventor citation with distance, the effect is dichotomous (within or outside the region) in the case of the dummy region. This is reflected in a sharp increase in the odds ratio of the region dummy as compared to *DistanceKM*. In the case of the country dummy, the effect obviously depends on the (average) country size in our sample. For large countries, this dummy does not imply a very strong localization effect, but for small countries it does. Note that in this case, we include, as in Table 3, citations involving patents outside the European countries for which we have regional data, and this increases the number of observations drastically. In this case, we still find a significant and fairly high localization effect. Thus, we conclude that the localization effect for inventor citations is robust for various definitions of localization.

The results for the other variables than localization are also fairly robust between the first three columns of Table 10. The differences between the first two columns (*DistanceKM* and region dummy) are relatively small, but the third column (within-country dummies as the measure of localization) yields larger deviation from one for the odds ratio, although the odds ratios in column three are on the same side of one as in the first two columns.

In column 4 we exclude the variable capturing the technology effect (*Diff_Tech*). In line with our argumentation above about co-location of citations and the effect of the pre-existing geographical specialization of R&D, we would expect that without this variable, a higher burden of explanation would come to rest on the *DistanceKM* variable. Without *Diff_Tech*, we would

[8] Unfortunately, we are unable to include random effects in the scobit models.

expect that *DistanceKM* will pick up the effect of pre-existing specialization patterns and that of localized spillovers. However, contrary to this expectation, the exclusion of *Diff_Tech* does not affect the odds- ratios of the *DistanceKM* variable (nor that of any of the other variables). This seems to imply that the interaction of pre-existing specialization of R&D with other variables in the model does not differ between examiner and inventor citations. In terms of geography, this means that if there is indeed an effect of pre-existing geographical specialization patterns on co-location of patent citations, it does not differ between inventors and examiners.

In column 5 we introduce a new dummy variable, *ClassXY*, which is equal to 1 if citation is type *X* or type *Y* and 0 otherwise, i.e. this variable captures jointly the two citation types that pose a serious threat to the novelty of the patent. We find that the odds-ratios of this variable is greater than 1, which implies that examiners are more likely to add the most 'dangerous' citations with respect to citations describing prior-art (citations type A) than inventors.

The last 3 columns in Table 10 also confirm the localization effect of inventor citations. Here we present results for specific sub samples. First, we exclude all citing patents for which all citations are examiner citations, next we exclude all citing patents for which the citations are all added by the inventor, and finally we exclude both previous types of citing patents. The reason why we exclude these types of patents is that citations where all citing patents are of one type only, might present cases where unobserved variables (e.g., personal characteristics of the examiner or inventor[9]) dominate the data, rather than a true localization effect. If this is a real feature in our data, the cases where one citing patent contains both examiner and inventor citations are much more reliable indicators for a localization effect (or its absence).

Naturally, as a result of dropping a number of citing patents, the number of observations in the regressions drops as well. This is most drastic when we drop all citing patents with only examiner citations. Still, the localization effect remains present and also the effect of the other variables. This holds true also when we exclude all citing patents with only inventor citations. Finally, we exclude both types of citing patents, and the number of observations drops most drastically. The localization effect remains significant, but we have an adverse effect for the citation lag and the two granted variables.

Concluding, what Tables 10 and 11 show in a general sense, is that the overall results in Tables 8 and 9 are robust to the variations that we apply. Stronger geographical concentration of inventor citations than examiner citations is a robust feature of our dataset, no matter what exact variables we use to indicate such concentration, and whether or not we exclude certain categories of data.

This robustness finding still holds in Table 11, which presents a summary of the estimations in separate technology classes. Here we find that in all technology classes the localization effect of inventor citations is significant. The *ClassX* variable is also significant (with odds ratios larger than one) in all technology fields. *Citation_lag* and *Diff_Tech* are significant with an odds ratio larger than one in the large majority of cases, while the odds-ratios of the *Cited_Granted* variable is consistently less than one and significantly.

7. A closer look at the effect of distance

So far, we have assumed that the effect of distance is linear, but it might be the case that the relation between the likelihood of examiner citation and distance is nonlinear. In particular, we

[9] For example, we might have inventors (applicants) that never cite anything, or examiners who have a very high tendency to scrap inventor citations.

would expect that at small distances, the increase in distance by a unit (km) would lead to a stronger effect on the likelihood of an examiner citation, than the same increase at longer distance. In order to test for this, we employed a non-parametric method that starts with eliminating the effect of variables other than distance from the likelihood of an examiner citation. To this end, we first estimate a random effects linear regression model, with *cits_examiner* as the dependent variable, and the variables in Tables 8 and 9 as the independent variables. We then calculate a residual from this regression as $r_i = e_i - \hat{e}_i$, where e stands for *cits_examiner*, and $\hat{e}_i = \hat{c} + \hat{\beta}X_i + \delta_i$. Here c and β are the parameters in our linear model, X is the vector of independent variables except *distanceKM*, δ is the random effect associated with the citing patent, and hats indicate estimated values. Note that the regressions from which we draw $\hat{c}$ and $\hat{\beta}$ did include *distanceKM* as an independent variable, but we do not include this variable in the calculation of the residual *r*. Hence *r* 'partials out' from *cits_examiner* all variables except distance.[10]

Next, we run a locally weighted regression (lowess) of *r* on *distanceKM* (we use a bandwidth of 0.8). This regression yields a smooth curve, of which each point corresponds the 'local' (for the value of *distanceKM*) effect of distance on the likelihood of an examiner citation. We document the results of this procedure in Figure 1.[11] Instead of the version of *distanceKM* that is standardized into units of 173 km, we use on the horizontal axis of this figure a distance variable with units of 1 km.

Figure 1 indeed confirms that the effect of distance is nonlinear. At short distances between the cited and citing patent, the likelihood of an examiner citation quickly increases with distance, but this effect wears off at larger distances. Beyond 1,000 km (which is, say, the distance between the Brussels and Vienna regions, or the Paris and Copenhagen regions), the marginal effect of distance on the likelihood of an examiner citation becomes rather low. The longest distance between two regions in our sample is around 4,000 km (between the northern Scandinavian and Southern Spanish regions) if we do not include the Canary Islands, and approximately 1,500 km more if we include them. This non-linear effect of distance is consistent with the results found in Bottazzi and Peri (2003).

********** INSERT FIGURE 1 AROUND HERE **********

Figure 2 documents the same relationship, but now for the sample of MNEs and based on a linear regression including the self-citation dummy. Now we see a relationship that is almost linear. While this confirms our finding in the previous section that distance is an issue even in technology flows between large MNEs, it also confirms that the nature of the effect of distance is special for this particular type of firms. In particular, we find that at shorter distances, the likelihood of a spillover (inventor citation) does not wear off as quickly for MNEs than for non-MNEs, which implies that MNEs are indeed less susceptible to distance than other firms.

[10] This method was proposed by Hausman and Newey (1995) and Bandiera and Rasul (2003).

[11] We also applied other methods to assess the potential nonlinear nature of the distance relationship, among were to estimate a step-function for *DistanceKM*, to estimate a linear spline function for *DistanceKM*, and to use kernel regression instead of locally weighted regression in the above procedure. These methods generally pointed in the same direction of the results that we document.

********** INSERT FIGURE 2 AROUND HERE **********

8. Concluding summary

The European patent database allows the identification of whether citations are added by the applicant/inventor (inventor citations) or the patent examiner. This information is available for the complete history of patent citations in the European patent system, and hence provides a rich source of data for assessing whether or not inventor citations indeed tend to be concentrated in geographical space. On the basis of this database, we have provided evidence based both on descriptive statistics and on the basis of multivariate econometrics. Both approaches yield a clear-cut conclusion: citations that originate from inventors/applicants are more concentrated in space than citations that originate from the patent examiners.

In our descriptive analysis, this holds both at the national level (inventor citations are more often to patents invented in the same country where the inventor is resident), and at the regional (i.e., sub-national) level in Europe (inventor citations are more often to patents invented in the same region where the inventor is resident). The econometric analysis controls for a number of other factors, such as the technological relatedness of the cited and citing patent, the citation lag (time elapsed between the cited and citing patent), the citation type (referring to state-of-the art, or citations that may compromise novelty), and whether or not the citing and cited patents are granted. We also apply different measures of distance and co-location of cited and citing patent, and we experiment with different sub samples and estimation methods. All econometric evidence points to a significant localization effect of inventor citations. Citations added by the examiner are rarely clustered, and span larger geographical distances between cited and citing patent. This result is completely robust across sub samples, the estimation methods and the various ways in which distance and co-location are measured.

Otherwise, we find that examiner citations more often involve citations that may compromise novelty, which points out that inventors may indeed have a tendency to omit relevant citations that may endanger their patent claims.

Our results point to two main conclusions. First, by benchmarking inventor citations against examiner citations, we find that knowledge flows (to the extent that they are indicated by patent citations) are indeed localized. We take the patterns of citation in the sample of examiner citations as somehow representative for the potential linkages between global R&D workers, and the inventor citations as the part of these potential flows that have indeed materialized. Interpreted in this way, our evidence suggests that the actual technology flows are more geographically concentrated than the potential flows, or in other words, that knowledge interaction is stronger at small distances than over long distance. Testing for potential non-linearity of this relationship, we find that an increase in distance has a stronger effect when citing and cited patent are close to each other. In other words, the effect of distance is strong initially, but wears off when distance becomes large.

Our econometric analysis also controls for whether or not the technology classes of the cited and citing patent are the same. If the main reason for inventor citations to be more concentrated in geographical space was that patents in the same technology class are more often co-located, we would have expected that the technology class variables would have been positively correlated with inventor citations. But this is not generally the case, except in a sub sample for large MNEs, and hence we conclude that the localization effect that we find for inventor citations

results from a source that is additional to the (potential) tendency of similar R&D activities to co-locate in space. In other words, the distribution of sectoral composition of R&D activities over space is not the prime responsible variable for the localization effect that we observe.

This does still not answer in a direct way the question what is behind this localization effect. It is obvious that patents are a source of codified knowledge, and hence the (often-used) hypothesis that the tacit nature of knowledge is responsible for the geographical concentration of knowledge flows is not immediately attractive. But we also cannot rule out the rule of tacit knowledge completely, because there might be tacit aspects related to the codified knowledge described in the patent. However, also the existence of common resources, such as a pool of skilled labour and the availability of (semi-)public research institutes and universities may also explain the localization effect. More explicit research into these reasons for knowledge flows to be concentrated remains necessary.

We have been able to construct a sub sample of citations between large MNEs in which we can identify separately whether or not a citation is to a patent of the same firm (self-citation). Controlling for this in addition to the other control variables, we still find a significant localization effect of inventor citations, although this is a smaller effect than in the large sample. This means that self-citations do not account for the geographical concentration effect found in our econometric analysis. We also find that the effect of distance in the MNE sample is much more linear than for the total sample, and hence that MNEs are less susceptible for the rapid increase of the effect of distance when cited and citing patent are already close.

But more importantly, it means that even in a sample of 'globalized firms', local concentration of knowledge flows is a relevant phenomenon, and this is our second main conclusion. Our finding indicates that rather than making the world unequivocally a smaller place, these large multinational firms add a specific factor to the global process of knowledge generation that at least partly reinforces geographical concentration. They have the ability to seek out locations where relevant knowledge is generated, and tap into this knowledge by means of a process that has been termed asset-augmenting R&D investment (Dunning and Narula, 1995). In this way, they tend to reinforce the technological capabilities of these locations, and hence the geographical differences in terms of knowledge generation.

References

AGRESTI, A., (2002), *Categorical data analysis*, 2nd ed. Wiley and Sons, Hoboken New Jersey.

AKERS, N., (2000), The referencing of prior art documents in European patents and applications, *World Patent Information*, 22: 309-315.

ALCACER, J., GITTLEMAN, M., (2004), "How do I know what you know? The role of inventors and examiners in the generation of patent citations", paper presented at the NBER conference, Boston.

BANDIERA, O., RASUL, I., (2003), Social networks and the adoption of new technology in Northern Mozambique, CEPR Discussion Paper No. 3341.

BOTTAZZI, L., PERI, G., (2003), Innovation and spillovers in regions: evidence from European patent data, *European Economic Review*, 47: 687-710.

BRESCHI, S., LISSONI, F., (2001), Knowledge spillovers and local innovation systems: a critical survey, *Industrial and Corporate Change*, 10: 975-1005.

DUNNING, J., NARULA, R., (1995), The R&D activities of foreign firms in the United States, *International Studies of Management and Organization*, 25: 39-73.

GREENLAND, (1994), Alternative models for ordinal logistic regression, *Statistics in Medicine*, 13: 1665-1677.
HAGGET, P., CLIFF, A. D., FREY, S., (1977), *Location Models*, Edwards Arnold, London.
HAUSMAN, J. A., NEWEY, W. K., (1995), Nonparametric estimation of exact consumers surplus and deadweight loss, *Econometrica*, 63: 1445-1476.
JAFFE, A., TRAJTENBERG, M., FOGARTY, M., (2002), The meaning of patent citations: report on the NBER/Case-Western Reserve survey of patentees. In: JAFFE, A., TRAJTENBERG, M. (Eds.), *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. The MIT Press, Cambridge, Massachusetts.
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, *Journal of Industrial Economics*, 46: 183-205.
JAFFE, A. B., TRAJTENBERG, M., (1996), Flows of knowledge from universities and federal labs: modelling the flow of patent citations over time and across institutional and geographical boundaries, NBER Working Paper No. 5712.
JAFFE, A. B., TRAJTENBERG, M., (1999), International knowledge flows: evidence from patent citations., *Economics of Innovation and New Technologies*, 8: 105-136.
JAFFE, A. B., TRAJTENBERG, M., HENDERSON, R., (1993), Geographic localization of knowledge spillovers as evidenced by patent citations, *The Quarterly Journal of Economics*, 108: 577-598.
MAURSETH, P., VERSPAGEN, B., (2002), Knowledge spillovers in Europe. A patent citations analysis, *Scandinavian Journal of Economics*, 104: 531-545.
MEYER, M., (2000), What is special about patent citations? Differences between scientific and patent citations, *Scientometrics*, 49: 93-123.
MICHEL, J., BETTLES, B., (2001), Patent citation analysis. A closer look at the basic input data from patent search reports, *Scientometrics*, 51: 795-816.
MORGAN, K., (2004), The exaggerated death of geography: learning, proximity and territorial innovation systems, *Journal of Economic Geography*, 4: 3-21.
MOULTON, B. R., (1990), An illustration of a pitfall in estimating the effects of aggregate variables on micro units, *The Review of Economic Statistics*, 72: 334-338.
NAGLER, J., (1994), Scobit: an alternative estimator to logit and probit, *American Journal of Political Science*, 38: 230-255.
OST, (2002), *Science et Technologie: Indicateurs*. Economica, Paris.
SAMPAT, B. N., (2004), Examining patent examination: an analysis of examiner and applicant generated prior art, paper presented at NBER conference, Boston.
THOMPSON, P., (2004), Patent citations and the geography of knowledge spillovers: what do patent examiners know?, Mimeo, Department of Economics, Florida International University.
VERSPAGEN, B., SCHOENMAKERS, W., (2004), The spatial dimension of patenting by multinational firms in Europe, *Journal of Economic Geography*, 4: 23-42.
WEBB, C., DERNIS, H., HARHOFF, D., HOISL, K., (2004), A first set of epo patent database building blocks for analysing European and international patent citations, OECD mimeo.

Appendix I. Distance calculations

A distance table between the European regions in our sample is not readily available. The approach taken here to calculate such a table is based on a computer map of Europe. This map was taken from Eurostat's classification server RAMON[12] but altered to take into account our customized regional breakdown. The map was divided into a dense set of cells (pixels). Each pixel was assigned either to a region or a border between municipalities. This was done on the basis of the borders drawn on the computer image of the map. Pixels assigned to borders were not included in the calculations. The distance between any two pixels on the map was defined as the Euclidean distance between them (the unit of measurement is kilometers). The fact that Euclidean distance on the flat computer map was used implies that no account was taken of the curvature of the globe. Also, no correction was made for the imperfections introduced by the projection of the map onto a flat space. The distance between two regions *i* and *j* was defined as the mean of the individual distance between all possible pairs of pixels, with one pixel located in *i*, and the other pixel located in *j*.

Because we report odds ratios in the documentation of regression results, a unit of 1 km is not very useful (it is too small to point out any discernable effect). Thus, we divide the distance in kilometres by 173, which is the distance that is found, on average, between two bordering regions on our map. We arrived at this 173 km distance by first defining a new variable *B*, in which B_{ij} for regions *i* and *j* is defined as the minimum number of borders one has to cross to reach *j* from *i* (or vice versa).[13] We then divide the distance in kilometres by the corresponding value of *B* and take the average over all pairs of regions, which yields 173 km. In cases where the citing and/or cited patents involve more than one inventor, we calculate an average distance value between all combinations of regions involved on the citing and cited side.

[12] http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html.

[13] In the geographical literature (e.g. Hagget *et al.*, 1977), this is rather common as a direct measure of distance. Note that in order for the distance variable to make sense, the regional map to which it is applied needs to be contiguous, i.e., every region must be reachable from every other region. In our European case, this requires us to deal with a number of sea passages, e.g., between the UK and continental Europe. In those cases, we have assumed that the sea area between our regions can be considered as a separate, artificial region, and so the map of regions becomes contingent. Details of this procedure are available on request, as are the resulting values for this variable.

Appendix II. The regions

For the following countries/regions, the NUTS classification has been used:

Austria	
AT11	Burgenland
AT12+AT13	Niederösterreich
AT21	Kärnten
AT22	Steiermark
AT31	Oberösterreich
AT32	Salzburg
AT33+AT34	Tirol And Vorarlberg
Belgium	
BE1	Brussels Hfdst.Gew
BE2	Vlaams Gewest
BE3	Region Wallonne
Germany	
DE1	Baden-Württemberg
DE2	Bayern
DE3	Berlin
DE4	Brandenburg
DE5+DE9	Bremen And Niedersachsen
DE6+DEF	Hamburg And Schleswig-Holstein
DE7	Hessen
DE8	Mecklenburg-Vorpommern
DEA	Nordrhein-Westfalen
DEB+DEC	Rheinland-Pfalz And Saarland
DED	Sachsen
DEE	Sachsen-Anhalt
DEG	Thüringen
Spain	
ES11	Galicia
ES12+ES13	Asturias And Cantabria
ES21+ES22+ES23	Pais Vasco, Navarra And Rioja
ES24	Aragon
ES3	Madrid
ES41	Castilla-Leon
ES42	Castilla-La Mancha
ES43	Extremadura
ES51	Cataluna
ES52	Valenciana
ES53	Baleares
ES61	Andalucia
ES62	Murcia
ES7	Canarias
Netherlands	
NL1	Noord-Nederland
NL21	Overijssel
NL22	Gelderland
NL23	Flevoland
France	
FR1	Ile De France
FR21	Champagne-Ardenne
FR22	Picardie
FR23	Haute-Normandie
FR24	Centre
FR25	Basse-Normandie
FR26	Bourgogne
FR3	Nord-Pas-De-Calais
FR41	Lorraine
FR42	Alsace
FR43	Franche-Comte
FR51	Pays De La Loire
FR52	Bretagne
FR53	Poitou-Charentes
FR61	Aquitaine
FR62	Midi-Pyrenees
FR63	Limousin
FR71	Rhone-Alpes
FR72	Auvergne
FR81	Languedoc-Roussillon
FR82	Provence-Alpes-Cote D'azur
FR83	Corse
Greece	
GR1	Voreia Ellada
GR2+GR3	Kentriki Ellada And Attiki
GR4	Nisia Aigaiou, Kriti
Italy	
IT1	Nord Ovest
IT2	Lombardia
IT31	Trentino-Alto Adige
IT32	Veneto
IT33	Friuli-Venezia Giulia
IT4	Emilia-Romagna
IT51	Toscana
IT52	Umbria
IT53	Marche
IT6	Lazio
IT7	Abruzzo-Molise
IT8	Campania
IT9	Sud
ITA	Sicilia
ITB	Sardegna

NL31	Utrecht
NL32	Noord-Holland
NL33	Zuid-Holland
NL34	Zeeland
NL41	Noord-Brabant
NL42	Limburg
Portugal	
PT11	Norte
PT12	Centro
PT13	Lisboa E Vale Do Tejo
PT14	Alentejo
PT15	Algarve
Sweden	
SE01+SE02	Stockholm And Östra Mellansverige
SE03+SE04	Småland And Sydsverige
SE05	Västsverige
SE06	Norra Mellansverige
SE07	Mellersta Norrland
SE08	Övre Norrland
United Kingdom	
UK1	North
UK2	Yorkshire And Humberside
UK3	East Midlands
UK4	East Anglia
UK5	South East
UK6	South West
UK7	West Midlands
UK8	North West
UK9	Wales
UKA	Scotland
UKB	Northern Ireland

For the following countries, a national classification has been used:

Norway	**Based on Fylken**
NO1	Akershus, Oslo
NO2	Hedmark, Oppland
NO3	Østfold, Busekrud, Vestfold, Telemark
NO4	Aust-Agder, Vest-Agder, Rogaland
NO5	Hordaland, Sogn og Fjordane, Møre of Romsdal
NO6	Sør-Trøndelag, Nord-Trøndelag
NO7	Nordland, Troms, Finnmark
Switzerland	**Based on Cantons**
CH1	Jura, Neuchâtel, Fribourg, Vaud, Geneva
	Argovia, Appenzell Inner-Rhodes, Appenzell Outer-Rhodes, Basel-Country-Basel-Town, Berne, Glarus, Lucerne, Nidwalden, Obwalden, St. Gallen, Schaffhausen,
CH2	Schwyz, Solothurn, Thurgovia, Uri, Zug, Zurich
CH3	Valais, Ticino, Grisons
Denmark	**Based on postal regions**
DK1	Hillerød, Helsingør, København
DK2	Fyn, Sjaelland ex. Hillerød, Helsingør, København
DK3	Jylland
Finland	**Based on municipalities**
FI11_12	Uusimaa+Etelä-Suomi
FI13	Itä-Suomi
FI14	Väli-Suomi
FI15	Pohjois-Suomi

The following countries have been included as a single region:

Ireland

Luxemburg

Figure 1. The relationship between distance and the likelihood of an examiner citation, total sample

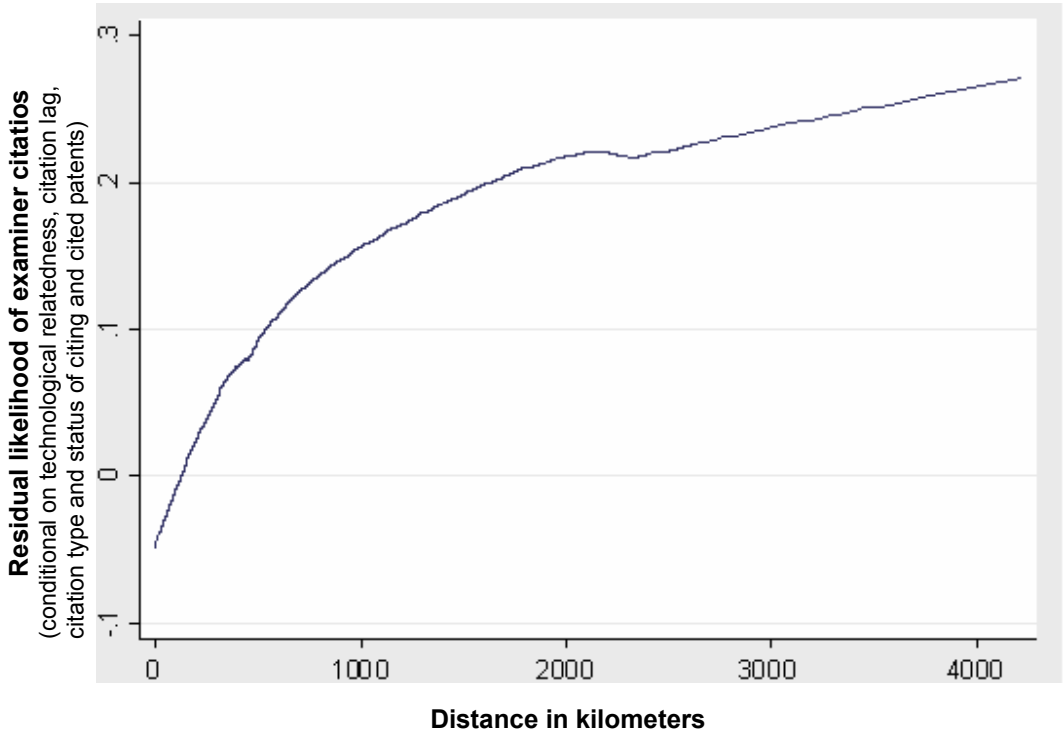

Figure 2. The relationship between distance and the likelihood of an examiner citation, MNE sample

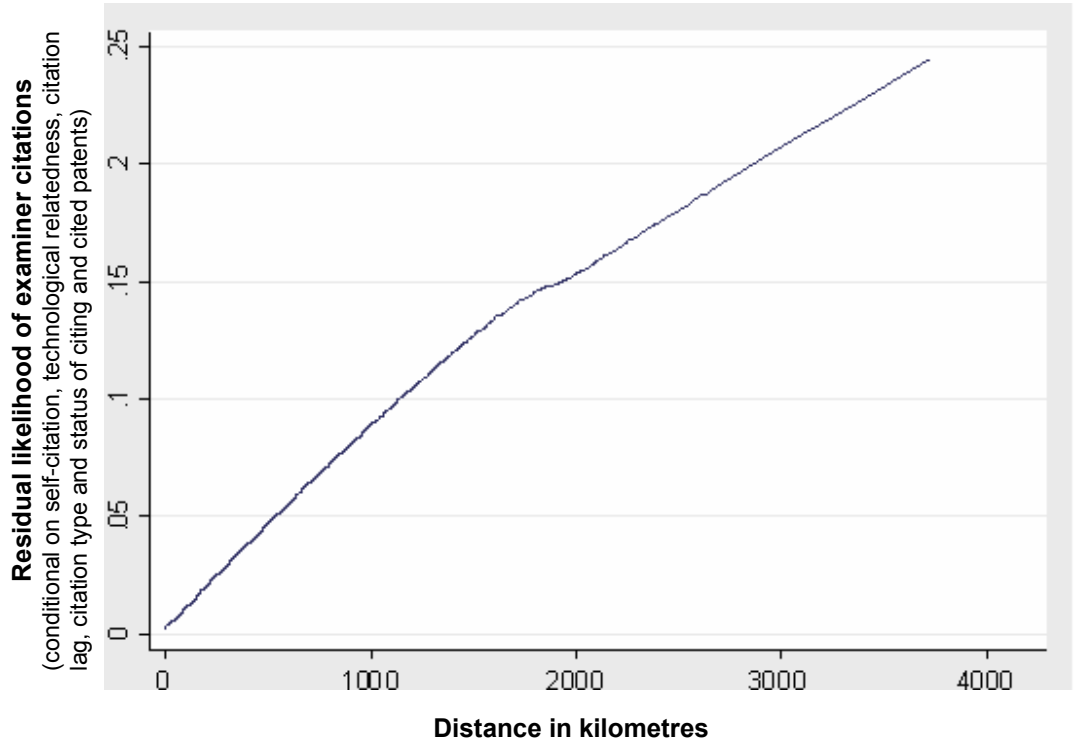

Table 1. Description of category of citations

Category of citations	Description	Fraction of all citations
X	Particularly relevant documents if taken alone; citations classified under these categories are such that when taken alone a claimed invention cannot be considered novel or cannot be considered to involve an inventive step.	0.62
Y	Particularly relevant documents if combined with another document, such a combination being obvious to a person skilled in the art.	0.20
A	Documents defining the state of the art and not prejudicing novelty or inventive step.	0.16
O	Documents that refer to a non-written disclosure.	0.09
P	Intermediate documents; Documents published on dates falling between the date of filing of the application being examined and the date of priority claimed.	0.04
T	Documents relating to the theory or principle underlying the invention.	0.01
E	Earlier patent documents, but published on, or after the filing date.	0.00
D	Documents cited in the application.	0.00
L	Documents cited for other reasons.	0.00

Source: EPO examination guides lines part B chapter X

Table 2. Descriptive statistics

Total Sample	
Number of citing patents	700,623
Number of citations	3,140,367
Citations per patent (mean)	4.48
Number of citing patents with all citations added by the examiner	530,842
Fraction of citing patents with all citations added by the examiner	75.8
Number of citing patents with all citations added by the inventor	16,617
Fraction of citing patents with all citations added by the inventor	2.37
Sample of within EPO citations	
Number of citing pats	490,230
Number of citations	1,263,642
Citations per patent (mean)	2.57
Number of citing patents with all citations added by the examiner	360,446
Fraction of citing patents with all citations added by the examiner	73.52
Number of citing patents with all citations added by the inventor	10,459
Fraction of citing patents with all citations added by the inventor	2.13

Table 3. Comparing the geographical distribution of inventor and examiner-citations (share of citations with inventors from different countries)

Technological sub-fields	All observations	Examiner citations	Inventor citations
Electrical Components Electronics	57.04	60.07	31.70
Audio–visual	50.53	51.52	30.62
Telecommunications	63.19	64.75	30.77
Information Technology	52.20	53.29	25.30
Semiconductors	53.41	55.03	26.55
Optical Instruments	47.69	49.53	33.92
Analytical, measurement & control instruments	59.00	61.77	32.07
Medical equipment	55.43	57.76	32.98
Nuclear technology	55.96	60.60	28.53
Organic chemistry	48.61	50.98	40.10
Macromolecular chemistry	49.51	51.80	35.98
Chemical processes: oil	49.19	51.34	36.69
Surface treatment	54.28	56.75	32.33
Materials–metals	53.54	56.79	36.43
Biotechnology	50.05	52.38	35.71
Pharmaceuticals–cosmetics	50.00	51.81	35.36
Food & agricultural products	56.30	58.10	44.51
Technological processes	57.14	60.38	33.09
Product handling printing	55.58	58.45	32.62
Agricultural machinery food processing	60.09	64.20	36.14
Materials handling	56.10	59.88	34.42
Environment–pollution	60.47	63.35	32.54
Machine tools	59.56	63.58	34.40
Motors–pumps–turbines	57.16	59.99	27.85
Thermal processes	60.69	64.26	27.10
Mechanical components	58.69	62.69	28.68
Transport	59.07	62.23	31.61
Space–arms	59.73	64.14	31.78
Household equipment and consumer goods	60.45	63.95	35.22
Building and public works	59.30	64.01	28.23
Overall	**54.70**	**57.32**	**34.14**

Table 4. Comparing the geographical distribution of inventor and examiner-citations (share of citations with inventors from different regions)

Technological sub-fields	All observations	Examiner citations	Inventor citations
Electrical Components Electronics	66.69	73.12	38.07
Audio–visual	65.26	70.46	33.41
Telecommunications	76.18	80.50	35.21
Information technology	72.22	76.77	37.41
Semiconductors	56.67	64.09	20.87
Optical Instruments	55.44	62.97	30.73
Analytical, measurement & control instruments	69.13	74.97	38.28
Medical equipment	72.44	77.38	45.60
Nuclear technology	59.20	67.04	28.88
Organic chemistry	42.63	47.16	31.75
Macromolecular chemistry	46.38	50.46	33.93
Chemical processes: oil	53.41	56.92	39.85
Surface treatment	58.82	65.09	33.79
Materials–metals	56.55	62.45	38.07
Biotechnology	51.89	56.22	35.90
Pharmaceuticals–cosmetics	59.26	64.21	39.94
Food & agricultural products	68.55	73.04	50.00
Technological processes	62.88	68.81	35.98
Product handling printing	68.72	74.47	41.03
Agricultural machinery food processing	68.73	72.91	47.24
Materials handling	61.26	67.66	37.73
Environment–pollution	72.48	76.69	43.48
Machine tools	65.32	71.78	39.39
Motors–pumps–turbines	62.38	67.50	33.70
Thermal processes	70.32	75.54	37.77
Mechanical components	67.58	73.63	36.07
Transport	70.04	74.56	43.19
Space–arms	67.41	74.40	39.06
Household equipment and consumer goods	68.64	74.17	40.22
Building and public works	72.23	78.14	40.10
Overall	72.79	68.57	37.27

Table 5. Variable definitions

Name	Definition
Cits_examiner	1 if examiner citation, 0 if applicant citation
DistanceKM	Average km distance between the citing and cited region, in units of 173 km
Diff_Regions	0 if at least one inventor in the citing and cited patent application are resident in the same region, 1 otherwise
Diff_Ctrys	0 if at least one inventor in the citing and cited patent application are resident in the same country, 1 otherwise
Diff_MNE	0 if citing and cited patents have at least a ultimate parent in common, 1 otherwise
Citation lag	Priority year of the citing patent application – priority year of cited patent application
Diff_Tech	0 if citing and cited patent application are classified in the same four-digit IPC code
Citing_Granted	1 if the citing patent application has been granted, 0 otherwise
Cited_Granted	1 if the cited patent application has been granted, 0 otherwise
ClassA	1 if the cited patent has been classified under category A, 0 otherwise
ClassY	1 if the cited patent has been classified under category Y, 0 otherwise
ClassX	1 if the cited patent has been classified under category X, 0 otherwise

Table 6. Descriptive statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max
Cits_examiner	233364	0.809	0.393	0	1
DistanceKM	233364	2.449	2.550	0	24.39
Diff_Regions	233364	0.626	0.484	0	1
Diff_Ctrys	866867	0.547	0.497	0	1
Diff_MNE	50945	0.381	0.486	0	1
Citation lag	233364	5.444	3.827	-15	23
Diff_Tech	233364	0.313	0.464	0	1
Citing_Granted	233364	0.527	0.499	0	1
Cited_Granted	233364	0.748	0.433	0	1
ClassA	233364	0.661	0.473	0	1
ClassY	233364	0.147	0.354	0	1
ClassX	233364	0.192	0.394	0	1

Table 7. Correlation matrix for the variables used in the regressions

Variable		1	2	3	4	5	6	7	8	9	10	11
Cits_examiner	**1**	1										
DistanceKM	**2**	0.2081	1									
		(0.000)										
Diff_Regions	**3**	0.2546	0.6576	1								
		(0.000)	(0.000)									
Self-citation	**4**	0.2222	0.5905	0.7701	1							
		(0.000)	(0.000)	(0.000)								
Citation lag	**5**	0.0328	0.1245	0.1838	0.1804	1						
		(0.000)	(0.000)	(0.000)	(0.000)							
Diff_Tech	**6**	0.0223	0.0467	0.0537	0.0498	0.0571	1					
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)						
Citing_Granted	**7**	-0.0529	-0.0709	-0.0605	-0.0629	-0.0939	-0.0308	1				
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)					
Cited_Granted	**8**	-0.1036	-0.0933	-0.0923	-0.0826	0.059	-0.0156	0.1179	1			
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)				
ClassA	**9**	-0.0577	-0.0141	-0.0017	-0.0008	0.0507	-0.0105	0.1021	0.0401	1		
		(0.000)	(0.000)	(0.403)	(0.854)	(0.000)	(0.000)	(0.000)	(0.000)			
ClassY	**10**	-0.0359	-0.0222	-0.0294	-0.0356	-0.0138	0.0167	-0.0145	0.0068	-0.5802	1	
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	0.001	(0.000)		
ClassX	**11**	0.1017	0.0369	0.0286	0.0337	-0.0485	-0.0024	-0.1097	-0.0544	-0.6805	0.0337	1
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.249)	(0.000)	(0.000)	(0.000)	(0.000)	

Significance levels of each correlation coefficient are reported below each coefficient.

Table 8. Results of different logit models using the full sample of European citations

	Random effects cloglog	Random effects logit	Logit with robust cluster errors	Scobit with robust cluster errors	Cloglog with robust cluster errors
DistanceKM	1.185	1.415	1.330	4.981	1.116
	(77.85)***	(85.32)***	(74.25)***	(62.93)***	(69.14)***
Citation lag	1.005	1.008	1.006	1.026	1.004
	(5.24)***	(4.25)***	(4.09)***	(3.32)***	(5.81)***
Diff_Tech	1.058	1.092	1.056	1.230	1.037
	(6.78)***	(5.71)***	(4.25)***	(3.20)***	(5.82)***
Citing_Granted	0.92	0.862	0.905	0.624	0.949
	(9.93)***	(9.64)***	(8.01)***	(7.49)***	(8.49)***
Cited_Granted	0.702	0.509	0.577	0.054	0.776
	(37.59)***	(37.55)***	(37.69)***	(33.80)***	(38.27)***
ClassX	1.578	2.384	2.024	44.31	1.377
	(40.86)***	(40.43)***	(37.65)***	(33.80)***	(39.22)***
ClassY	0.917	0.855	0.913	0.637	0.956
	(8.05)***	(8.01)***	(5.78)***	(5.89)***	(5.48)***
Observations	233364	233364	233364	233364	233364
Number of citing pats	149546	149546			
Log-likelihood	-103603	-103339	-105217	-105031	-105864
AIC	207224.3	206696.1	210450.3	210079.6	211744.7
BIC	207317.5	206789.3	210533.2	210172.8	211827.6
Min cited per citing	1	1			
Avg cited per citing	1.56	1.56			
Max cited per citing	23	23			
Wald χ^2	8246.7	10310.38			
Degrees of freedom	7	7			
ρ	0.27	0.36			
χ^2	4522.45	3756.25			
α	1.001	1.002	1.002	0.147	1.001
χ^2				372.75***	

Absolute value of z statistics in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 9. Results of different logit models using the sample of MNEs' patents

	Random effects clog-log	Random effects Logit	Random effects Logit	Logit with robust cluster errors	Scobit with robust cluster errors	Cloglog with robust cluster errors
DistanceKM	1.103	1.443	1.206	1.161	1.138	1.069
	(17.58)***	(37.54)***	(17.65)***	(16.90)***	(6.28)***	(17.33)***
Diff_MNE	1.842		2.945	2.283	2.069	1.56
	(27.19)***		(26.62)***	(25.07)***	(7.53)***	(28.04)***
Citation lag	0.979	0.98	0.964	0.974	0.977	0.985
	(8.84)***	(5.11)***	(8.96)***	(8.39)***	(6.21)***	(8.20)***
Diff_Tech	0.962	0.93	0.921	0.932	0.942	0.974
	(2.27)**	(2.44)**	(2.76)***	(2.90)***	(2.36)**	(1.99)**
Citing_Granted	0.943	0.906	0.905	0.938	0.945	0.966
	(3.37)***	(3.27)***	(3.27)***	(2.66)***	(2.52)**	(2.66)***
Cited_Granted	0.719	0.525	0.546	0.613	0.653	0.789
	(16.17)***	(17.68)***	(16.42)***	(16.68)***	(6.48)***	(16.04)***
ClassX	1.399	1.83	1.815	1.581	1.499	1.272
	(15.19)***	(15.31)***	(14.92)***	(13.63)***	(7.15)***	(14.26)***
ClassY	0.87	0.778	0.79	0.856	0.87	0.916
	(6.18)***	(6.59)***	(6.10)***	(5.04)***	(4.37)***	(4.98)***
Observations	50945	50945	50945	50945	50945	50945
Number of citing pats	34065	34065	34065			
Log-likelihood	-26742.4	-27118.5	-26726.8	-27223	-27222.5	-27236.5
AIC	53504.78	54255.06	53473.61	54465.08	54491.08	53504.78
BIC	53593.17	54334.61	53562	54553.47	54570.62	53593.17
Min cited per citing	1	1	1			
Avg cited per citing	1.5	1.5	1.5			
Max cited per citing	17	17	17			
Wald χ^2	2215.23	1764.06	2481.81			
Degrees of freedom	8	7	8			
ρ	0.27	0.28	0.36			
χ^2	988.3	1051.08	992.39			
α					1.29	
χ^2					0.92	

Absolute value of z statistics in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 10. Results of different specification of the random effects logit model

	1	2	3	4	5	6	7	8
DistanceKM	1.415 (85.32)***			1.418 (86.16)***	1.416 (85.50)***	1.206 (47.09)***	1.348 (62.07)***	1.216 (37.50)***
Diff_Region		5.067 (97.54)***						
Diff_Ctrys			3.629 (127.46)***					
Citation lag	1.008 (4.25)***	0.995 (2.33)**	0.99 (7.98)***	1.006 (3.00)***	1.009 (4.53)***	1.001 (0.48)	1 (0.02)	0.995 (1.74)*
Diff_Tech	1.092 (5.71)***	1.084 (5.17)***	1.151 (14.17)***	1.082 (5.16)***		1.024 (1.31)	1.071 (3.75)***	1.026 (1.20)
Citing_Granted	0.862 (9.64)***	0.851 (10.34)***	0.677 (36.92)***	0.841 (11.30)***	0.86 (9.77)***	0.966 (2.03)**	0.992 (0.49)	1.065 (3.09)***
Cited_Granted	0.509 (37.55)***	0.519 (36.14)***	0.516 (60.12)***	0.504 (38.38)***	0.509 (37.59)***	0.62 (23.37)***	0.548 (27.29)***	0.607 (19.54)***
ClassX	2.384 (40.43)***	2.437 (40.87)***	2.273 (62.45)***		2.383 (40.43)***	1.729 (23.18)***	1.868 (24.74)***	1.541 (14.91)***
ClassY	0.855 (8.01)***	0.869 (7.10)***	0.965 (2.85)***		0.857 (7.91)***	0.883 (5.38)***	0.8 (10.04)***	0.79 (8.66)***
ClassXY				1.414 (22.42)***				
Observations	233364	233364	866867	233364	233364	68939	208205	43780
Log-likelihood	-103339	-102232	-282439	-104131	-103355	-42883.1	-60055.5	-28771.2
Number of citing pats	149546	149546	444670	149546	149546	37743	127972	16169
Min cited per citing	1	1	1	1	1	1	1	2
Avg cited per citing	1.56	1.56	1.95	1.56	1.56	1.83	1.63	2.71
Max cited per citing	23	23	29	23	23	23	23	23
Wald χ^2	10310.38	12127.7	24792.31	9600.76	10300.66	3631.27	5669.63	2236.35
degrees of freedom	7	7	7	6	6	7	7	7
ρ	0.36	0.37	0.46	0.35	0.36	0	0.16	0
χ^2	3756.25***	3958.7***	21813.7***	3722.73***	3744.09***	0	728.09***	0

Absolute value of z statistics in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 11. Estimate of the random effects logit model using 30 sub-samples determined by the technological class of the citing patent

	Parameters			
	Odds ratios Mean	**number of non-significant odds ratios**	**number of odds ratios>1**	**t-stat Mean***
DistanceKM	1.416	0	30	35.760
Citation lag	1.020	12	25	73.940
Diff_Tech	1.153	11	24	9.703
Citing_Granted	0.920	13	10	9.592
Cited_Granted	0.546	2	0	8.232
ClassX	3.667	1	30	3.281
ClassY	0.865	16	7	7.968

*This is computed as the average parameter divided by the average standard error