WORKING PAPER
Wageningen School of Social Sciences

What is the effect of spatial proximity on research collaboration in a small country?
A gravity model for co-authored publications

Pieter W. Heringa, Edwin Horlings, Wim J. M. Heijman

WASS Working PAPER No. 6
2013

Wageningen School
of Social Sciences

Hollandseweg 1, 6706 KN Wageningen,
The Netherlands
Phone: +31 317 48 41 26
Fax: +31 317 48 47 63
Internet: http://www.wass.wur.nl
e-mail: wass@wur.nl

Working Papers are interim reports on work of Wageningen School of Social Sciences (WASS) and have received only limited reviews[1]. Each paper is refereed by one member of the Editorial Board and one member outside the board. Views or opinions expressed in them do not necessarily represent those of WASS.

WASS's researchers are based in two departments: 'Social Sciences' and 'Environmental Sciences' and two institutes: 'LEI, Agricultural Economics Research Institute' and 'Alterra, Research Institute for the Green World'. In total WASS comprises about 250 researchers.

WASS promotes top-quality research that increases our understanding of social processes and design practices around challenges and opportunities in the sphere of agriculture, food, health, environment and development. WASS provides a home for internationally oriented scholars with diverse disciplinary expertise and research traditions, and wishes to create an enabling environment for disciplinary, interdisciplinary and trans-disciplinary work. WASS strives to make high-quality academic contributions and also to critically engage in societal debates and contribute to societal problem solving and innovation. WASS offers in-depth PhD training that provides students with a suitable background for a career in academic and applied research, policy-making, or other leading societal positions.
Comments on the Working Papers are welcome and should be addressed directly to the author(s).

Pieter Heringa	Rathenau Institute, p.heringa@rathenau.nl
Edwin Horlings	Rathenau Institute, e.horlings@rathenau.nl
Wim Heijman	Wageningen University, wim.heijman@wur.nl

[1] Working papers may have been submitted to other journals and have entered a journal's review process. Should the journal decide to publish the article the paper no longer will have the status of a WASS Working Paper and will be withdrawn from the WASS website. From then on a link will be made to the journal in question referring to the published work and its proper citation.

What is the effect of spatial proximity on research collaboration in a small country? A gravity model for co-authored publications

Pieter W. Heringa[1,2,3](Corresponding author), Edwin Horlings[1], Wim J. M. Heijman[4]

[1)] *Science System Assessment Department, Rathenau Institute, PO box 95366, 2509 CJ The Hague, the Netherlands;* [2)] *KWR Watercycle Research Institute, Nieuwegein, the Netherlands;* [3)] *Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands;* [4)] *Agricultural Economics and Rural Policy Group, Social Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands*

T: +31 70 342 1542

F: +31 70 363 3488

p.heringa@rathenau.nl; e.horlings@rathenau.nl; wim.heijman@wur.nl

Abstract

In this study we fit a gravity model for knowledge production, where the number of co-authored publications is explained by the size of the affiliated organisations and the physical distance between them. We analyse 2247 publications on drinking water and wastewater in the period 2006-2008 with at least one author affiliation in the Netherlands. At this small spatial level we find a robust and significant effect in the expected direction: the larger the distance between two organisations, the less publications they co-author

together. We extend the model and show that organisations of the same type (academia, governmental bodies, etc.) also collaborate more.

Key words

Research collaboration; geographical proximity; gravity model; Netherlands; co-authorships; negative binomial distribution

JEL codes

O 31; R12; R15;

1. INTRODUCTION

Since about five decades, there is an increasing interest for the phenomenon of collaboration in scientific research. Early work by e.g. Smith (1958) and De Solla Price and Beaver (1966) was followed by many others until today, trying to measure collaboration, discern underlying patterns and find drivers of collaboration. Meanwhile, it has become commonplace to encourage collaboration among researchers and institutions in various ways (Katz and Martin, 1997). However, knowledge and knowledge production are geographically clustered (Malecki, 2010). It is known that people collaborate more with geographically proximate partners both at the micro level of one building (Allen, 1977) and at the macro level of very large countries like the USA or entire continents (Katz, 1994; Hoekman et al., 2010). Little is known so far about a level in between: that of small countries. In this paper we look at a small country, the Netherlands. Moreover, we use data from a sector that tends to organise itself regionally: the water sector.

Policies that encourage research collaboration pay attention to the role of geographical distance in two ways. On the one hand some actively encourage the co-location of researchers (and sometimes other stakeholders) for example in science parks to promote knowledge exchange and spill-overs, and to share large facilities. On the other hand the EU for example actively promotes collaboration across long geographical distances, with the idea that a longer radius improves the chance of finding relevant collaborators for shared knowledge production. More insight in the role of geographical proximity may improve such investments and policies.

The remainder of this article is organised as follows: in section two we give an overview of the earlier literature on drivers of collaboration, with specific attention to the role of geographical proximity. In section three we describe the methodology we employed to construct a dataset on collaboration and the statistical methods to test the effect of distance. In section four we present our findings and results. In section five give our conclusions and discuss the implications for future research.

2. CURRENT LITERATURE

There exists quite some literature already on patterns in collaborative research. There are three streams within this literature on which we specifically build: the use of co-authorships as a proxy for collaboration, factors that are known to drive collaboration, and the role of geographical distance in collaboration. We give an overview of each of them below.

Mapping Co-authorships

The current policies that promote collaboration and the research into the factors that induce collaboration implicitly assume that research collaboration is well understood, that it can be measured and that more collaboration is always better, either for knowledge development or for the effective exploitation of knowledge (Katz and Martin, 1997). However, what research collaboration actually entails and how it can be made operational is less obvious than it may seem at first sight. Still, remarkably little is done so far to conceptualise the idea somewhat further. The everyday use of the word collaboration suggests that research collaboration involves the working together of researchers to achieve a common goal (most likely conducting research and developing new scientific knowledge). That does not reveal how closely researchers should work together to consider it collaboration. It is hence not easy for an external person to assess who should be counted as collaborators (Katz and Martin, 1997).

An alternative for assessing the contributions of collaborators is to use the names mentioned as co-authors on scientific papers that publish the results of research. We consider the use of co-authorships as the most viable way of collecting data on collaboration for our analysis. It is virtually impossible to obtain an equally large and reliable dataset using any other approximation of collaboration. As Melin and Persson

(1996) already stated, we simply have to accept some uncertainty, and we can hope that significant research collaboration will generally lead to co-authoring, as authors will want to claim priority.[1]

Factors Influencing Collaboration

Before we try to discern statistical patterns, let us first look at the motivations authors can have for collaboration with others. Beaver (2001) lists no fewer than 18 purposes for which people collaborate. We have aggregated them to a few main categories:

- Access to resources (expertise, equipment, funds)
- Efficiency/effectiveness (more (rapid) progress, tackle bigger problems, find flaws)
- Learning purposes (obtain/share (tacit) knowledge, educate others, advance knowledge)
- Personal purposes (fun, satisfy curiosity, reduce isolation, build a network, increase scientific recognition and visibility or strengthen the own career).

There is also an extensive literature on factors that stimulate or encourage collaboration. We distinguish two main categories:

- Factors that have to do with developments in science: funding patterns, rationalisation of scientific manpower, demands for instrumentation, increasing specialisation and professionalisation of science, etc.
- Conditions that facilitate interaction and collaboration: spatial proximity, infrastructural developments, absence of social, cultural, linguistic or political barriers, etc. (Acedo et al., 2006; Katz, 1994; Katz and Martin, 1997).

The Impact of Geographical Distance

Although we are aware that the collaboration pattern is the result of a complex interplay of at least all the factors mentioned above, in this article we single out one specific factor, namely spatial (or geographical) distance[1]. Distance is a known determinant of collaboration patterns. Scientific collaboration becomes increasingly interinstitutional and international. There are at least two reasons why more insight in the role of physical distance in collaborative knowledge production is important: First, as we have argued above, collaboration has many benefits. Some of these benefits might be larger for collaboration across larger distances; if the search radius increases, the odds become higher that one finds relevant partners with a supplementary knowledge base, leading to new knowledge. Second: significant investments have been made in the past years on the one hand to stimulate long-distance collaboration (such as the so-called Framework Programmes of the European Union, intended to promote collaboration across member states), while on the other hand large investments are made to co-locate researchers closely together to stimulate collaboration (the idea behind many science parks for example). More insight in the relevance of distance can improve the rationale behind such investments (Hoekman et al., 2010).

Intuitively, one would argue that beneficial geographical factors would lead to more collaboration. Most collaborations are initiated in informal settings, and geographical proximity facilitates such settings, for example because face-to-face meetings are easier to organize (e.g. Katz, 1994).

However, too much geographical proximity may hinder the processes of knowledge production; if actors in a specific region become too much inward-looking. This may result in geographical lock-in, weakening the learning capacities of the actors. This is a risk in particular if the actors are also similar in other dimensions (e.g. cognitively), leading to a small knowledge base (Boschma, 2005).

[1] We define geographical distance as a broader category than spatial or physical distance alone. Some authors have studied the effect of national or regional borders for example; this is a form of geographical distance but not of spatial/physical distance.

Various researchers have conducted empirical studies on the role of geographical proximity, with different results. Hagstrom (1965) was the first to suggest that geographical distance matters for collaborative work. He stated that, although face-to-face contact often is not necessary for collaboration, researchers often agree to collaborate after informal communication, and this informal communication is greatly promoted by spatial propinquity. Allen (1977) showed on a microlevel (within buildings) that the probability of communicating between potential collaborators declines sharply as distance increases. Kraut and Egido (1988) show similar findings for researchers that collaborate already. One might think that these findings are simply an artefact, as researchers who share important characteristics (such as research interests) will often be located close to each other. However, Kraut and Egido controlled for organisational proximity and research similarity, and still found an independent effect for spatial proximity.

In the early nineties, two studies show patterns of coauthorship among European countries. Narin et al. (1991) do not include an indicator for physical distance, they find that the patterns are strongly affected by linguistic, historical and cultural factors. Andersson and Persson (1993) include travel time by air between two countries as a proxy for distance, and add other geographical variables such as language similarity. They find a rather strong effect for both.

Katz (1994) seems to be the first to isolate geographical effects from other factors for inter-organisational collaboration within a country. Using datasets on intranational university-university collaboration, he finds that the frequency of collaboration between domestic universities declines exponentially with distance between the partners.

Nagpaul (2003) investigates collaboration patterns among 45 countries, using country as the observational unit, and using the price of airline tickets between capital cities as a proxy for distance between countries. He tests for what he calls thematic proximity, socio-economic proximity and geographical proximity. Even with the rather crude proxy for geographical distance he finds that geographical proximity is the most important determinant of these three.

Sutter and Kocher (2004) study a sample of articles in economic top journals from five years between 1977 and 1997. The analysis is restricted to the US by only including articles where all authors listed at least one US affiliation. They test for several geographical variables (spatial distance, but also being in the same or an adjacent state). Strikingly, they find that none of these geographical factors is significant.

Hoekman et al. (2009) employ a gravity model for biotechnology and semiconductors in subnational regions in Europe. Testing both patents and publications, they find that physical distance continues to play a role in collaboration; in addition they find a strong bias towards collaboration within nations.

Matthiessen et al. (2010) analysed the data of the hundred largest cities in the world by research output. They find that the research connectivity (strong co-authorship links) between these cities is influenced by the geographical proximity.

A popular hypothesis in the literature on the relevance of geographical distance is that the importance of distance decreases over time, as modern infrastructures would enable researchers to overcome the barrier caused by distance. Smith and Katz (2000) investigate this for the UK in the periods 1981-1983 and 1992-1994. Comparing the two time intervals, they find that the average distance between collaborators in the life sciences has indeed increased during the years. However, in the natural sciences, engineering and multidisciplinary research it remained more or less stable over time. Havemann et al. (2006) conducted an analysis for a sample of German immunological institutes for the time span of 1992 till 2002. The remarkable result is that distance did not matter, neither in the beginning nor at the end of this period. Collaboration was a bit higher with institutes in the same town, but outside the own town distance had no effect. This did not change over the years.

Hoekman et al. (2010) conduct a study on the role of geographical proximity for collaboration among institutions in European regions over the years 2000-2007. They use multiple indicators for geographical proximity, including physical distance between the regions and dummies for being in the same nation, in adjacent regions, in the same

language area, etc. They find that the effect of physical distance did not increase over time, but the effect of territorial borders did decrease.

Our study deviates from earlier work in at least two ways. First, in spatial level: most studies so far either used data on a supranational level (for example an entire continent) or at a very small scale (one organisation or building). If the level is one nation, than it is usually a relatively large one (Germany in Havemann et al., 2006; UK, Canada and USA in Katz, 1994; USA in Suttermann and Kocher, 2004; UK in Smith and Katz, 2000). Our level is a small European country, the Netherlands. As far as we know, the only other study on a small country (coincidentally also the Netherlands) is the one by Ponds et al. (2007). However, they use the so-called NUTS3 regions as unit of analysis, and average travelling time between those regions as proxy for distance. It is not unlikely that geographical proximity has a different effect in small countries than in large ones. Distances between any two cities are so (relatively) small here that it cannot be compared to for example distance between American or even German or French cities. This may also influence the perception people have of distances (and the efforts required to overcome them). As we measure the distance between locations at city level (rather than regional or national level as others do), we can assess the effect of small differences in distance. Second, our study uses empirical data of a very specific research field, the water sector. Although the water sector is ubiquitous in the sense that (almost) every country in the world produces knowledge on the production and transport of drinking water and wastewater. However, a lot of this knowledge is contextualized for specific local conditions. This may form an additional incentive for actors in the water sector to collaborate with local partners.

3. METHODOLOGY

One of the great advantages of using co-authorships as a proxy for collaboration is that one can easily construct datasets that are large enough for quantitative (bibliometric) analysis. Statistical methods are commonly employed to infer relationships between counts of the co-authorships and explanatory variables. The methods can vary in the level of analysis, the way the co-authorships are counted, etc. Such differences are discussed below. In addition, we explain the choices we have made.

Although analysing co-authorships is the most common way of studying collaboration, it is not unproblematic. Not all forms of (fruitful) collaboration result in joint papers, and the mere fact that several people are listed as authors does not imply that they did collaborate during the research phase. There have been a few attempts to quantify the extent of these limitations. In a small-scale study at Umeå University, Melin and Persson (1996) found that less than 5% of the authors indicated that they had experienced situations where collaborative work did not result in co-authored articles. The main reason was that the contribution was considered too minor. More specifically, Laudel (2002) shows on the basis of interviews and bibliometric research that whether or not a collaboration results in co-authoring depends on what the collaboration entails. She distinguishes six types of collaboration: a division of labour, service collaboration, provision of access to research equipment, transmission of know-how, mutual stimulation, and trusted assessorship. Almost all collaborations based on division of labour resulted in co-authorship; in the exceptional cases where it did not this was most likely because the collaborative work failed to produce publishable results. All other forms of collaboration were rewarded with co-authorship rarely (service collaboration) or not at all (other categories). It is hence likely that our dataset mainly contains information on (successful) collaborations where a division of labour was made.

An additional issue can be important if one analyses collaboration at an institutional level (i.e. assessing whether people affiliated with institute A collaborate

with people affiliated with organisation B). Especially in science it is not uncommon for people to be affiliated to more than one institute. It is questionable whether or not this should be counted as a collaboration from a conceptual perspective. From a practical perspective it is often close to impossible to avoid including dual institutional identities. Until recently, the Web of Science did not directly link authors to their affiliations, but instead provided separate lists of authors and research addresses per article. Katz and Martin (1997) report counts of papers listing more institutes than authors for a dataset taken from the Science Citation Index. The set contains papers from UK, Canada and Australia, in a broad range of scientific fields. The outcomes differ strongly per field, and range from less than 5% of the papers having more institutes than authors in fields like chemistry and engineering, to over 40% in clinical medicine.

If one is to assess the role of geographical distance in collaboration patterns (as we do in this article), there is one more point with co-authorships to bear in mind. It may well happen that authors from different countries all have an affiliation with one specific institute, and list only that institute when writing an article together. Indeed, the reverse could also happen if one researcher has more affiliations in distant places (see also Katz and Martin, 1997; Wagner and Leydesdorff 2005).

Level of Analysis

Co-authorships are often initiated and carried out at individual level. However, it does make sense to analyse the patterns of co-authorship at higher levels of aggregation as well, for example at the level of research groups, departments, institutions, regions or countries. Most policies regarding collaboration in research aim at such levels rather than the level of interindividual collaboration (Katz and Martin, 1997). As our main focus is on the relevance of geographical distance, it seems most relevant to conduct an analysis at organisational level, as it is not possible to retrieve data on addresses at a more detailed level than organisations. Of course, theoretically it would still be possible to link

institutional addresses to observations per individual researcher. However, this leads to practical problems, both because of the amount of data processing that this requires, and because most datasets do not link addresses to individuals (i.e. per article a list of authors and a list of organisations is given, but it is not possible to see which affiliation belongs to whom). In practice, most studies with attention for geographical variables conduct analyses at organisational level (see e.g. Katz, 1994; Havemann et al., 2006).

Retrieve Data

In principle, data on co-authorship can be retrieved from almost any extensive bibliographic database. Thomson Reuters Web of Science (WoS) is believed by some experts to be the most reliable source for a comprehensive survey of co-authored publications (e.g. Wagner and Leydesdorff, 2005; Melin and Persson, 1996). That is why we used the WoS to retrieve our data. Like every other bibliographic database, the WoS has a bias: it underestimates the social sciences and humanities does not cover Asia as well as other databases do. By including the WoS conference proceedings indexes, we improve coverage of the technical sciences. We have not distinguished between types of output (journal articles, letters, reviews, proceedings) as we were interested in connections between people, not in the scientific status of those connections (see Wagner and Leydesdorff, 2005, for a comparable argument).

First, a topic search was carried out using the search terms “drinking water”, “water treat*” and “desalinat*” (where the asterisk is a boolean operator for unknown characters). Experts were consulted to validate the initial dataset and adjust the search terms. Based on the dataset that was initially recovered for the period 1969-2008, it was determined that five journals published the largest part of these publications: Desalination, Water Research, Environmental Science & Technology, Water Science and Technology, and the Journal of the American Water Works Association. All articles published in these journals were downloaded. The keywords mentioned in the articles

were used to develop a more refined set of keywords for topic search. Based on this topic search the final set of publications was generated. To keep the amount of data within workable limits, only the publications from 2006 to 2008 were used. From the final dataset, we extracted publications that include at least one author with an affiliation in the Netherlands. The result is a set of 2,227 publications from 307 organisations (the nodes in the network), representing 646 co-authorship links (the edges in the network) between organisations from the Netherlands. The number of publications per organisation has been used as a proxy for the size (or capacity) of the institute as far as water-related knowledge production is concerned. The 646 co-authorship links were used in determining the relevance of distance.

All publications in the dataset contain details on the affiliations of the authors, including the address of the organisation. There are often minor variations in the way names and addresses are written. All institutes have been given a unique name and reference code that harmonises the address information provided by the WoS. The institutional affiliations have been accepted at face value. No attempt has been made to exactly reconstruct the organisational structure of universities and research institutes. For example, universities faculties are identified explicitly as part of a university; research labs may belong to a faculty but are often mentioned only as part of the university; and some inter-university research groups and university spin-offs are mentioned as separate institutions rather than as part of larger organisations.

Distance Matrix

After generating the dataset with the organisations, total number of publications per organisation, and number of co-authored articles between any possible combination of two organisations, the addresses were used to determine the town where the organsation is located. Sometimes one organisation appeared to have locations in several places. In such cases, the town that was most frequently mentioned on the articles was selected as

location for publications from that organisation. This was done since the analysis is carried out at organisation level, hence it would yield biases if some organisations were split up. It turned out that the 307 institutes had their locations in 97 towns throughout the Netherlands. These locations were georeferenced (i.e. longitude and latitude were collected). The result can be projected on a map, see figure 1. The quickest route between any combination of the locations was determined, and collected in a distance matrix. This distance matrix is linked to a matrix containing all possible ($307^2/2$) combinations of co-authorships between organisations (so including the ones that have zero co-publications). Moreover all institutes were categorised to their type of institute (universities, medical research centres, (semi)public research organisations, consultancies, production industry, governmental bodies or other). In as far as the name did not reveal a category, the website of the organisation was accessed to get information for categorisation.

Counts

Before one can analyse the linkages between institutions, one first needs to determine a way to count the linkages. For the questions under consideration in this article, we are interested in a measure that indicates how 'attractive' (for collaboration) institution X is to all other institutions in the dataset. This can be measured by counting the number of 'pairings' an institution has with other institutions. In other words: the two-way collaborations are counted. To give an example, if a paper lists four affiliations, A, B, C and D, this will be counted as six collaborations with value '1' each: A-B, A-C, A-D, B-C, B-D and C-D. If there are more authors from one institution involved, the collaboration will still be counted as one (see Katz, 1994 for further elaborations on this counting technique). There are different ways to assign weights to the different relations (see Katz, 1994 and Luukkonen et al., 1993 for overviews). We use integer counting, i.e. if there are, say, 4 different articles with authors from institute A and authors from institute B, then the link A-B has value 4. Some other techniques (e.g. Jaccard's index,

Salton's index) correct for the size (in terms of scientific output) of the institutes; this is not necessary in our case as we include size of the institutes as a variable in our model.

It is important to take all possible pairs of institutions into consideration, including the pairs with zero co-authored publications. The argument is similar to the one in other gravity models (for example on international trade): excluding zero-flows from the analysis implies an important loss of information on low levels of interaction. The explanatory variables may also (partially) explain why some organisations have no co-authorships with each other at all (Eichengreen and Irwin, 1998; Havemann et al., 2006).

Model

There is an increasing literature on the role of spatial and geographical factors in research collaboration. However, most of the studies so far seem to be descriptive in nature (Hoekman et al., 2010). The estimated model is usually a variant of the one by Katz (1994) who fitted a regression line $y=ae^{-bd}$ where y denotes the frequency of bilateral cooperation, d is distance, *e* is the base of the natural logarithm, and a and b are the parameters to be estimated.

We chose to use a so-called gravity model. The general idea behind this model is that some phenomena in the social sciences can be described by an analogy of Newton's gravitation law, namely that the gravitational force between two entities can be explained by the mass of these entities and the distance between them. The modern use is popularised by Stewart (1948) and a few years later the model was improved by (among others) Isard (1960). It is used to explain phenomena ranging from marriages to phone calls, and has been very popular especially in theories on international trade, initiated by work from Jan Tinbergen in the early 1960s (Hoekman et al., 2010; Santos Silva and Tenreyro, 2006). It has also been popular in quantitative geography, where it formed the core of a large body of literature on spatial interaction models (Murray, 2010). The rationale for using a gravity model to estimate the effect of physical distance is threefold:

first, including the mass (size) of the collaborating organisations in the model makes it much more realistic: organisations that produce more publications will naturally have more co-authorships (in absolute numbers). Second, the multiplicative nature of a gravity model has proven to provide a better fit to empirical data than additive (linear) models for many different phenomena. Third, the gravity model is analytically convenient. Its intuitive basic specification can easily be enriched by adding other relevant variables, such as size and quality indicators of the organisations, cultural variables (e.g. main language), etc (see also Sutter and Kocher, 2004).

Others have also used a gravity model in the analysis of co-autorship patterns. Beckmann (1994) builds a theoretical argument why a gravity would be suitable to explain research collaboration. However, he does not empirically test whether it exists. Sutter and Kocher (2004) seem to be the first to empirically test a gravity model. They try to find geographical patterns in the distribution of co-authorships in the economic departments of universities in the United States. Hoekman et al. (2010), who conducted an analysis at NUTS2-level in Europe, developed a more sophisticated composite variable for distance. It comprises five different dimensions: region, country, language area, spatial distance, and thematic similarity of the research. However, some of the dimensions have very high correlation rates in their research already, and do not make much sense in an analysis at institutional level for one (relatively small) country.

The most general form of the gravity model can be described as follows:

$$(1) \qquad N_{ij} = A(i)B(j)F(d_{ij})$$

Where N_{ij} denotes a measure of the interactions between origin *i* and destination *j*, A is an function of the origin, B is a function of the destination, and F is a function of the separation (or distance) between *i* and *j* (Sen and Smith, 1995).

More specifically for our case:

$$(2) \qquad co_{ij} = \beta_0 d_{ij}^{\beta 1} m_i^{\beta 2} m_j^{\beta 3}$$

Where co_{ij} is the number of co-authorships between institute i and j, d_{ij} is the geographical distance between these institutes, and m_i and m_j are the mass of institute i and j respectively (measured as total scientific output on the watercycle of the institute). The unknown parameters to be estimated are β_0 to β_3. For a Newtonian gravity model, it would hold that β_1=-2 and β_2= β_3=1.

It used to be very common to turn the equation into a log-additive model for analytical convenience. In other words: at both sides of the equation the logarithm of all terms is taken; this makes the task of estimating the unknown parameters a lot easier, as all the exponents disappear because of standard mathematical rules (Sen and Smith, 1995). Doing so yields:

$$(3) \qquad \ln(co_{ij}) = \ln(\beta_0) + \beta_1 \ln(D_{ij}) + \beta_2 \ln(m_i) + \beta_3 \ln(m_j) + \mu_{ij}$$

where μ_{ij} is an independent random variable which is normally distributed with zero mean and identical variance σ^2. In other words, it is an error term needed for empirical estimation.

In spite of the large number of studies that employ a version of this model, estimating it by applying ordinary least squares (OLS) regression is not as straightforward as it may seem at first sight. The underlying reason is in what is called Jensen's inequality, implying that E (ln (y)) ≠ ln (E (y)). In other words: the expected value of the logarithm of a random variable is not equal to the logarithm of the expected value of that same variable (Santos Silva and Tenreyro, 2006). If one uses a log-linear model as described above, the regression will produce estimates of the logarithm of μ_{ij}, not of μ_{ij} itself; the antilogarithms of these estimates are biased estimates of μ_{ij}. Ignoring that problem leads to systematic under-prediction of large values of the dependent variable (Flowerdew and Aitkin, 1982).

A second problem concerns the error term. If one is to apply OLS on this equation, it must be assumed that μ_{ij} is normally distributed. That in turn implies that the values of co_{ij} are log-normally distributed around the estimate. However, as co_{ij} is

measured as binary co-authorships (with integer counts), they must be nonnegative integers, hence their distribution will not be log-normal.

A third issue is related to the assumption that the variance of the error term is identically distributed. In other words, the expected difference between the log of the observed and the log of the estimated value is the same for every pair of institutions. In practice, this implies that the probability of finding an observed value of 2 with an estimated value of 1 is equal to the probability of finding an observed value of 200 with an observed value 100. As most linkages in our dataset have low observed and expected values, small absolute differences may give large differences between the observed and expected values in logarithmic form.

Last but not least it is important to realize that the logarithm of a value 'zero' is not defined. Deleting all observations with zero co-publications is no option, as it leaves out important information on extremely low levels of collaboration. This gives biased results, particularly if the zero-valued observations are non-randomly distributed, as is the case in our dataset, see figure 3 (Eichengreen and Irwin, 1998; Burger et al., 2009). This problem is often circumvented by adding a small positive number to all observations. However, if many of the observations have a value of zero (as is the case in our study for the number of coauthorships), the exact value of that added constant does have considerable impact on the coefficients and explanatory power of the model. It can even be shown that you can generate any desired parameter estimate by adapting the value of the added constant (Flowerdew and Aitkin, 1982; King, 1988).

These problems can be overcome by assuming a different distribution. Recall that each observation of co_{ij} is a nonnegative integer, and hence co_{ij} can be considered as having a discrete probability distribution. If there is a (small) constant probability P_{ij} that organisation i and j co-author a publication (and if co-authorships can be assumed to be independent of each other), then the number of copublications of i and j follows a so-called Poisson distribution. Denoting the mean of the distribution as λ_{ij}, it can be derived that the probability that i and j have exactly k copublications is

$$(4) \qquad P(co_{ij} = k) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{k}}{k!}$$

(see Flowerdew and Aitkin, 1982 for more details). Poisson distributions do not assume equal variation, but variation dependent on the conditional mean λ_{ij}, one can hence apply a weighted least squares (WLS) regression, with a weight function of the (unknown) λ_{ij}. In an iterative (converging) procedure, the value of λ_{ij} can be estimated. This turns out to be equivalent to using a maximum likelihood (ML) estimation for Poisson random variables. The model is therefore also known as the Poisson pseudo-maximumlikelihood (PPML) estimator. PPML is a viable alternative for log-normal models; it keeps the multiplicative structure of the gravity model, but it does not suffer from the shortcomings of log-normalizing (Flowerdew and Aitkin, 1982; Santos Silva and Tenreyro, 2006).

One of the basic assumptions of the Poisson distribution is that the variance of the dependent variable is equal to its mean. In our dataset the variance (.159) is much larger than the mean (.027), in other words there is overdispersion. However, in the Poisson family there is also a distribution that allows for variance higher than the mean. This is the negative binomial distribution, which is standard available in econometric software packages like Stata. The expected value of co_{ij} will remain the same as in a Poisson model, but the variance has one more free parameter: it is a function of the conditional mean λ_{ij} and a dispersion parameter α. The dispersion parameter can model between-subject heterogeneity, and in this way overdispersion can be taken care of (Burger et al., 2009).

The probability mass function of a negative binomial model is:

$$(5) \qquad P(co_{ij} = k) = \frac{\Gamma(co_{ij}+\alpha^{-1})}{co_{ij}!\,\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda_{ij}}\right)^{\alpha^{-1}} \left(\frac{\lambda_{ij}}{\alpha^{-1}+\lambda_{ij}}\right)^{co_{ij}}$$

Where co_{ij} again is the number of coauthorships, Γ denotes the gamma function, α is the dispersion parameter, and λ denotes the conditional mean. This estimator is also known as a negative binomial pseudo-maximum likelihood model (NBPML). A likelihood ratio

test can be used to test whether α is significantly different from zero and hence whether a negative binomial distribution is preferred over a Poisson distribution (Burger et al., 2009).

4. RESULTS

In this section we present our findings. We first show a visualisation of the geographical position of all organisationans in our dataset. Then we visualise the co-authorships between the organisations. After that we show with boxplots that the observations without co-authorships (the zeros) are non-randomly distributed in physical distance. Last we show the results of two specifications of the gravity model, one with and one without institutional dummies.

Mapping Co-authorships

We devised our dataset in such a way that only publications with all affiliations in the Netherlands are included. It contains 97 different cities in total. Once we added their geographical coordinates, we can print the locations of all involved institutes on a map (Figure 1).

INSERT FIGURE 1 ABOUT HERE

Moreover, once we have combined the data on the locations of the institutes with the data on co-authorships, we can map all institutes on their geographical position, and map all collaborations between them (Figure 2). In this figure, the size of the node and the font size of its name represent the mass of the institute (as measured in number of publications in our dataset); the thickness of an edge and the saturation of its colour represent the number of co-authorships. The resulting image already indicates that most co-authorships occur among organisations in a relatively small part of the country. There is a ‘belt’ of

dense collaborations from west to east in the middle of the country. The image does not suggest the existence of regional clusters everywhere in the country. In other words: the picture reveals that collaborative research does not seem organised solely in regions (like the covered area of drinking water companies), yet there is a clear effect of distance.

INSERT FIGURE 2 ABOUT HERE

Comparing the Groups With and Without Collaboration

After visualising the effect of distance on the occurrence of co-authorships, we analytically prove the effect and measure its impact. As a first exercise we compare the distance of the organisations that have one or more co-authorships with the distance of all pairs of organisations in the dataset that have no shared co-authorships. If distance would not play any role in developing co-authorships, these distances should be more or less equal. The data on distance are not normally distributed, so we cannot employ any parametric tests. Hence we use a Mann-Whitney test to see if there is a genuine difference between the distance between organisations with actual co-authorships and the distance between random combinations of organisations. The set with pairs of organisations that do not have any co-authorships consists of 46035 observations, the sets with actual co-authorships consists of 646 observations. There is a clear difference between the two groups; the median of the distance in the group without co-authorships is 105.0 km; the median of the group with co-authorships is 76.5 km ($Z= -10.052$, $p = .000$). The distribution of distances in the two groups is shown with boxplots (Figure 3).

INSERT FIGURE 3 ABOUT HERE

There is hence a clear difference in distance between the group with co-authorships and the group without; more proximate organisations have higher odds of collaboration. This

confirms that the selection of a gravity model is appropriate. We will now turn to the results of that model.

Gravity Model

We first show the results of the basic gravity model (Table 1). This model has three explanatory variables: the mass of organisation A (measured by the total number of publications of that organisation in the dataset), the mass of organisation B, and the distance between A and B. All three explanatory variables are highly significant ($p<.001$). The direction is as expected: physical distance (distance_ab) is negative: the larger the distance, the less co-authorships ($Z= -13.31$; $p<0.001$); the size of organisation A and B (mass_a and mass_b) are both positive: the larger an organisation, the more co-authorships it has ($Z=36.48$; $p<.001$ and $Z=37.25$; $p<.001$ respectively). The coefficients can be interpreted as follows: for an increase of one unit in the explanatory variable, the log of the expected count of the dependent variable is to change with the respective regression coefficient. So if for example the natural logarithm of distance increases with one unit, the model predicts a decrease in the log of co-publications with .526 unit. The dataset is constructed in such a way that every possible combination of two organisations in the set occurs exactly once. It is hence arbitrary whether an organisation is mentioned as "A" or "B". The fact that the coefficient of mass A is a bit higher is merely coincidence.

As explained in the methodology section, this model employs a negative binomial distribution with a pseudo-maximum likelihood estimator. The R^2 is therefore not defined. The pseudo-R^2 is 0.318.

INSERT TABLE 1 ABOUT HERE

One may argue that with an NBPML model a so-called Eicker-White robust covariance matrix estimator should be used because of potential heteroskedasticity in the model (Santos Silva and Tenreyro, 2006). This does not alter the conclusions of the model and it hardly changes the findings. The first three decimals of all p-values remain the same; the Z-value change a tiny bit: distance to -12.43, mass A to 33.71 and mass B to 39.54.

In the methodology section we explained that the choice for a negative binomial model stems from the overdispersion in our data. A likelihood ratio (LR) test can be performed to test if the additional parameter α is significantly different from zero. This is indeed the case ($p < .001$), which confirms that a negative binomial distribution is indeed more appropriate than a 'common' Poisson distribution because of the overdispersion. The model is robust however for slightly different specifications of the model; specifying it as a Poisson model does not alter the conclusions. The model can also be specified as a Generalized Linear Model (GLM) with Maximum Likelihood (ML) optimization, if the distribution is set to negative binomial and the "link" as logarithmic. This leads to the same conclusions.

Extended Gravity Model

One of the advantages of the gravity model is that it can be extended with additional variables very easily. How many co-authorships two organisations have together is not only determined by their size and the physical distance between them. One of the other relevant factors is probably whether or not the two institutions belong to the same category of institutions (firms, universities, governmental bodies, etc). Different types of organisations have different cultures, incentive systems, a different kind of knowledge base, etc. In fact, such differences can also be seen as dimensions of distance (see Boschma, 2005 for an overview of such dimensions). To test empirically whether this makes a large difference, we have extended our gravity model with dummies for different types of organisations. Seven different types of organisations are distinguished:

universities, academic hospitals, (semi)public research organisations, consultancy firms, government bodies, industrial firms, and others. The dummies are constructed in such a way that they take a value of one if the two organisations both belong type of organisation (say universities), and a value of zero otherwise. All observations with a pair of two different types of organisations serve as a baseline. The results are presented in table 2. The variables from the basic model hardly change. The direction of distance is still negative and very significant (Z=-12.90; p<.001). The size of the organisations is positive and strongly significant (Z=35.35; p<.001 and Z=35.96; p<.001 respectively). Universities and (semi) public research organisations and others do not significantly differ from the baseline. The category of academic hospitals ("medical") is significantly higher (Z=8.10; p<.001); consultancies are significantly higher (Z=5.95; p<.001); industrial are significantly higher (Z=4.84; p<.001); and governmental bodies are also significantly higher (Z=3.32; p=.001). This means that a pair of two organisations with a given size and a given distance between them will collaborate more if they are both medical organisations, both consultancies, both industrial firms or both governmental bodies, compared to two organisations of a different type. If they are both universities, (semi) public research organisations or others, they do not collaborate more than two organisations of different types. Again, using a model with robust standard errors only causes slight modifications in the Z-values. P-values only change for the three dummies that were insignificant already, their p-values become even higher. The LR test confirms that in this situation a negative binomial distribution is again appropriate.

INSERT TABLE 2 ABOUT HERE

However, the impact of all the dummies together is very limited: the pseudo R^2 increased only slightly (pseudo R^2 =.330). That does not imply that it does not matter whether or not collaborating organisations are of the same type. It does suggest however that the

dummies may explain part of the variance that is already explained by the variables of the basic model. To verify this, we have constructed a correlation matrix, see table 3.

INSERT TABLE 3 ABOUT HERE

The correlation matrix shows that although many correlations are significant, they are not very strong. Apparently, for our data physical proximity and the size of the organisations have a much stronger impact on collaboration than whether or not (potential) collaborators belong to the same type of organisation.

5. CONCLUSIONS AND DISCUSSION

The production of scientific knowledge is increasingly a collaborative effort. One of the factors that explain the intensity of collaboration between organisations, is physical distance. Physical distance can work differently at different spatial levels. There is a growing body of empirical literature on the effect of distance, using co-authorships of publications as a proxy for collaboration. However, so far the literature basically focused on two levels: the micro level of one organisation or building, and the macro level of a large group of countries or an entire continent. In this study we show that at least in a geographically delineated sector as the water sector there is also an effect of physical distance at a level between the micro and macro level, namely that of a small country. Employing a gravity model, we show that, controlling for the size of organisations (i.e. the total amount of papers they contribute to in the water sector), there is a clear negative relation between physical distance and the number of co-authorships organisations have. This is in line with the findings of Ponds et al. (2007), who find that the travel time between regions has a negative effect on the intensity of collaboration. In other words: if distance between two organisations increases, the number of co-authored publications to which they both contributed decreases.

In a more elaborate specification of the gravity model we have tested whether two organisations that belong to the same type of societal organisation (universities, governmental bodies, etc.) also have more intensive collaboration. Although for most types of organisations this is the case, the effect is not very strong. The effect of distance remains equally significant and equally strong.

Other studies have shown that not only geographical distance, but also other dimensions of distance (such as cognitive, social, organisational, institutional) matter for research collaboration, although empirical evidence of the interactions among such dimensions is scarce and very scattered (Boschma, 2005; Knoben and Oerlemans, 2006). The fact that two collaborators belong to the same type of organisation is an indicator of organisational distance, albeit a rather crude one. However, Ponds et al. (2007) use more or less the same indicator (they distinguish academic institutions, commercial organisations and governmental organisations). That does not cause increases in the pseudo R^2 of their models, but surprisingly, the effect of traveltime between regions (the proxy for physical proximity that Ponds et al. use) weakens in the physical sciences and even becomes insignificant in a few specific subsectors, especially for academic-academic collaborations. This effect does not occur in the life sciences. They suggest that geographical proximity helps to overcome institutional or organisational differences between academic and non-academic organisations (and more so in the physical sciences, because it has a more mature structure with longer established relations between actors). However, the Dutch water sector has a long tradition of collaboration between different types of organisations (academic, semi-public, commercial, governmental), and still the effect of geographical proximity (and organisation size) is much stronger than the effect of whether or not the collaborators belong to the same type. This may be due to the contextualised knowledge effect, or due to historical factors: actors stick with their established, geographically proximate collaborations.

There are two specific reasons that may help explain why we find a rather strong effect with these data. First, there may be a “contextualized knowledge effect”: the

knowledge that is disclosed in the co-authored publications is not universal, but adapted to special local questions and circumstances. This phenomenon occurs in many research areas, but the water sector is known for its specificity in knowledge, because of the dependency on local environmental conditions. A relatively applied area of research like the water sector may also require quite some non-codified (tacit) knowledge. This may of course induce a tendency to collaborate mainly with geographically proximate collaborators.

Second, there may be a "small country effect": in small countries, people probably perceive distance relatively quickly as prohibitively large for collaboration, as they are used to a geographically dense network, with many partners in close proximity. This holds probably even more for the Netherlands as it is a densely populated area. Moreover, if people are to collaborate over larger distances, they may have incentives to do so across national borders. International research collaboration is more prestigious and is known to have a larger citation impact (Narin et al., 1991; Katz and Martin, 1997). The open borders in the European Union and modern communication means make international distances relatively easier to overcome. Earlier research on the impact of geographical distance has shown that the impact of borders in Europe on co-authorships has decreased, but the impact of absolute physical distance has not (Hoekman et al., 2010). More prestigious research may be less prone to the effect of distance: a study that only included publications in top economic journals (and only if the authors had an affiliation in the United States), finds no effect of distance (Sutter and Kocher, 2004). However, there in general there is still a bias to collaborate domestically (Frenken et al., 2009). Together with our findings this suggests that researcher may either opt for long-distance, international collaboration, or for intranational collaboration; if they choose for the latter they strongly prefer to collaborate across small distances.

Three issues deserve more elaboration in future research. First, more research is needed on the underlying causes of the effect geographical proximity has. Is it a deliberate choice of researchers to have local collaborators? We can think of at least three

causes: First they may be convinced that their questions are so much contextualised and localised that only local partners can be of use in answering them. Second, they could be so much inward-looking that they do not meet potential partners from less proximate places. Third, transaction costs of maintaining a long-distance collaboration could be prohibitive. The second issue is related to that: there are both factors that promote proximate and factors that promote long-distance collaboration. In this study we have shown that the former are stronger at the spatial scale of a small country. However, there is very little insight in the interactions between these "push and pull" mechanisms. The third issue that deserves future research is the interaction with other dimensions of proximity (see also Frenken et al., 2009). Our analysis reveals that indicators for the type of organisation to which collaborators belong are relevant for the intensity of collaboration. This is confirmed by other studies. Yet, our understanding of the interplay of the different dimensions of proximity should be extended.

FOOTNOTES

1) A second reason to use co-authorships is that the outcomes are easier to verify: other researchers can obtain the same dataset, hence they should be able to replicate the results (Smith and Katz, 2000; Subramanyam 1983).

2) We define geographical distance as a broader category than spatial or physical distance alone. Some authors have studied the effect of national or regional borders for example; this is a form of geographical distance but not of spatial/physical distance.

REFERENCES

Acedo, F. J., C. Barroso, C. Casanueva & J. L. Galán (2006) Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis. *Journal of Management Studies,* 43**,** 957-983.

Allen, T. J. (1977) *Managing the flow of technology: Technology transfer and the dissemination of technological information within the research and development organization.* Boston, MA: Massachusetts Institute of Technology.

Andersson, Å. E. & O. Persson (1993) Networking scientists. *The Annals of Regional Science,* 27**,** 11-21.

Beaver, D. D. (2001) Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics,* 52**,** 365-377.

Beckmann, M. J. (1994) On knowledge networks in science: collaboration among equals. *The Annals of Regional Science,* 28**,** 233-242.

Boschma, R. (2005) Proximity and innovation: a critical assessment. *Regional Studies,* 39**,** 61-74.

Burger, M., F. Van Oort & G. J. Linders (2009) On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. *Spatial Economic Analysis,* 4**,** 167-190.

De Solla Price, D. J. & D. Beaver (1966) Collaboration in an invisible college. *American psychologist,* 21**,** 1011.

Eichengreen, B. & D. A. Irwin (1998) The role of history in bilateral trade flows. In: Frankel, J.A. (Ed.) (1998) *The Regionalization of the World Economy.* Chigago: University of Chigago Press, pp. 33-63.

Flowerdew, R. & M. Aitkin (1982) A Method of Fitting The Gravity Model Based on the Poisson Distribution. *Journal of Regional Science,* 22**,** 191-202.

Frenken, K., S. Hardeman & J. Hoekman (2009) Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics,* 3**,** 222-232.

Hagstrom, W. (1965) *The Scientific Community*. New York: Basic Books.

Havemann, F., M. Heinz & H. Kretschmer (2006) Collaboration and distances between German immunological institutes–a trend analysis. *Journal of Biomedical Discovery and Collaboration,* 1**,** 6.

Hoekman, J., K. Frenken & R. J. W. Tijssen (2010) Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy,* 39**,** 662-673.

Hoekman, J., K. Frenken & F. Van Oort (2009) The geography of collaborative knowledge production in Europe. *The Annals of Regional Science,* 43**,** 721-738.

Isard, W. 1960. *Methods of regional analysis: an introduction to regional science*. Cambridge, MA: MIT Press.

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

Katz, J. S. & B. R. Martin (1997) What is research collaboration? *Research Policy,* 26**,** 1-18.

King, G. (1988) Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science***,** 838-863.

Knoben, J. & L. Oerlemans (2006) Proximity and inter-organizational collaboration: A literature review. *International Journal of Management Reviews,* 8**,** 71-89.

Kraut, R., C. Egido & J. Galegher. 1988. Patterns of contact and communication in scientific research collaboration. In: *Proceedings of the Conference on Computer-Supported Cooperative Work,* Portland OR.

Laudel, G. (2002) What do we measure by co-authorships? *Research Evaluation,* 11**,** 3-15.

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

Maggioni, M. A. & T. E. Uberti (2009) Knowledge networks across Europe: which distance matters? *The Annals of Regional Science,* 43**,** 691-720.

Malecki, E. J. (2010) Everywhere? The Geography of Knowledge. *Journal of Regional Science,* 50**,** 493-513.

Melin, G. & O. Persson (1996) Studying research collaboration using co-authorships. *Scientometrics,* 36**,** 363-377.

Murray, A. T. (2010) Quantitative Geography. *Journal of Regional Science,* 50**,** 143-163.

Nagpaul, P. (2003) Exploring a pseudo-regression model of transnational cooperation in science. *Scientometrics,* 56**,** 403-416.

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

Paci, R. & S. Usai (2009) Knowledge flows across European regions. *The Annals of Regional Science,* 43**,** 669-690.

Ponds, R., F. Van Oort & K. Frenken (2007) The geographical and institutional proximity of research collaboration. *Papers in Regional Science,* 86**,** 423-443.

Sen, A. & T. E. Smith (1995) *Gravity models of spatial interaction behavior*. Heidelberg: Springer.

Silva, S. & S. Tenreyro (2006) The log of gravity. *Review of Economics and Statistics,* 88**,** 641-658.

Smith, D. & J. S. Katz (2000) Collaborative approaches to research. *A report to the Higher Education Funding Council for England.* Centre for Policy Studies in Education, University of Leeds.

Smith, M. (1958) The trend toward multiple authorship in psychology. *American psychologist,* 13**,** 596.

Stewart, J. Q. (1948) Demographic gravitation: evidence and applications. *Sociometry,* 11**,** 31-58.

Subramanyam, K. (1983) Bibliometric studies of research collaboration: A review. *Journal of information Science,* 6**,** 33-38.

Sutter, M. & M. Kocher (2004) Patterns of co-authorship among economics departments in the USA. *Applied Economics,* 36**,** 327-333.

Wagner, C. S. & L. Leydesdorff (2005) Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation,* 1**,** 185-208.

TABLE 1: Negative binomial regression for the basic gravity model

	Coefficient	Z-score
Ln (distance)	-.5260984***	-13.31
Ln (mass A)	.9718572***	36.48
Ln (mass B)	.8658232***	37.25
Dependent variable	Co-authored pubs	
N	46681	
Log likelihood	-2848.0934	
Pseudo R^2	0.3181	
LR chi2	2657.39 ***	
LR $\alpha = 0$	592.51 ***	

*** Two-sided significance at 1%-level.

TABLE 2: Extended gravity model with dummies for pairs of organisations of the same type

	Coefficient (p-value)	Z-score
Ln (distance)	-.5084501***	-12.90
Ln (mass A)	1.006832***	35.35
Ln (mass B)	.9222609***	35.96
Both university	.1834895	0.90
Both (semi) public res	.1482864	0.78
Both medical	1.672508***	8.13
Both consultancy	1.532714***	5.98
Both industrial	1.269188***	4.86
Both governmental	1.474479***	3.33
Both other	.3824729	0.99
Dependent variable	Co-authored pubs	

N	46681
Log likelihood	-2800.1604
Pseudo R^2	0.3296
LR chi2	2753.26
LR $\alpha = 0$	582.03(.000)

*** Two-sided significance at 1%-level.

TABLE 3: Correlation matrix of the dummies of the extended model against the variables of the basic gravity model.

	Ln (distance)	Ln (mass A)	Ln (mass B)
Both university	.004	.071***	.122***
Both (semi) public res	-.001	.057***	.038***
Both medical	-.024***	.010**	-.012***
Both consultancy	-.009**	-.045***	-.071***
Both industrial	.039***	-.048***	-.057***
Both governmental	-.010**	-.007	-.006
Both other	-.033***	-.020***	-.033***

** Two-sided significance at 5%-level.

*** Two-sided significance at 1%-level.

FIGURE 1: The locations of the organisations in our dataset; each dot depicts a city where one or more organisations are located

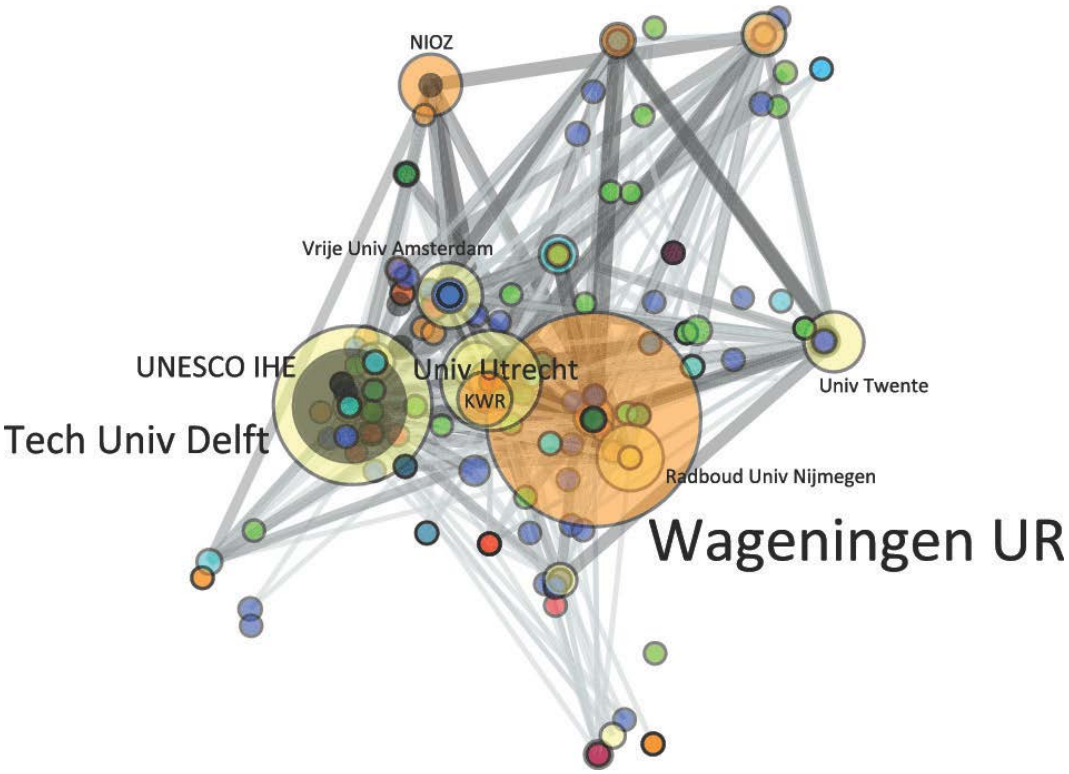

FIGURE 2: Map of the co-authorship patterns in our dataset; the size of the nodes represents the number of publications of the organisation, the thickness of the edges represents number of co-authorships

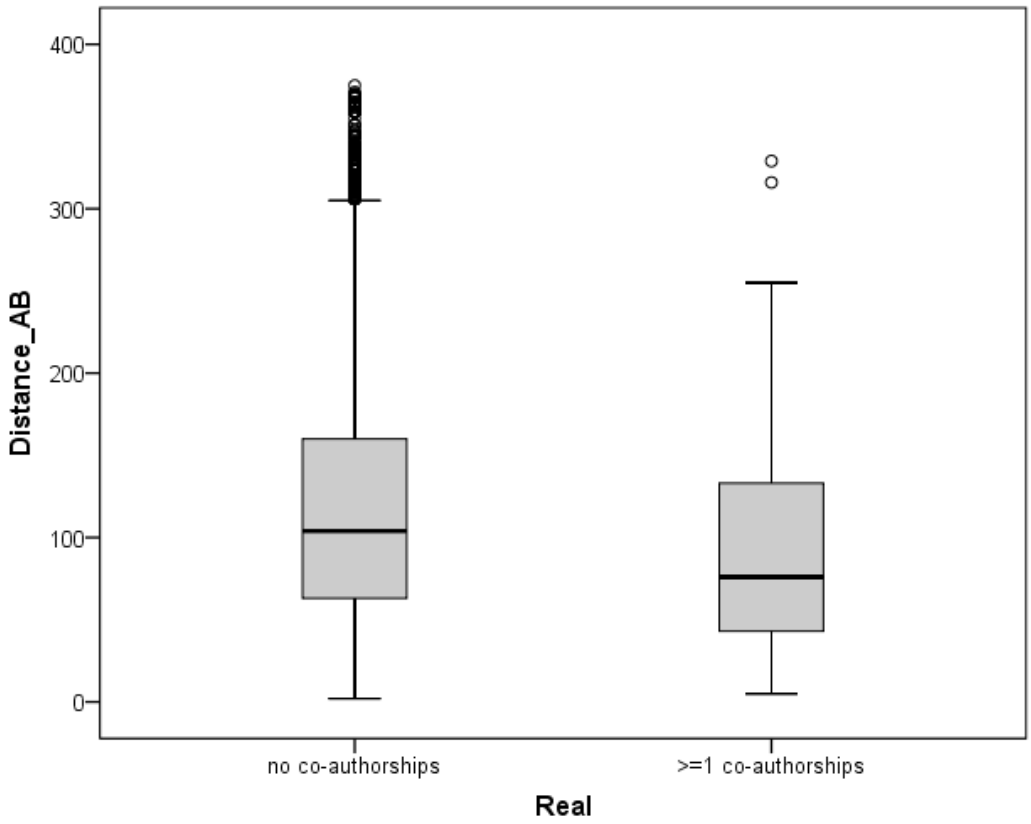

FIGURE 3: Boxplots of the distances between organisations with and without co-authorships respectively.