Globalization of science: Geographical distance measurements of research collaboration

Nees Jan van Eck, Ludo Waltman and Robert J.W. Tijssen

{ecknjpvan, waltmanlr, tijssen}@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands)

Abstract
We study the globalization of science by measuring geographical distances between collaborating researchers. Our analysis covers all publications in the period 1980–2009 indexed in the Web of Science database. It turns out that during the last three decades collaboration distances have increased more or less linearly over time. For instance, the share of publications involving collaboration over a distance of more than 1000 km displays a linear growth from about 6% in 1980 to about 23% in 2009. We also make comparisons of globalization trends in different fields of science and in different countries.

Introduction
Research collaboration has become an important contributing factor for productive and successful scientific research. At the same time, the physical distance between partners has become increasingly irrelevant in contemporary science, which is driven more and more by improved ICT facilities (internet), common language (English) and convergence of shared themes and urgent problems of global relevance ('grand challenges' research agenda). Obviously, research collaboration is driven by a myriad of determinants, part of which are distance-dependent. Hence, we distinguish between collaboration processes across the geographical scale: 'regionalization' (within the same sub-national region), 'nationalization' (within nation states) and 'globalization' (cross national). These processes and their driving forces are difficult, if not impossible, to disentangle analytically within an increasingly interconnected worldwide network of research partnerships. However, a macro-level analysis of collaboration patterns enables an examination of general ('structural') characteristics at the aggregate level of countries and fields of science.
The OECD handbook on measuring globalization (OECD, 2005) identifies a range of gaps as regards to measuring globalization processes within scientific research and technological development. The modes for capturing empirical data on international research collaboration are indeed numerous: participation in international research organizations; coordination and joint programming of research activities; mobility migration flows of R&D personnel; (electronic) communication between research partners; shared physical resources and facilities; allocation of national R&D budgets; cross-border contracts and flows of funding; tangible outputs of physical or virtual collaboration. In all but one case, comparative global data are lacking for a truly comprehensive analysis across and within institutional and geographical borders. The exception is: joint research publications co-authored by collaborating researchers.
This paper addresses the analytical potential of this source within the context of an indicator-based framework. The methods that we introduce allow us to produce, for the first time ever, unobtrusive distance-based measurements of globalization processes within and across national borders for all countries worldwide. Our analysis is guided by the following research questions: What was the globalization rate in recent decades? Which countries are leading the process of globalization at present, and which ones are lagging behind? And are there different trends across fields of science?

Methodology and research design

The march of globalization through the landscape of science is documented in the addresses provided by researchers in their publications in the open scientific literature. We used the millions of publications indexed in the CWTS version of the Web of Science (WoS) database, produced by Thomson Reuters, to generate statistics derived from bibliographic information. Our empirical analysis can be nested within the research program of spatial scientometrics (Frenken et al., 2009) and builds on a rapidly expanding body of scientometric studies in which internationalization and globalization processes are examined (e.g., Narin et al., 1991; Luukkonen et al., 1993; Katz, 1994; Glänzel, 2001).

We selected all publications in the WoS database that were published between 1980 and 2009, that are of the document type 'article' or 'review' and that have at least one address. There are 21.4 million publications that satisfy these three criteria. For each of the selected publications, an attempt was made to find the geographical coordinates (i.e., the latitude and the longitude) of the addresses mentioned in the publication's address list.[18] Finding the geographical coordinates of an address is referred to as geocoding.

We employed the following geocoding procedure (cf Leydesdorff & Persson, 2010). First, all 39.0 million addresses of the selected publications were reduced to a city and a country.[19] Other address elements, such as organization names, streets and postal codes, were disregarded. Next, for each unique address, the number of times it occurs in the address lists of the selected publications was counted. Performing geocoding for all unique addresses turned out to be infeasible, and we therefore restricted our attention to about 11 000 addresses that occur most frequently. The remaining addresses were not taken into account in the geocoding procedure, and their coordinates were considered unknown. For the selected addresses, coordinates were obtained using the website www.gpsvisualizer.com/geocoder/. This website relies on geocoding services provided by Google and Yahoo. Comparing the two services, we found that they sometimes yield quite different results and also that they sometimes fail to recognize an address. Furthermore, although both services make errors, Google seemed to be somewhat more accurate than Yahoo. Based on these observations, we decided to take the following approach. For each address, the Google-Yahoo distance was calculated, that is, the distance between the coordinates provided by Google and the coordinates provided by Yahoo. An address was verified manually if the Google-Yahoo distance is larger than 50 km and the address occurs more than 200 times in the address lists of the selected publications. In some cases, the verification of an address revealed that both the coordinates of Google and the coordinates of Yahoo were incorrect. Usually, the correct coordinates could then be found manually, but in a small number of cases the correct coordinates remained unknown. An address was also verified manually if the Google-Yahoo distance is larger than 100 km and the address occurs less than 200 times. In these cases, however, the verification of an address was done in a more cursory way. If the correctness of the coordinates of either Google or Yahoo could not be easily established, the coordinates of

[18] The WoS database distinguishes between the ordinary addresses associated with the author(s) of a publication and the so-called reprint address of a publication. We disregarded the reprint addresses of all publications that appeared after 1997. Starting from 1998, the reprint address of a publication is usually also mentioned in the ordinary address list of the publication. When the reprint address is not mentioned in the ordinary address list, it seems that in most cases the corresponding author of the publication moved to a new organization after the research reported in the publication was finished.

[19] The distinction between cities and provinces is not always clearly indicated in an address. What we refer to as cities may therefore sometimes be provinces. In the case of US and Canadian addresses, we took into account not only the city and the country indicated in an address but also the state or the province. State or province information seems to be provided consistently in a large majority of the US and Canadian publications.

an address were simply considered unknown. Addresses that did not satisfy one of the above two criteria (about 90% of the selected addresses) were not verified manually. For these addresses, the coordinates provided by Google were taken as the correct ones. In the end, our geocoding procedure yielded coordinates for 98.6% of the 39.0 million addresses of the selected publications.

To assess the accuracy of our geocoding procedure, we manually verified the coordinates of a limited number of addresses. Out of the 11 000 addresses that were taken into consideration in the geocoding procedure, a random sample of 150 addresses was taken. For each of the 150 addresses, we manually identified the geographical coordinates. We then compared the manually identified coordinates with the coordinates obtained using the geocoding procedure. There turned out to be four addresses for which the distance between the manually identified coordinates and the geocoding coordinates was larger than 50 km. In three of the four cases, this was caused by the presence of multiple cities with the same name in a country. In the fourth case, this was caused by an error in the WoS database. The four addresses with incorrect geocoding coordinates are all relatively unimportant. Together, the addresses occur in 343 publications.

Using the results of our geocoding procedure, we calculated the *geographical collaboration distance* (GCD) of each selected publication. We define the GCD of a publication as the largest geographical distance between two addresses mentioned in the publication's address list.[20] If a publication's address list contains only one address, the GCD of the publication is defined as zero. As mentioned earlier, publications that do not have any address at all were not taken into consideration in our analysis. Due to the limitations of the geocoding procedure, the coordinates of some of the addresses of a publication may be unknown. This turned out to be the case for 2.3% of the selected publications. For these publications, the addresses with unknown coordinates were disregarded and the GCD was calculated based on the remaining addresses. Notice that this may cause the GCD of these publications to be biased downwards.

Based on the GCD of a publication, we define the following four indicators of scientific globalization:

- *Mean geographical collaboration distance* (MGCD): average GCD of a set of publications;
- *Percentage of medium and long distance collaborations* (%MLDC): percentage of publications with a GCD of more than 200 km;
- *Percentage of long distance collaborations* (%LDC): percentage of publications with a GCD of more than 1000 km;
- *Percentage of very long distance collaborations* (%VLDC): percentage of publications with a GCD of more than 5000 km.

These indicators can be calculated for any set of publications as defined according to some delineation criterion, either geographical (e.g., country, region or city), institutional (e.g., university, research institute or company) or cognitive (e.g., field of science or research topic). When counting publications and calculating our indicators, we take a fractional counting approach. For instance, a publication with addresses from two countries is treated as belonging half to each country.

[20] Alternatively, we could have defined the GCD of a publication as the average geographical distance between all pairs of addresses mentioned in the publication's address list. A drawback of this definition would have been that the GCD of a publication may depend heavily on various details of the way in which address data are processed. For instance, if a publication has two or more addresses in the same city (perhaps even belonging to the same organization), are these addresses treated as one single address or as multiple different addresses? Because of issues such as these, we prefer to define the GCD of a publication as the largest geographical distance between two addresses mentioned in the publication's address list.

Our analysis covers mainstream science as a whole, that is, across all publications indexed in the WoS database. A further breakdown is made into four broad fields of science, namely *Engineering Sciences and Technology* (ET), *Medical Sciences, Life Sciences and Agricultural Sciences* (MLA), *Natural Sciences, Computer Sciences and Mathematics* (NCM), and *Social Sciences, Humanities and Arts* (SHA). These fields were obtained by grouping WoS journal subject categories. Each subject category was assigned to one of the four broad fields.

Selected results and observations

Science has globalized at a fairly steady rate. The average collaboration distance (MGCD) has increased more or less linearly over the past three decades, from about 300 km in 1980 to more than 1500 km in 2009 (see Figure 1, left panel).[21] This indicates a fundamental phase-shift from a nation-state science based on predominantly nation-oriented partnerships to an internationally networked global science system. The process of increased interconnectedness seems to have occurred in different speeds at different geographical scales (see Figure 1, right panel). The share of medium and long distance collaborations (%MLDC) has grown by more than a factor three between 1980 and 2009, and the share of long distance collaborations (%LDC) has grown by almost a factor four. The fraction of very long distance partnerships (%VLDC) has increased almost fivefold. Hence, collaboration occurs more and more across larger distances. This clearly shows the process of globalization of science.

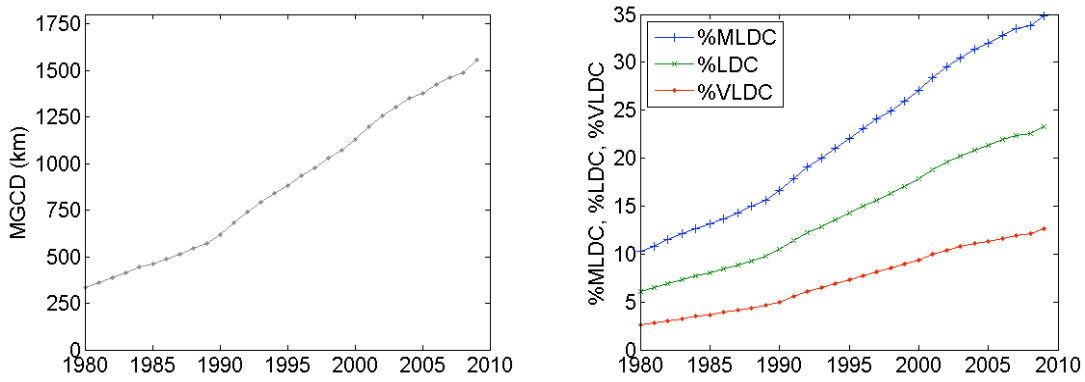


Figure 1. Development of MGCD (left panel) and %MLDC, %LDC and %VLDC (right panel) over time for all WoS publications.

The growth and evolution of world science is not only driven by socio-economic and political factors, but also by the cognitive dynamics within scientific fields, such as the rise of the biomedical sciences, nanoscience and ICT. Figure 2 captures the differences in globalization dynamics for four broad scientific fields. The field of *Natural Sciences, Computer Sciences and Mathematics* (NCM) was, and still is, the most globalized of the four. This is at least partly the result of a long tradition of cross-border, resource-intensive 'big science' collaboration (physics and astronomy), in which large research facilities are shared by scientists spread across the globe. The field of *Medical Sciences, Life Sciences and Agricultural Sciences* (MLA), however, with only two-third of NCM's MGCD level in 1980, has almost caught up with NCM's level of globalization in 2009. MLA has also become global 'big science' in terms of research partnership proximities. The field of *Engineering*

[21] As can be seen in Figure 1, there is a sudden increase in the growth of the MGCD around 1990. We suspect this sudden increase to be a database artefact rather than a true effect. Furthermore, for the purpose of comparison, it may be interesting to know that between 1980 and 2009 the share of internationally co-authored publications has jumped from 5% to 21% and the average number of authors per publication has risen from 2.5 to 4.5.

Sciences and Technology (ET) was engaged in the same catching-up process, but apparently its research network expansion has not been able to keep up with MLA's steep growth rate since 2003. In contrast, recent years have shown a remarkable catching up of the field of *Social Sciences, Humanities and Arts* (SHA), the field least prone to collaboration between individual researchers and scholars. SHA is still significantly behind the others but is closing in fast.

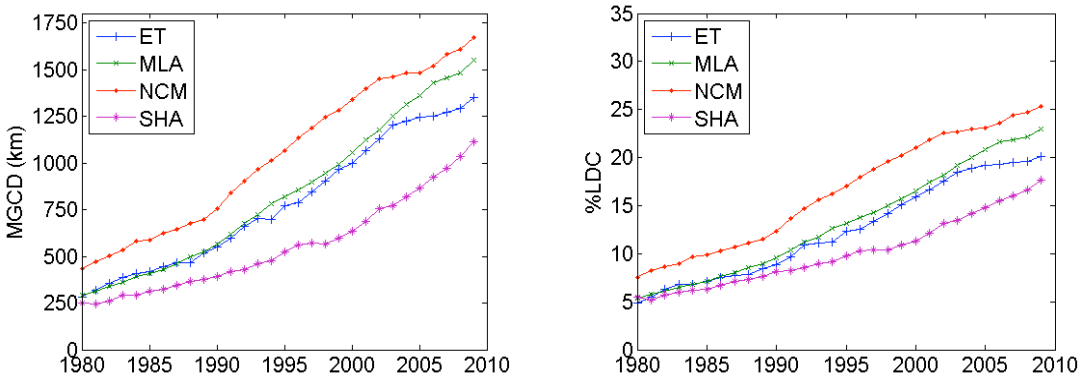
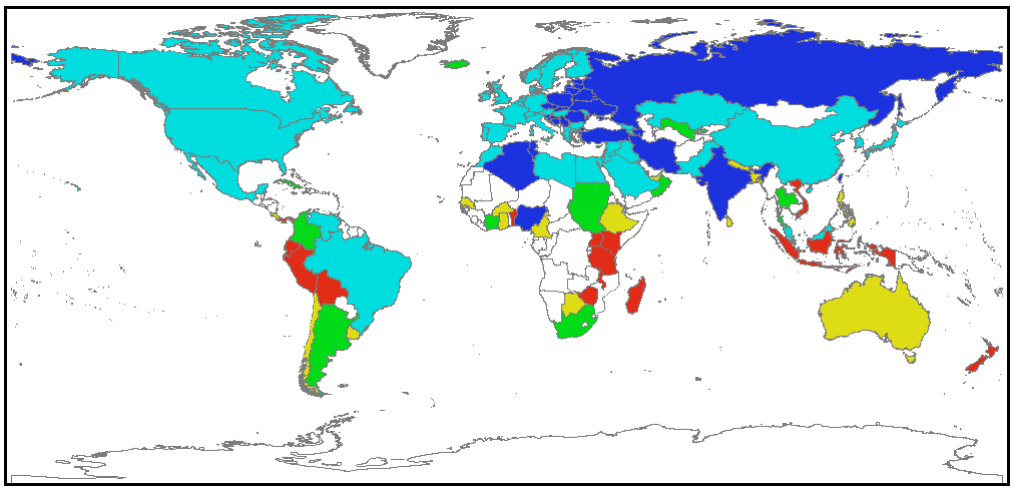


Figure 2. Development of MGCD (left panel) and %LDC (right panel) over time for publications in four broad scientific fields.

Location matters in science. The distances between research partners are obviously also influenced by the geographical location of research. Those located within the center of science-intensive countries, regions or continents have less need for long distance partners than those working at the geographical periphery of world science. The effect of a country's location on the globe is aptly illustrated in Figure 3, which displays 113 color-coded countries according to their MGCD level in the period 2007–2009. Each of the 113 countries has an output of at least 200 WoS publications in this period. As expected, 'peripheral' countries in the southern hemisphere are characterized by the largest collaboration distances, with New Zealand as an extreme case with an MGCD level of more than 4000 km. Perhaps more surprising is that several countries in or near the tropics also surpass the 4000 km mark. These are typically developing countries with long distance partners on either the northern or the southern hemisphere.



Color coding: Dark blue: MGCD < 1000 km; Light blue: MGCD between 1000 and 2000 km; Green: MGCD between 2000 and 3000 km; Yellow: MGCD between 3000 and 4000 km; Red: MGCD > 4000 km.

Figure 3. World map with colors indicating countries' MGCD level in the period 2007–2009.

More detailed statistics for some selected countries are reported in Table 1. In the selection of the countries, only countries with an output of more than 3000 WoS publications in 2009 were considered. Perhaps the most notable observation is that the 'catching up' countries, which are experiencing a rapid growth of their publication output, tend to have a very small or even negative growth of their MGCD level. Apparently, these countries achieve their rapid output growth mainly by means of publications that do not involve long distance collaboration. It seems likely that long distance collaboration will increase at a later 'global networking' stage in the development of the science systems of these countries.

Table 1. Publication output and MGCD statistics for some selected countries.

Country	Category	2009		Annual growth rate 2000–2009	
		Output	MGCD	Output	MGCD
USA	Top 5 output	271 383	1 883	1.5%	4.7%
China	Top 5 output	108 202	1 302	17.1%	1.1%
Japan	Top 5 output	64 362	1 152	-0.4%	3.4%
United Kingdom	Top 5 output	63 355	1 681	0.5%	6.7%
Germany	Top 5 output	61 290	1 360	1.3%	4.6%
New Zealand	Top 5 MGCD	4 515	4 154	2.9%	5.1%
Australia	Top 5 MGCD	27 298	3 604	4.7%	4.4%
Chile	Top 5 MGCD	3 180	3 128	9.6%	1.0%
South Africa	Top 5 MGCD	5 098	2 898	6.2%	3.9%
Singapore	Top 5 MGCD	5 832	2 828	7.2%	6.9%
Iran	Top 5 output growth	12 547	806	30.4%	-2.6%
Malaysia	Top 5 output growth	3 344	1 541	19.9%	-2.2%
China	Top 5 output growth	108 202	1 302	17.1%	1.1%
Turkey	Top 5 output growth	19 340	542	16.8%	-1.8%
Thailand	Top 5 output growth	3 450	2 674	16.4%	-1.2%
Ireland	Top 5 MGCD growth	3 969	1 459	7.4%	7.2%
Singapore	Top 5 MGCD growth	5 832	2 828	7.2%	6.9%
United Kingdom	Top 5 MGCD growth	63 355	1 681	0.5%	6.7%
Norway	Top 5 MGCD growth	5 876	1 522	5.3%	5.5%
New Zealand	Top 5 MGCD growth	4 515	4 154	2.9%	5.1%
World		1 134 979	1 553	4.1%	3.6%

Acknowledgement

We gratefully acknowledge the research assistance provided by Suze van der Luijt (CWTS).

References

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: towards a cumulative research program. *Journal of Informetrics*, *3*(3), 22–232.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69–115.

Narin, F., Stevens, K., & Whitlow, E.S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics*, *21*(3), 313–323.

Katz, J.S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31–43.

Leydesdorff, L., & Persson, O. (2010). Mapping the geography of science: distribution patterns and networks of relations among cities and institutes. *Journal of the American Society for Information Science and Technology*, *61*(8), 1622–1634.

Luukkonen, T., Tijssen, R.J.W., Persson, O., & Sivertsen G. (1993). The measurement of international scientific collaboration. *Scientometrics*, *28*(1), 15–36.

OECD (2005). *OECD handbook on economic globalisation indicators*. Paris: Organisation for Economic Cooperation and Development.

Wildavsky, B. (2010). *The great brain race: how global universities are reshaping the world*. Princeton and Oxford: Princeton University Press.