Research Policy 48 (2019) 1771–1780

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

The geography of scientific citations

Mignon L. Wuestman*, Jarno Hoekman, Koen Frenken

Innovation Studies, Copernicus Institute of Sustainable Development, Utrecht University, the Netherlands

ARTICLE INFO

Keywords:
Geography of scientific knowledge
Spatial scientometrics
Citation analysis
Knowledge

ABSTRACT

Science's main norms prescribe scientists to use citations as acknowledgements of cognitive content irrespective of geographical location. Previous studies, however, suggested that there is a considerable geographical bias in scientific citations. We argue that this geographical bias does not, in itself, falsify the notion that citations reflect acknowledgement of cognitive content, because cognitively related knowledge may be geographically concentrated as well. We analyse the role of organizational, regional and national co-location on citation likelihood for 5.5 million article pairs, and find that the geographical bias in citations is weak once cognitive relatedness is accounted for. Furthermore, we find that the effect of co-location on citation likelihood is strongest at the organizational level, weaker at the regional level, and weakest at the national level. In addition, we show that geographical co-location particularly increases the citation likelihood between two papers when knowledge relatedness between articles is low, suggesting that interdisciplinary research benefits most from co-location. Finally, we find that, when knowledge relatedness is high, the effect of geographical co-location on citation likelihood is non-existent. We discuss the implications regarding policies aimed to discourage strategic citations and to foster interdisciplinary research.

1. Introduction

Accumulation of scientific knowledge is often considered to be a global communal endeavour. Science's norms prescribe scientists to build on each other's work based on mere cognitive considerations and irrespective of personal and geographical qualities (Hagstrom, 1965; Merton, 1942). This is reflected in normative theories of citations which consider a scientific citation as acknowledgment of intellectual debt and cognitive influence (Baldi, 1998; Furman et al., 2006; Small, 1978). It would follow that citations are based on the topical, methodological or intellectual content of research findings, and are selected irrespective of geographical location of the cited authors. In this sense, citations would be considered 'placeless' (Livingstone, 2003).

In spite of the assumed 'placelessness' of citations, a number of studies have shown a geographical bias in scientific citations. For example, Matthiessen et al. (2002) found that citations occur most frequently within countries rather than between countries. Börner et al. (2006) showed that citations are concentrated within organizations, and also provided evidence of distance- decay in citations between research organizations. And, Pan et al. (2012) found that citation likelihood decreases with distance between cities.

One possible explanation for these findings holds that co-location between teams of scientists within organizations, regions and countries facilitates face-to-face interaction between scientists necessary to transfer tacit knowledge that is required to accumulate upon existing knowledge (Collins, 2001). This interpretation is in line with previous research on the geography of patent citations (Breschi and Lissoni, 2009). Geographical biases may equally be understood from social constructivist theories that stressed that citations are not merely a sign of intellectual acknowledgement, but also tools of persuasion (Baldi, 1998; Gilbert, 1977). In this view, scientists may preferentially cite eminent scholars and reputed organizations over equally relevant sources from scholars and organizations less well-known. Hence, to the extent that the reputation of scholars and organizations is geographically bounded, for example, being rather specific to particular countries or language areas, citation flows are likely to display a geographical bias.

We argue, however, that the current understanding of the geography of scientific citations is ill-founded. The simple fact that the probability of a scientific citation between two articles declines with geographical distance does not necessarily imply that citations are made for other reasons than acknowledging codified intellectual debt. For such a conclusion to hold, one should properly disentangle the effect of geography on citations from other relevant factors. To the extent

* Corresponding author at: Innovation Studies, Copernicus Institute of Sustainable Development, Utrecht University, Princetonlaan 8A, 3584CB, Utrecht, the Netherlands.
E-mail address: m.l.wuestman@uu.nl (M.L. Wuestman).

https://doi.org/10.1016/j.respol.2019.04.004
Received 12 March 2018; Received in revised form 25 March 2019; Accepted 5 April 2019
Available online 27 April 2019
0048-7333/

that places concentrate their research efforts on certain topics (Boschma et al., 2014), citations reflecting intellectual debt will also be more geographically concentrated (Head et al., 2018). Hence, the geographical bias in citations may, in principle, be fully explained by the geographical concentration of intellectually related knowledge.

Moreover, while literature on patent citations has shown a geographical bias once controlled for cognitive factors (Breschi and Lissoni, 2009; Jaffe et al., 1993), it is doubtful whether these findings can be extrapolated to the scientific domain. Patent citations and scientific citations are distinctly different (Dasgupta and David, 1994). First, patent citations are meant to indicate what is similar but excluded from the patent's claims, rather than what is included (Duguet and Macgarvie, 2003; Meyer, 2000). Second, patent examiners often add citations to similar patents, so that a citation does not imply the author's awareness of the referenced knowledge (Alcácer and Gittelman, 2006; Duguet and Macgarvie, 2003; Meyer, 2000).

The aim of our study is to understand the geographical bias in scientific citations. In order to do so, we estimate the effect of co- location between teams of authors of scientific articles at multiple geographical levels: the organizational, regional and national level, while taking into account the relatedness of the knowledge codified in articles. In addition, we study whether the effect of co-location on scientific citations at multiple geographical levels differs depending on the relatedness of the knowledge in the cited and citing articles. In particular, we are interested in the question whether citations between otherwise unrelated articles are more likely between co-located authors. As a methodological contribution, we isolate the geographical bias in citations at each of the three geographical levels (organizational, regional, and national), by introducing a new way to characterize the geographical co-location between teams, designed so that the organizational, regional and national level are fully independent of each other.

The insights we gain regarding the geography of scientific citations are threefold: first, while we find a geographical bias in citations, this effect is small when the effect of knowledge relatedness is controlled for. Second, the effect of geographical co-location between teams is strongest at the organizational level and weakest at the national level. Third, we show that co-location particularly increases the likelihood of citation when knowledge relatedness is low. For highly related articles, geographical effects are non-existent or even negatively associated with citation likelihood. The outcomes of our research suggest that the institutionalization of science in localized campuses may be especially beneficial for fostering the creation of scientific links for producing unrelated knowledge (Wang et al., 2017).

2. Theoretical background

Scientific knowledge is produced in almost every country across the globe. Scientists are organized in global epistemic communities that codify their knowledge in peer- reviewed articles published in specialist international journals. Citations to these articles are understood as a normative institution, representing acknowledgement of the intellectual content described in articles for everyone to be seen (Baldi, 1998; Kaplan, 1964). This view implies that the accumulation of scientific knowledge is a placeless process, where scientists have the means to access, and to accumulate upon, any cognitive, methodological or topical content that they consider relevant, and decide on the relevance of content irrespective of its author or place of origin. As such, citations provide an indication that previous knowledge is deemed relevant, and support the relevance of the referencing article. Citing thus underlies the accumulation of relevant scientific knowledge. As citing is highly selective, the process, by implication, also leads to the erosion of less relevant and less useful knowledge. An important question therefore becomes what determines the relevance of scientific knowledge for a citing scientist.

To be able to judge the relevance of knowledge introduced in a scientific article, knowledge relatedness seems to be by far the most important factor (Breschi and Lissoni, 2009). In relation to scientific articles, knowledge relatedness can be understood to mean the extent to which the reader of the article has the necessary knowledge to de-code the codified knowledge stored in the article and to actively use it when developing new knowledge (Breschi et al., 2003; Frenken, 2010). Only when knowledge relatedness is sufficiently high, the reader will be able to use the knowledge as input in the production of new knowledge. In this sense, knowledge relatedness is consistent with the idea that knowledge can be transmitted as information as long as an author and a reader use the same code of communication, to code and decode the message (Cowan et al., 2000).

Besides the relatedness of knowledge, geography may in practice also influence a reader's judgment of the relevance of knowledge, thus affecting citations, or the absence hereof, that follow from such judgments. While scientific articles are written in order to codify knowledge, tacit knowledge is not transmitted with the text. Tacit knowledge in this context has been defined as: 'knowledge or abilities that can be passed between scientists by personal contact but cannot be, or have not been, set out or passed on in formulae, diagrams, or verbal descriptions and instructions for action.' (Collins, 2001, p. 72). Shared tacit knowledge is useful for the effective use of codified knowledge as an input in the reader's own research, for example, to be able to use similar instruments, software, materials used in the author's previous research (Balconi et al., 2007; Collins, 1985; Nightingale, 2004). For two scientists to get in contact, however, they have to spend time and resources. As the costs of contact roughly increase with geographical distance, low geographical distance may support tacit knowledge transfer benefitting a reader's ability to judge and use knowledge. And, since co-located scientists often have many colleagues in common, the local network also provides an infrastructure to cross- check the validity of knowledge claims.

Geographical biases may also result from strategic considerations. As tools of persuasion (Gilbert, 1977), scientists may select the relevant citations 'based on the location of a cited paper's author within the stratification structure of science rather than on the worth or contents of the paper itself so as to enlist the support of eminent authors and thus convince readers of the validity of their arguments' (Baldi, 1998, p. 832). This implies that the small minority of authors who enjoy high status and visibility will be preferentially cited over the large majority of low-status scholars, a phenomenon known as the Matthew effect (Merton, 1968). Importantly, an author's level of eminence will vary among scholars worldwide, and is expected to be, at least to some extent, specific to their location(s) of work. Generally, then, eminence will decline with geographical distance, suggesting that citation flows driven by considerations of eminence display a geographical bias indeed.

Both from the perspective of shared tacit knowledge and the perspective of the role of status, one can derive that citations are more probable when authors are co-located. Because we expect that both the transfer of tacit knowledge and the effect of status are functions of distance, and that the effect of shared coordination and institutions is stronger at lower levels, the effect of geographical co-location of scientific teams is expected to vary across different geographical levels. As organizations are nested within regions, which are in turn nested within countries, we expect that the effect of co- location between teams of authors is largest at the organizational level, and smallest at the national level.

Furthermore, the effect of co-location on citation may also vary with different levels of knowledge relatedness. In particular, literature on economic geography suggests that geographical co-location can compensate for a lack of knowledge relatedness (Boschma, 2005). Scientists generally lack the interpretative skills and the tacit knowledge required to understand a knowledge claim in a field unrelated to their own field. Yet, such unrelated knowledge can still be explained through face-to-face interaction. Moreover, whereas global communities are organized around bodies of related knowledge that are discipline or topic-specific,

making scientists aware of new developments in their field, local serendipitous encounters make scientists aware of developments in fields unrelated to their own field. It is likely that scientists would otherwise remain unaware of these unrelated bodies of knowledge. Also note in this context that regional and national research policies may be explicitly directed to bring unfamiliar scientists together (Fitjar and Rodríguez-Pose, 2017). In all, we expect that co-location is especially important for citations to occur between unrelated knowledge.

3. Research design and data

3.1. Data construction

Our research design is centred around estimating the probability that a scientific article cites another scientific article. Our data was derived from two bibliometric sources. Data from both sources was collected in October 2016. First, we used the NIH National Library of Medicine PubMed database to extract data on a comprehensive list of life sciences and medicine articles. We collected this data using an API from the 'RISmed'- package in R (Kovalchik, 2016). We chose to start the data construction from this database because it offers a comprehensive and coherent body of literature, reflecting a broad range of scientific disciplines, methods and topics. Moreover, the PubMed database has high quality indexing of journals, journal issues, and author names. We subsequently used the Institute for Scientific Information's Web of Science database to extract additional information on all publications derived from PubMed. This data included author affiliations and addresses and the full reference lists of all publications.

We limited data extraction to all English language journal articles that were indexed in both databases and were published in a journal in either 2012 or 2014. We excluded i) review articles, since citations in these articles are distinctly different from non-review articles (Biscaro and Giupponi, 2014; Vanclay, 2013), ii) articles for which the reference list was missing or did not include any DOI, and iii) articles for which address details on the author level were missing. Additionally, for each included article, single references that did not include a DOI, for example because they are references to non-published material, were excluded. On average, 74.92 per cent (sd 16.57) of all references per 2012 article include a DOI, compared to 77.24 (sd 16.09) per cent for 2014. These missing DOI values do not affect the dependent variable because all scientific articles covered by PubMed and Web of Science have a DOI. They do, however, have some effect on the estimation of *bibliographic coupling* between articles (see below) by underestimation actual *bibliographical coupling*. We expect this underestimation of *bibliographic coupling* to be random because reference lists in Web Of Science are standardized, so that any cited document that has a DOI in one reference list is expected to also have a DOI in another reference list, and because the percentage of references without DOI is similar across articles, as reflected by the low standard deviations.

For each author listed on an article, address details consist of one or multiple affiliations, with their zip code, city and country. For some countries, such as the United States, the author's state is also mentioned. To capture affiliation, city and country, all addresses were geolocated on the city-level in order to compute geographical distances between addresses. First, all addresses were matched to a list of countries using the R 'Maps'-package (Becker, Wilks, & Brownrigg, 2017), which uses data from the Natural Earth Project (NaturalEarth, 2017). In case the Web of Science database used other country names than the R Package, country names were manually recoded. The 'Maps'-package also contains a list of all cities with over 40,000 inhabitants, including their names and coordinates. Addresses in the United States (US) were further matched to a list of US states to prevent mistakes due the occurrence of the same city names in multiple states. Second, for each country or state (in case of US), addresses were matched to the list of cities within that country available in the 'Maps' package. During matching, longer names were prioritized over shorter names. This way, 93.28% of all addresses were matched to a city and its coordinates. Addresses that were not matched were excluded from the analysis. Next, affiliation data was extracted from each of the addresses. Because many, but not all, affiliations use standardized abbreviations ('univ' instead of 'university'; 'hosp' instead of 'hospital'), we replaced those entries that use the full term with its abbreviation to improve consistency. Because we distinguish between affiliations which may, together, be part of an overarching affiliation but which are located in different areas, we combine affiliation data with city data to get unique affiliations per city. A manual check of 1000 addresses containing affiliations, cities and countries revealed no errors.

To study citations between articles, couples of articles consisting of one article published in 2014 and one published in 2012 were chosen to serve as the unit of analysis in this study. We therefore constructed a dataset of article couples that allowed us to model the likelihood that a 2014 article cites a 2012 article. A two-year window was taken to ensure that authors of an article published in 2014 could be aware of the existence of the article published in 2012, while at the same time minimising biases in the measurement of co-location due to the fact that authors might have been mobile between the two periods. Ultimately, our dataset consists of 5,591,571 couples

3.2. Dependent variable

Our dependent variable Cit_{ij} is defined as the presence of a citation link between an article i from 2012 and an article j from 2014. Since all journal articles are accompanied by a unique DOI, a citation link from j to i was considered to be present when the DOI-based reference list of article j from 2014 contained the DOI of article i from 2012. Cit_{ij} was coded 1 when a citation link was present, and 0 when it was not.

3.3. Independent variables

3.3.1. Knowledge relatedness

To measure the extent to which the knowledge described in one article is related to another article, we compute the absolute number of references that are shared by articles i and j. This measure is known as *bibliographic coupling* (*BC*) and was originally proposed as a method for grouping scientific documents based on similarity (Kessler, 1960) and for document retrieval (Kessler, 1962). Bibliographic coupling can be used to measure a cognitive relationship between two articles (Jarneving, 2007; Peters et al., 1995). A key advantage of bibliographic coupling over other measures holds that there is no delay for its calculation, since all information is present upon publication and will remain constant over time (Glanzel and Czerwon, 1996; Thijs et al., 2015). For our data, bibliographic coupling was calculated by comparing the DOI-based reference lists of the 2014 article and the 2012 article for all couples in our sample. *Bibliographic coupling* is calculated as follows:

$$BC_{ij} = \sum_{r=1}^{n} E_{ri} \times E_{rj} \quad (1)$$

where BC_{ij} represents bibliographic coupling for a couple comprising article i from 2012 and article j from 2014; $r = 1, \ldots, n$ represents the total set of references; and E is a matrix of references by article.

3.3.2. Co-location at the organizational, regional and national level

3.3.2.1. Organizational coupling. To measure co- location between two teams at the organizational level, we introduce a measure of *organizational coupling* (OC). *Organizational coupling* is defined as the number of author affiliations that are shared, relative to the maximum possible number of author affiliations that may be shared, by articles i from 2012 and j from 2014. As authors can have more than one affiliation, we consider multiple addresses as separate unique 'author affiliations'. Affiliations that are situated in more than one location are

considered different affiliations, fitting with the nestedness of our variables. For the set of couples, it is calculated as follows:

$$OC_{ij} = \frac{\sum_{s=1}^{w} F_{si} \times F_{sj}}{L_i \times L_j} \quad (2)$$

where OC_{ij} represents organizational coupling for a couple comprising article i from 2012 and article j from 2014; $s = 1, \ldots, w$ represents the total set of author affiliations; and F is a matrix of author affiliations by article, based on the exact affiliation names extracted from the Web of Science data. L_i and L_j represent the total number of author addresses of articles i and j respectively. As an example, consider a couple with an article i from 2012 with one author A from MIT and one author B from Harvard, and an article j from 2014 with one author C from MIT and another D from Boston University. The maximum possible number of shared authors affiliations is four, because author affiliations can be shared between A and C, A and D, B and C and B and D. Authors A and C are coupled, while the other three potential couples remain uncoupled. Then, OC_{ij} equals $\frac{1}{4}$.

3.3.2.2. Regional coupling. Co-location at the regional level is measured using a similar measure of *regional coupling* (RC). We define *regional coupling* as the number of author locations that are shared, relative to the maximum possible number of author locations that may be shared, by articles i from 2012 and j from 2014, minus the organizational coupling between them. In doing so, we isolate organizational coupling from regional coupling, because affiliations are nested within regions. An author location is considered to be shared if the location of an author listed on the 2014 article is within a distance of 20 km from the location of an author listed on the 2012 article and within the same country. In other words, regional coupling expresses the share of authors that are situated in the same region comprising a 20 km radius and within the same country, yet without being affiliated to the same organization. Again, one authors' multiple locations are considered unique 'author locations'. *Regional coupling* is calculated as follows:

$$RC_{ij} = \frac{\sum_{t=1}^{p} G_{ti} \times GK_{tj}}{L_i \times L_j} - OC_{ij} \quad (3)$$

where RC_{ij} is regional coupling for a couple with article i from 2012 and article j from 2014; $t = 1, \ldots, p$ represents the total set of author cities; G is a matrix of author cities by article of length n; and GK is a matrix of author cities by article, where authors are counted when they are located in a city or in cities within 20 km of their listed city and within the same country. L_i and L_j again represent the total number of author addresses of articles i and j respectively. In the example we used previously, both authors A and B of the 2012 article are located in Cambridge MA, while author C of the 2014 article is located in Cambridge MA and author D in Boston MA. A and B are thus both coupled with C, but as Boston is within a 20 km vicinity of Cambridge, A and B are also coupled with D. Further note that we subtract $\frac{1}{4}$ for OC_{ij}, because authors A and C are already coupled at the organizational level. RC_{ij} thus equals $\frac{4}{4}$ minus OC_{ij}, equals $\frac{3}{4}$.

Given that the nestedness of the coupling measures introduces dependency between the measures, we also test whether the effect of regional coupling is robust when nestedness is not assumed, by including two alternative measures of *RC*. First, we subtract the number of shared author affiliations from the denominator, so that regional coupling is only relative to the maximum possible number of coupled authors, which were not already coupled at the organizational level. Second, we compute RC_{ij} without subtracting OC_{ij}.

3.3.2.3. National coupling. For co-location of teams at the national level, we use a measure of *national coupling* (NC), which we define as the number of authors' countries that are shared, relative to the maximum possible number of unique author countries that may be shared, by articles i from 2012 and j from 2014, minus the regional and organizational coupling between them. *National coupling* is, thus, the share of authors that are listed in the same country, without being listed in the same region or, which follows, the same organization. *National coupling* is calculated as follows:

$$NC_{ij} = \frac{\sum_{u=1}^{q} H_{ui} \times H_{uj}}{L_i \times L_j} - RC_{ij} - OC_{ij} \quad (4)$$

where NC_{ij} is *national coupling* for a couple with article i from 2012 and article j from 2014; $u = 1, \ldots, q$ represents the total set of author countries; and H is a matrix of author countries by article. Again, L_i and L_j represent the total number of author addresses of articles i and j respectively. All authors in our earlier example are located in the United States, so both authors A and B from the 2012 article are coupled with both authors C and D from the 2014 article, so that all possible shared author countries are indeed shared. However, we subtract RC and OC, so that we only include those situations in which authors are co-located at the national level, but not at the regional or the organizational level. Thus NC_{ij} is $\frac{4}{4} - \frac{3}{4} - \frac{1}{4}$, equals 0.

Again, we also examine the effect of national coupling when nestedness is not assumed. We include a measure of *NC* where we subtract the number of shared author affiliations and locations from the denominator, so that national coupling is only relative to the maximum possible number of coupled authors, which were not already coupled at the organizational or regional level. Alternatively, we include a measure of NC_{ij} where RC_{ij} and OC_{ij} are not subtracted from NC_{ij}.

3.4. Control variables

3.4.1. Shared authors

We control for self-citations, as several studies have shown that such citations differ from other citations. For example, Zaggl (2017) found that the way the scientific system is organized provides incentives towards unjustified self-citation (i.e. citations with overlap in authors, but no overlap in content), and that mechanisms that prohibit such manipulation are limited. Bartneck and Kokkelmans (2011) found similar results when simulating how self-citations inflate an author's h-index, and Seeber et al. (2017) found that scientists use self-citations to strategically respond to citation metrics. While it is noteworthy that not all self-citation is, of course, unjustified, we control for this effect by including *shared authors (SA)* as a measure of self- citation alongside our measure of *BC*.

Since author names are not standardized in our dataset, for each couple, we computed the sum of the 2014 authors whose last name and first initial exactly match with those of 2012 authors, and divided this by the total number of authors of the 2014 article. Thus, this variable accounts for a potential higher probability of citation when an article couple shares author names, i.e. self-citation. Because of potential inaccuracy in this measure, we replaced the sum of exact matches with the sum of the 2014 authors whose last name and first initial have a Jaccard similarity of $> .8$ in a robustness check. This measure takes into account the proportion of common letters, as well as individual letters' positions within names.

3.4.2. Times cited

We included the number of citations that the 2012 article has received *(times cited (TC))* as a control. Highly-cited articles are more visible and more highly regarded in the scientific literature, which might influence further citation likelihood. Note that this measure captures both perceived quality differences between articles and the sheer Matthew Effect, where further scientific rewards accrue to prominent scientists based on their status alone (Azoulay et al., 2013; Barabási and Albert, 1999; Merton, 1968).

3.4.3. Product of number of references and product of number of authors

As the number of references per article varies and the length of the reference lists of both articles in a couple affects the likelihood that references are shared, the product of the length of the reference lists of article *i* from 2012 and article *j* from 2014 was included as a control variable, and is referred to as *references product*. Similarly, the product of the number of authors of article *i* from 2012 and article *j* from 2014 was included as a control variable as well, and is referred to as *authors product*.

3.5. Sampling

The large number of observations in our dataset (616,490*674,196 couples) combined with the fact that there are very few $Cit_{ij} = 1$ observations relative to $Cit_{ij} = 0$ observations (approximately 1 out of each 3000 couples), makes meaningful estimation using all article couples or a random sample impossible. To create a sample, we therefore used choice-based sampling on the $Cit_{ij} = 0$ observations while weighting each sampled observation by the number of elements it represents (Singh, 2005). This approach ensures that no bias is introduced in our model, even though stratification is carried out on our dependent variable. For each $Cit_{ij} = 1$ couple, three unique $Cit_{ij} = 0$ couples were sampled, conditional on these three couples having the same 2014-article as the $Cit_{ij} = 1$ couple, and on the 2012-article being published in the same journal issue as the 2012-article in the $Cit_{ij} = 1$ couple, so that there is some empirical basis for citation of the sampled $Cit_{ij} = 0$ couples.[1] The weight of the sampled $Cit_{ij} = 0$ observations is relative to the total number of 2012- articles published in the same journal issue as the cited article, so that the total weight of the sampled $Cit_{ij} = 0$ observations equals the total number of articles published in the same journal issue as the cited article. As all $Cit_{ij} = 1$ are included, the weight of $Cit_{ij} = 1$ is always 1. This approach resulted in 5,598,687 couples derived from 490,163 unique 2014 articles and consisting of 4,177,895 couples for which $Cit_{ij} = 0$, and 1,413,676 couples for which $Cit_{ij} = 1$.

In order to control for any variation at the level of the 2014 article that is not captured in our independent variables, we introduce the 2014 article as a random effect term. By including this random effect, we also resolve the fact that the same 2014 article can occur in multiple sampled couples, causing dependence across these couples.

3.6. Analysis

Because our dependent variable is dichotomous and we include random effects, mixed-effect logistic regression is the most appropriate modelling technique to test citation likelihood (Faraway, 2016). All of our mixed effect models include weights for each observation and use the 2014 article as random effect term, as a result of our sampling strategy. Data was analysed using R (R Core Team, 2017), and the 'lme4' package (Bates et al., 2015).

In our first model, we include our control variables *times cited*, *shared authors*, *authors product* and *references product*. Second, we test the effect of *bibliographic coupling* after including all control variables. Third, we fit a 'naïve' model, in which we include our three co-location variables without including *bibliographic coupling*, so as to show the geographical bias in citations when *bibliographic coupling* is not taken into consideration. Fourth, we include both *bibliographic coupling* and our three geographical coupling variables. Fifth, we add interactions between the three geographical coupling variables and *bibliographic coupling*. Including *bibliographic coupling* as a continuous variable in the interaction term, however, does not allow us to fully understand whether the effect of geography is stronger in some situations than in others. As a sixth step, we therefore model citation likelihood for situations where knowledge relatedness is absent ($BC_{ij} = 0$), low ($0 < BC_{ij} < 5$), medium ($5 < BC_{ij} < 10$), or high ($BC_{ij} > 10$), using separate logistic models for each of these four levels, in which we include our control variables and random effects as well. To aid interpretation of these results, we distinguish between situations where there is none ([0]), low ((0-0.33]), medium ((0.33-0.66]), high ((0.66–1)) and full ([1]) coupling for *organizational*, *regional* and *national coupling*, respectively.

4. Results

4.1. Descriptive statistics

Descriptive statistics for all variables can be found in Table 1, and correlations between all variables in Table 2 . In Table 1, one can see that the mean value of each of our variables is higher when $Cit_{ij} = 1$ than when $Cit_{ij} = 0$. Regarding the difference in the means of *times cited*, it is worth noting that all 2012 articles in $Cit_{ij} = 1$ couples have, by definition, been cited at least once. In Table 2, we see that there is a positive correlation between *bibliographic coupling* and the co-location measures, especially *organizational coupling*, suggesting that more related scientific knowledge is also more geographically concentrated. This finding suggests that without controlling for knowledge relatedness, any geographical effect found may in fact be due to a geographical concentration of related knowledge. Further note that the correlation between the different geographical coupling variables is low, following from our nested conceptualization of geographical co-location at organizational, regional and national levels.

4.2. Regression models

We present the results of the regression models in Table 3. In each of the models, all direct effects of independent variables are positive and significant ($p < .0001$), except for *authors product*, which is significant at the $p < .001$ level. The interaction effects introduced in Model 7 are negative and significant ($p < .0001$). The inclusion of all variables increases each model's fit, as expressed by the significant Chi^2-test relative to the previous model. The Chi^2-test reported for Model 3 is relative to Model 1.

Model 1 includes each of the control variables *times cited, shared authors, authors product* and *references product*, and the random effects. While the effect of *times cited* is small, it is worth noting that as some articles get thousands of citations, this variable may still have a considerable impact on the probability of citation. With an odds ratio (OR) of $e^{18.461}$, the effect of *shared authors* on the probability of a citation is large: when *shared authors* increases from 0 to 1, citation probability P ($Cit_{ij} = 1$) increases from 0.012 to 0.999 when all other variables are conditioned on 0.

In Model 2, which further includes *bibliographic coupling*, the effect of *bibliographic coupling* is positive and significant (OR = $e^{2.508}$ =& #8201;12.280), and the model has considerable explanatory power with a conditional R^2 of 0.941. Inclusion of *bibliographic coupling* lowers the coefficient of *shared authors*, which indicates that self-citations are partly the result of knowledge relatedness. It shows that, while P ($Cit_{ij} = 1$) is 0.006 when $BC_{ij} = 0$, a BC_{ij} value of 3 is enough for P($Cit_{ij} = 1$) to approach 1.

Model 3 shows the effects of co-location of teams at different geographical levels on citations without controlling for knowledge relatedness. With a conditional R^2 of 0.558, the models' explanatory power is considerable. The geographical bias in citation is mostly observed at the organizational level (OR = $e^{3.838}$ =& #8201;46.433), and is weaker at the regional (OR = $e^{1.967}$ =& #8201;7.149) and national

[1] However, some journal issues may contain too few unique articles to be able to sample three unique $Cit_{ij} = 0$ articles per $Cit_{ij} = 1$ article. This is especially likely when a 2014 article cites more than one 2012 article from the same journal issue. In these cases, all available $Cit_{ij} = 0$ articles were sampled. Weights of the sampled observations were adjusted accordingly.

Table 1
Descriptive statistics.

	Cit_{ij} = 1 (N = 1,415,402)			Cit_{ij} = 0 (N = 4,366,734)		
	Mean (SD)	Min	Max	Mean (SD)	Min	Max
Bibliographic Coupling	3.225 (3.863)	0	116	0.022 (0.237)	0	75
Organizational Coupling	0.105 (0.246)	0	1	0.002 (0.032)	0	1
Region Coupling	0.037 (0.178)	0	1	0.005 (0.055)	0	1
National Coupling	0.221 (0.359)	0	1	0.114 (0.283)	0	1
Times Cited	67.170 (311.400)	1	6,759	16.600 (31.610)	0	5,747
Shared Authors	0.105 (0.224)	0	4	0.001 (0.014)	0	1.33
Authors Product	90 (133)	1	1,924	81.5 (93.9)	1	1,670
References Product	2,403 (3470)	2	504,036	2,457 (2148)	2	570,064

Table 2
Correlation matrix.

	BC	OC	RC	NC	TC	SA	AP
Organizational Coupling (OC)	.363***						
Regional Coupling (RC)	.094***	−.010***					
National Coupling (NC)	.051***	.015***	−.061***				
Times Cited (TC)	.028***	−.006***	−.001*	.031***			
Shared Authors (SA)	.442***	.704***	.206***	.046***	−.008***		
Authors Product (AP)	.002***	−.000	.0001	−.000	.002***	.001*	
References Product (RP)	.145***	.006***	−.015***	.001	.031***	−.001*	−.001

Note: p < .0001 ***, p < .001 **, p < .01 *.

Table 3
Logistic regression model coefficients.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	−4.240***	−5.093***	−4.442***	−5.084***	−5.110***	−5.311***	−5.341***
	(0.002)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Bibliographic Coupling (BC)		2.508***		2.473***	2.475***	2.521***	2.658***
		(0.002)		(0.002)	(0.002)	(0.002)	(0.002)
Organizational Coupling (OC)			3.838***	3.054***	3.307***	3.469***	4.998***
			(0.012)	(0.017)	(0.017)	(0.017)	(0.016)
Regional Coupling (RC)			1.967***		1.709***	1.900***	2.265***
			(0.009)		(0.013)	(0.013)	(0.014)
National Coupling (NC)			0.920***			0.847***	0.962***
			(0.003)			(0.005)	(0.005)
BC*OC							−2.433***
							(0.005)
BC*RC							−1.023***
							(0.011)
BC*NC							−0.342***
							(0.004)
Times Cited	0.007***	0.007***	0.007***	0.007***	0.006***	0.007***	0.006***
	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Shared Authors	18.461***	12.394***	13.868***	9.794***	9.027***	8.663***	9.343***
	(0.020)	(0.027)	(0.022)	(0.029)	(0.029)	(0.030)	(0.029)
Authors product	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
References product	0.000***	−0.0001***	0.000***	−0.0002***	−0.0002***	−0.0002***	−0.0002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Likelihood	−6,010,431	−3,499,078	−5,918,564	−3,480,175	−3,474,730	−3,454,079	−3,415,576
Chi^2		5,021,785***	183,735***	38,727***	10,891***	41,301***	77,006***
observations	5,591,571	5,591,571	5,591,571	5,591,571	5,591,571	5,591,571	5,591,571
N groups	489,667	489,667	489,667	489,667	489,667	489,667	489,667
Marginal R2	0.573	0.870	0.558	0.866	0.870	0.864	0.834
Conditional R2	0.709	0.941	0.701	0.942	0.940	0.942	0.922

Note: Chi^2 of Model 3 is relative to Model 1. Dependent variable: Cit. p < .0001 ***, p < .001 **, p < .01 *.

(OR = $e^{0.920}$ =& #8201;2.509) level.

Comparing Model 3 with Model 6, in which we do control for knowledge relatedness, we see that the inclusion of *bibliographic coupling* results in a large drop in the coefficient of *organizational coupling* and *regional coupling*, and a smaller drop in the coefficient of *national coupling*. The odds ratio of an increase from no *organizational coupling* (OC_{ij} = 0) to full *organizational coupling* (OC_{ij} = 1) drops from 46.433 ($e^{3.838}$) in Model 3 to 32.105 ($e^{3.469}$) in Model 6, whereas the odds ratio for *national coupling* drops from 2.509 ($e^{0.920}$) to 2.333 ($e^{0.847}$). The drop in coefficients and odds ratios confirms the idea that part of the geographical bias in citations is in fact due to a concentration of related knowledge in organizations, regions, and, to a lesser extent, countries. Furthermore, we see that the explanatory power of Model 6 is only minimally higher than that of Model 2, with a conditional R^2 of 0.942 compared to 0.941. While the model has a better fit than Model 2, the additional variance explained by including the geographical variables

when we control for knowledge relatedness is small. Comparing Models 4, 5, and 6, we find that inclusion of one coupling variable does not affect the effect of another, since they were computed to be independent from each other.

4.3. Interaction effects

In Model 7 we include interaction effects between *bibliographic coupling* and the geographical variables. As our three geographical co-location measures are computed on similar scales, we can compare their coefficients and see that, when $BC_{ij} = 0$, the coefficient of *organizational coupling* is about twice as high as the coefficient of *regional coupling*, which is again twice as high as the coefficient of *national coupling*. Looking at the interaction terms, we find a negative and significant ($p < .0001$) effect for all interactions. The negative effect provides us with a first indication that the effect of co-location on citation is smaller when *bibliographic coupling* is high compared to when *bibliographic coupling* is low.

We use separate logistic regression models to further explore this effect and model citation likelihood for no ($BC_{ij} = 0$), low ($0 < BC_{ij} < 5$), medium ($5 < BC_{ij} < 10$), or high ($BC_{ij} > 10$) *bibliographic coupling*, distinguishing between none ([0]), low ((0-0.33]), medium ((0.33- 0.66]), high ((0.66–1)) and full ([1]) coupling for *organizational*, *regional* and *national coupling*. The thresholds used do not affect our results. The odds ratios for citation, and the corresponding 95% confidence intervals, for the different levels of *bibliographic coupling* and the geographical variables are presented in Figure 1. Note that the odds ratios represent the odds of citation for a certain level of *organizational, regional* or *national coupling,* relative to no *organizational, regional or coupling*, and conditional on a level of *bibliographic coupling*. For each of the geographical co-location variables *organizational*, *regional* and *national coupling* we see that the effect of co-location is largest and positive when *bibliographic coupling* is absent or low. We further find that when *bibliographic coupling* is absent or low, *organizational coupling* has a bigger effect on citation likelihood than *regional coupling*, while *national coupling* has the weakest effect. For example, when $BC_{ij} = 0$, citation is 14.788 (OR = $e^{2.694}$) times more likely when the sets of authors are organizationally fully coupled compared to when there is no *organizational coupling*, 4.244 (OR = $e^{1.445}$) times more likely when the sets of authors are fully coupled at the regional level compared to when there is no *regional coupling*, and only 2.311 (OR = $e^{0.838}$) times more likely when the sets of authors are fully coupled at the national level compared to when there is no *national coupling*. When *bibliographic coupling* is medium or high, however, we find that the effects of geographical coupling are much smaller. In these cases the effect of *regional* and *national coupling* may even be completely absent. For instance, when *bibliographic coupling* is high, the odds ratio for full coupling relative to no coupling is 1.114 (OR = $e^{0.109}$) for *organizational coupling*, 1.074 (OR = $e^{0.071}$) for *regional coupling*, and 0.956 (OR = $e^{-0.045}$) for *national coupling*. None of these are significant. Our results further suggest that, with medium or, especially, high *bibliographic coupling*, citation likelihood is higher when the sets of authors are highly coupled, but not fully coupled, than when the sets of authors are fully coupled. Surprisingly, when *bibliographic coupling* is high, odds of citation when sets of authors are fully coupled at the organizational, regional or national level seem to be equal to the odds of citation in the reference case, when *organizational, regional* or *national coupling* is absent (odds ratio = 1), suggesting that in those cases there is no effect of co-location on citation likelihood.

We ran variations on Model 8 in which we included robustness checks Table 4. The results of these checks can be found in Table 4. We computed *regional coupling* based on a range of 50 km rather than 20 km, and adjusted *national coupling* accordingly (Model (8)). Next, we ran a model that includes coupling measures for which coupling is only relative to the number of authors that have not already been coupled at a more specific level (see above) (Model (9)), and one in which we include our co-location variables without subtracting OC_{ij} from RC_{ij}, and without subtracting OC_{ij} and RC_{ij} from GC_{ij} (Model (10)). Finally, we replaced our *shared authors* variable with one based on a Jaccard similarity of > .8 (Model (11)). The robustness checks do not affect our overall conclusions. While the effect of *organizational coupling* still exceeds that of *national* and *regional coupling*, the alternative coupling measures reduced the effect size of *national coupling* and *regional*, which is to be expected due to the structure of the data. While the coefficients

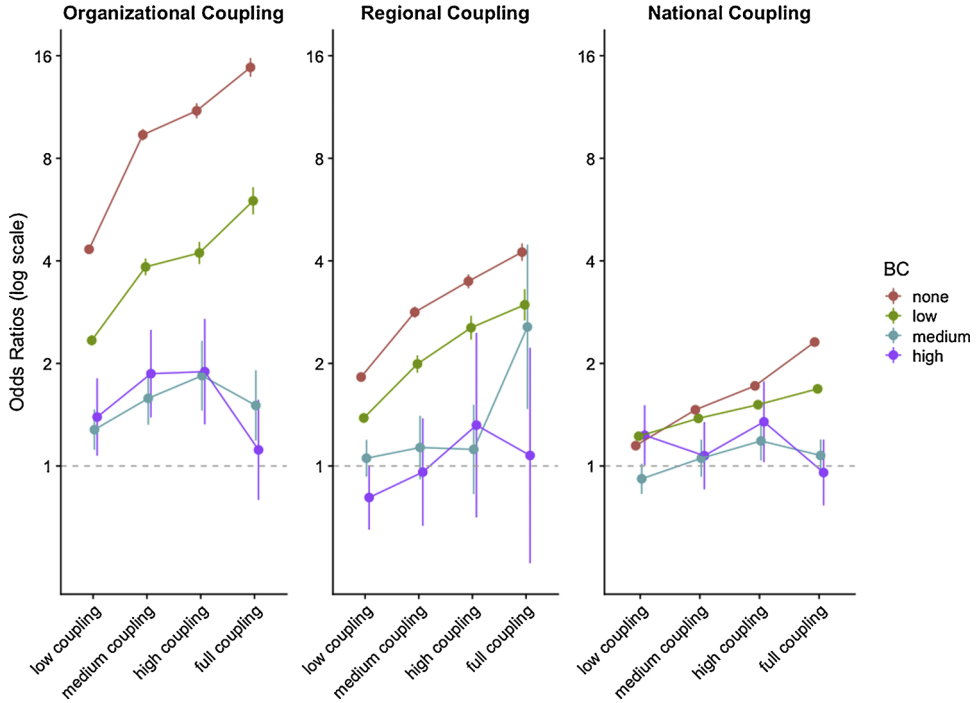

Fig. 1. Effect of geographical variables for different levels of *bibliographic coupling (BC)*. The odds ratios represent the odds of citation for a certain level of *organizational, regional* or *national coupling,* relative to no *organizational, regional or coupling*, and conditional on a level of *bibliographic coupling*.

Table 4
Logistic regression model coefficients: Robustness checks.

	50 km RC (8)	Alternative Coupling 1 (9)	Alternative Coupling 2 (10)	.8 Shared Authors (11)
Intercept	−5.383***	−5.314***	−5.341***	−5.319***
	(0.004)	(0.004)	(0.004)	(0.004)
Bibliographic Coupling (BC)	2.705***	2.622***	2.658***	2.647***
	(0.002)	(0.002)	(0.002)	(0.002)
Organizational Coupling (OC)	5.034***	4.373***	4.998***	6.220***
	(0.016)	(0.016)	(0.016)	(0.015)
Regional Coupling (RC)	2.271***	1.534***	0.022***	2.507***
	(0.014)	(0.014)	(0.000)	(0.013)
National Coupling (NC)	0.966***	0.952***	0.009***	0.886***
	(0.005)	(0.005)	(0.000)	(0.005)
BC*OC	−2.467***	−2.237***	−1.410***	−2.260***
	(0.005)	(0.005)	(0.012)	(0.006)
BC*RC	−1.035***	−0.638***	−0.680***	−0.810***
	(0.011)	(0.008)	(0.011)	(0.011)
BC*NC	−0.347***	−0.333***	−0.343***	−0.289***
	(0.005)	(0.004)	(0.004)	(0.004)
Times Cited	0.006***	0.006***	0.006***	0.006***
	(0.000)	(0.000)	(0.000)	(0.000)
Shared Authors	9.493***	9.402***	9.342***	3.563***
	(0.029)	(0.029)	(0.029)	(0.014)
Authors product	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
References product	−0.0002***	−0.0002***	−0.000***	−0.0002***
	(0.000)	(0.000)	(0.000)	(0.000)
Log Likelihood	−3,411,064	−3,433,076	−3,415,917	−3,464,075
observations	5,591,571	5,591,571	5,591,571	5,591,571
N groups	489,667	489,667	489,667	489,667
Marginal R2	0.83	0.83	0.83	0.83
Conditional R2	0.92	0.92	0.92	0.92

Note: dependent variable: Cit. p < .0001 ***, p < .001 **, p < .01 *.

of *shared authors* computed with similarities of > .8 was slightly lower than in the original model, neither variable affected any of our results. Otherwise, signs, significance and relative importance of all variables remained the same.

5. Discussion

This paper investigated how scientific citations are affected by geographical co-location of scientific teams at the organizational, regional and national level, and tested how these effects vary for different levels of knowledge relatedness. While similar questions have already been addressed within the context of patenting (Breschi and Lissoni, 2009; Jaffe et al., 1993), we believe that we are the first to discuss the geography of citations specifically within the context of scientific citations. Our findings differ from those related to patent citations particularly with respect to strategic citation, which is distinctly different in patenting due to the role of the patent examiner (Alcácer and Gittelman, 2006; Duguet and Macgarvie, 2003; Meyer, 2000).

Similar to findings in the context of patent citations, we found that while there is indeed a geographical bias in scientific citations, this bias is heavily reduced when including knowledge relatedness – a very strong driver of citations. In situations where knowledge relatedness is high, we even find that co- location has no positive effect on citations. Hence, the largest part of the geographical bias in scientific citations is explained by the geographical concentration of related knowledge rather than by geographical biases in citation as such. Thus, our findings shed new and critical light on previous research that found a strong geographical bias in citation, but failed to control for knowledge relatedness.

Furthermore, we also showed that, as expected, the effect of co-location is strongest at the organizational level and weakest at the national level, with intermediate effects at the level of the region. We find that the effect of co-location at the regional and national level is relatively small, when controlling for co-location between teams at the organizational level. These findings are consistent with a previous study that found a distance decay in the probability of citation (Börner et al., 2006). It is also worth noting that our inclusion of three levels of co-location adds to other previous studies that considered only one geographical level (e.g. Matthiessen et al., 2010; Pan et al., 2012).

Further exploration of the effects of geographical co-location under different cognitive circumstances showed that sharing a geographical location particularly increases the chances of citation when knowledge relatedness is low. The observed effect may be due to serendipitous encounters, local buzz (cf. Bathelt et al., 2004; Storper and Venables, 2004), or projects organized by organizations or local agencies, purposefully designed to overcome unrelatedness (Fitjar and Rodríguez-Pose, 2017). It can, however, also be caused by selective citations between scientific teams that are co-located (Baldi, 1998; Gilbert, 1977). As citations have become instrumental in determining the value, quality and potential of scientific articles and, by extension, scientists, the ability to cite other articles may be a strategic asset (Bartneck and Kokkelmans, 2011; Kostoff, 1998). When scientists work on highly related topics, while also working for the same organization or in the same region, they are likely to be direct competitors for grants, jobs and intellectual property granted by organizations and regional governments. This may give scientists an incentive to boost their own citations and to not cite co-located colleagues who work on similar topics. We find evidence of strategic citation in that an article is much more likely to cite another article if the two articles share authors, even when knowledge relatedness is controlled for. However, we did not find evidence of strategic non-citation of co-located colleagues, because when knowledge relatedness is high, fully co-located teams have a similar citation likelihood as teams without co-location. Our findings may motivate future research into strategic motivations for citation decisions (Seeber et al., 2017; Zaggl, 2017) and raise the question whether strategic citations are reciprocated by the receiving authors – a dynamic that is referred to as 'intellectual backscratching' (Baldi, 1998).

Our findings are policy-relevant given debates about concentrating research organizations on campuses and in cities (Grossetti et al., 2014; Matthiessen et al., 2010). Here, we can distinguish between policy implications with regard to citations as acknowledgement of intellectual debt, and citations as strategic tools. Regarding the first, our results suggests that geographical co-location between teams at the organizational, regional and national level particularly enhances citation when knowledge is unrelated. As such, geography can play an important role in the objective to foster excellence in research, especially as breakthroughs come relatively often from interdisciplinary crossovers of previously unrelated domains (Fleming, 2001; Wang et al., 2017). Possibly, this principle can be fruitfully extended towards a cluster of organizations surrounding the university, public research organisation or multinationals as the 'anchor tenant' in a cluster (Agrawal et al., 2009; Feldman, 2003). At the same time, our results also suggest that when authors are co-located the potential for using citations as strategic tools increases. Thus, any policies aimed at detecting and preventing selective citation behaviour could take into account the effect of geography on citations. For example, while our data does not reflect temporary co-location in the form of research visits and conferences, temporary co-location is similar to permanent co-location in that both facilitate serendipitous encounters and the exchange of tacit knowledge (Torre, 2008). On the other hand, temporarily co- located authors are not affected by local competition in the same way as permanently co-located authors are. Stimulating temporary co-location can thus benefit knowledge accumulation without it being hampered by local competition.

6. Limitations

Several limitations of the data and methods used in this study are worth noting. First, while this paper has a focus on understanding the geography of citations, social dynamics between scientists are also likely to play an important role in shaping citations, whereas we have only considered it for as far as social ties are geographically bounded (Head et al., 2018). Nevertheless, we believe that our framework can serve as a template for future research aimed at analysing geographical effects on knowledge accumulation. Second, we do not distinguish between various types of citations. Negative and positive citations are, in this study, considered part of citation networks in the same way. However, since the number of negative citations in scientific articles is limited (Catalini et al., 2015), we do not expect this to have a major impact on our results. Third, the articles studied are all related to life sciences and medicine, which might limit the generalizability of our findings. Engineering and social sciences, for example, are more geographically bounded in their collaboration patterns compared to life sciences (Hoekman et al., 2010). Hence, citation patterns in such fields may be more affected by geographical co-location as well. Moreover, medical research has a relatively strong national bias (Hoekman et al., 2010). The effect of national co-location may thus be smaller for other scientific fields. Fourth, this study is based on a particular time frame: only couples of articles consisting of one article published in 2014 and one published in 2012 were considered. This limits the generalizability of our findings, since the effects considered here may change over time or may vary when time between couples is longer or shorter than two years. Such dynamic analyses are left for future research.

Acknowledgements

The authors thank Gaston Heimeriks, Kyle Siler and the reviewers for their helpful comments and suggestions. M. Wuestman and K. Frenken are financed by the Vici grant (453- 14-014), awarded by the Netherlands Organisation for Scientific Research (NWO). J. Hoekman is financed by a Veni grant (451-15- 037), also awarded by NWO. No potential conflict of interest was reported by the authors.

References

Agrawal, A.K., Cockburn, I.M., Rosell, C., 2009. Not Invented Here? Innovation in Company Towns. NBER Working Paper Series.
Alcácer, J., Gittelman, M., 2006. Patent citations as a measure of knowledge flows: the influence of examiner citations. Rev. Econ. Stat. 88 (4), 774–779. https://doi.org/10.1162/rest.88.4.774.
Azoulay, P., Stuart, T., Wang, Y., 2013. Matthew: effect or fable? Manage. Sci. 60 (1), 92–109.
Balconi, M., Pozzali, A., Viale, R., 2007. The 'codification debate' revisited: a conceptual framework to analyze the role of tacit knowledge in economics. Ind. Corp. Change 16 (5), 823–849. https://doi.org/10.1093/icc/dtm025.
Baldi, S., 1998. Normative versus social constructivist processes in the allocation of citations : a network- analytic model. Am. Sociol. Rev. 63 (6), 829–846.
Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.
Bartneck, C., Kokkelmans, S., 2011. Detecting h-index manipulation through self-citation analysis. Scientometrics 87 (1), 85–98. https://doi.org/10.18637/jss.v067.i01.
Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48.
Bathelt, H., Malmberg, A., Maskell, P., 2004. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Prog. Hum. Geogr. 28 (1), 31–56. https://doi.org/10.1191/0309132504ph469oa.
Biscaro, C., Giupponi, C., 2014. Co-authorship and bibliographic coupling network effects on citations. PLoS One 9 (6). https://doi.org/10.1371/journal.pone.0099502.
Börner, K., Penumarthy, S., Meiss, M., Ke, W., 2006. Mapping the diffusion of information among major U.S. research institutions. Scientometrics 68 (3), 415–426. https://doi.org/10.1007/s11192-006-0120- 2.
Boschma, R., 2005. Proximity and innovation: a critical assessment. Reg. Stud. 39 (1), 61–74. https://doi.org/10.1080/0034340052000320887.
Boschma, R., Heimeriks, G., Balland, P.A., 2014. Scientific knowledge dynamics and relatedness in biotech cities. Res. Policy 43 (1), 107–114. https://doi.org/10.1016/j.respol.2013.07.009.
Breschi, S., Lissoni, F., 2009. Mobility of skilled workers and co- invention networks: an anatomy of localized knowledge flows. J. Econ. Geogr. 9 (4), 439–468. https://doi.org/10.1093/jeg/lbp008.
Breschi, S., Lissoni, F., Malerba, F., 2003. Knowledge-relatedness in firm technological diversification. Res. Policy 32 (1), 69–87.
Catalini, C., Lacetera, N., Oettl, A., 2015. The incidence and role of negative citations in science. Proc. Natl. Acad. Sci. 112 (45), 13823–13826. https://doi.org/10.1073/pnas.1502280112.
Collins, H., 1985. Changing order: replication and induction in scientific practice. Philos. Soc. Sci. Vol. 19. https://doi.org/10.1177/004839318901900112.
Collins, H., 2001. Tacit knowledge, trust, and the Q of sapphire. Soc. Stud. Sci. 31, 71–85.
Cowan, R., David, P.A., Foray, D., 2000. The explicit economics of knowledge codification and tacitness. Ind. Corp. Change 9 (2), 211–253.
Dasgupta, P., David, P.A., 1994. Toward a new economics of science. Res. Policy 23 (5), 487–521. https://doi.org/10.1016/0048-7333(94)01002- 1.
Duguet, E., Macgarvie, M., 2003. How well do patent citations measure flows of technology? Evid. French Innov. Surv. (September), 1–30.
Faraway, J.J., 2016. Extending the Linear Model With R, 2nd ed. Boca Raton, FL: Taylor & Francis.
Feldman, M., 2003. The locational dynamics of the US biotech UIndustry: knowledge externalities and the anchor hypothesis. Ind. Innov. 10 (3), 311–328. https://doi.org/10.1080/1366271032000141661.
Fitjar, R.D., Rodríguez-Pose, A., 2017. Nothing is in the air. Growth Change 48 (1), 22–39. https://doi.org/10.1111/grow.12161.
Fleming, L., 2001. Recombinant uncertainty in technological search. Manage. Sci. 47 (1), 117–132. https://doi.org/10.1287/mnsc.47.1.117.10671.
Frenken, K., 2010. Geography of scientific knowledge: a proximity approach. Eindhoven Centre for Innovation Studies (ECIS) Working Paper 10- 01, Eindhoven: Technische Universiteit Eindhoven.
Furman, J.L., Kyle, M.K., Cockburn, I., Henderson, R., 2006. Public & Private Spillovers, Location and the Productivity of Pharmaceutical Research. NBER Working Paper Series.
Gilbert, G.N., 1977. Referencing as persuasion. Soc. Stud. Sci. 7, 113–122.
Glanzel, W., Czerwon, H.J., 1996. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. Scientometrics 37 (2), 195–221. https://doi.org/10.1007/BF02093621.
Grossetti, M., Eckert, D., Gingras, Y., Jégou, L., Larivière, V., Milard, B., 2014. Cities and the geographical deconcentration of scientific activity: a multilevel analysis of publications (1987–2007). Urban Stud. 51 (10), 2219–2234. https://doi.org/10.1177/0042098013506047.
Hagstrom, W.O., 1965. The Scientific Community. Basic Books, New York.
Head, K., Li, Y.A., Minondo, A., 2018. Geography, Ties, and Knowledge Flows: Evidence From Citations in Mathematics. CEPR Discussion Paper No. DP12942. Available at. SSRNhttps://doi.org/10.2139/ssrn.3163679. Https://Ssrn.Com/Abstract = 318240 5.
Hoekman, J., Frenken, K., Tijssen, R.J.W., 2010. Research collaboration at a distance: changing spatial patterns of scientific collaboration within Europe. Res. Policy 39 (5), 662–673. https://doi.org/10.1016/j.respol.2010.01.012.
Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 108 (3), 577–598.
Jarneving, B., 2007. Bibliographic coupling and its application to research-front and other core documents. J. Informetr. 1 (4), 287–307. https://doi.org/10.1016/j.joi.2007.07.004.
Kaplan, N., 1964. The norms of citation behaviour: prolegomena to the footnote. J. Assoc. Inf. Sci. Technol. 16 (3), 179–184.
Kessler, M.M., 1960. An Experimental Communication Center for Scientific and Technical Information. Lincoln Laboratory: Massachusetts Institute for Technology.
Kessler, M.M., 1962. An Experimental Study of Bibliographic Coupling Between Technical Papers. Lincoln Laboratory: Massachusetts Institute for Technology.
Kostoff, R.N., 1998. The use and misuse of citation analysis in research evaluation. Scientometrics 43 (1), 27–43.
Kovalchik, S., 2016. RISmed: Download Content From NCBI Databases. R Package Version 2.1.6. Retrieved from. https://cran.r- project.org/package = RISmed.
Livingstone, D.M., 2003. Putting Science in Its Place: Geographies of Scientific Knowledge. University of Chicago Press, Chicago.
Matthiessen, C.W., Schwarz, A.W., Find, S., 2002. The top-level global research system, 1997-99: centres, networks and nodality. An analysis based on bibliometric indicators. Urban Stud. 39 (5–6), 903–927. https://doi.org/10.1080/0042098022012837.
Matthiessen, C.W., Schwarz, A.W., Find, S., 2010. World cities of scientific knowledge: systems, networks and potential dynamics. An analysis based on bibliometric indicators. Urban Stud. 47 (9), 1879–1897. https://doi.org/10.1177/0042098010372683.
Merton, R.K., 1942. The ethos of science. In: Sztomka, P. (Ed.), Social Structure and Science, pp. 115–126.
Merton, R.K., 1968. The matthew effect in science. Science 159 (3810), 56–63. Retrieved from. http://www.jstor.org/stable/172341 4.
Meyer, M., 2000. What is special about patent citations? Differences between scientific and patent citations. Scientometrics 49 (1), 93–123.
NaturalEarth, 2017. Natural Earth. Retrieved 1 January 2017, from. http://www.naturalearthdata.com/a bout/.
Nightingale, P., 2004. Technological capabilities, invisible infrastructure and the un-social construction of predictability: the overlooked fixed costs of useful research. Res. Policy 33 (9), 1259–1284.
Pan, R.K., Kaski, K., Fortunato, S., 2012. World citation and collaboration networks: uncovering the role of geography in science. Sci. Rep. 2 (902). https://doi.org/10.1038/srep00902.
Peters, H.P.F., Braam, R.R., van Raan, A.F.J., 1995. Cognitive resemblance and citation

relations in chemical engineering publications. J. Am. Soc. Inf. Sci. 46 (1), 9–21. https://doi.org/10.1002/(SICI)1097-4571(199501) 46:1<9::AID-ASI2>3.0.CO;2-3.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. Retrieved from. Austria: R foundation for Statistical Computing, Vienna. https://www.r-project.org/.
Seeber, M., Cattaneo, M., Meoli, M., Malighetti, P., 2017. Self-citations as strategic response to the use of metrics for career decisions. Res. Policy 1–14. https://doi.org/10.1016/j.respol.2017.12.004. December 2016.
Singh, J., 2005. Collaborative networks as determinants of knowledge diffusion patterns. Manage. Sci. 51 (5), 756–770. https://doi.org/10.1287/mnsc.1040.0349.
Small, H.G., 1978. Cited documents as concept symbols. Soc. Stud. Sci. 8 (3), 327–340. https://doi.org/10.1177/030631277800800305.
Storper, M., Venables, A.J., 2004. Buzz: face-to-face contact and the urban economy. J. Econ. Geogr. 4 (4), 351–370. https://doi.org/10.1093/jnlecg/lbh027.
Thijs, B., Zhang, L., Glanzel, W., 2015. Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes. Scientometrics 105 (3), 1453–1467. https://doi.org/10.1007/s11192-015-1641- 3.
Torre, A., 2008. On the Role Played by Temporary Geographical Proximity in Knowledge Transmission On the Role Played by Temporary Geographical Proximity in Knowledge Transmission. pp. 3404. https://doi.org/10.1080/00343400801922814.
Vanclay, J.K., 2013. Factors affecting citation rates in environmental science. J. Informetr. 7 (2), 265–271. https://doi.org/10.1016/j.joi.2012.11.009.
Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: a cautionary tale for users of bibliometric indicators. Res. Policy 46 (8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006.
Zaggl, M.A., 2017. Manipulation of explicit reputation in innovation and knowledge exchange communities: the example of referencing in science. Res. Policy 46 (5), 970–983. https://doi.org/10.1016/j.respol.2017.02.009.