



Delft University of Technology

Using toponym co-occurrences to measure relationships between places review, application and evaluation

Meijers, Evert; Peris, Antoine

DOI

[10.1080/12265934.2018.1497526](https://doi.org/10.1080/12265934.2018.1497526)

Publication date

2018

Document Version

Final published version

Published in

International Journal of Urban Sciences

Citation (APA)

Meijers, E., & Peris, A. (2018). Using toponym co-occurrences to measure relationships between places: review, application and evaluation. *International Journal of Urban Sciences*, 23 (2019)(2), 246-268. <https://doi.org/10.1080/12265934.2018.1497526>

Important note

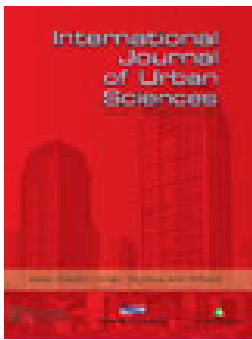
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Using toponym co-occurrences to measure relationships between places: review, application and evaluation

Evert Meijers & Antoine Peris

To cite this article: Evert Meijers & Antoine Peris (2018): Using toponym co-occurrences to measure relationships between places: review, application and evaluation, International Journal of Urban Sciences, DOI: [10.1080/12265934.2018.1497526](https://doi.org/10.1080/12265934.2018.1497526)

To link to this article: <https://doi.org/10.1080/12265934.2018.1497526>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 10 Jul 2018.



Submit your article to this journal [↗](#)



Article views: 97



View Crossmark data [↗](#)

Using toponym co-occurrences to measure relationships between places: review, application and evaluation

Evert Meijers and Antoine Peris

Faculty of Architecture and the Built Environment, Delft University of Technology, Delft, The Netherlands

ABSTRACT

While there is consensus that network embeddedness of cities is of great importance for their development, the precise effect is difficult to assess because of a lack of consistent information on relations between cities. This paper presents, applies and evaluates a rather novel method to establish the strength of relationships between places, a method we refer to as 'the toponym co-occurrence method'. This approach builds the urban system on the basis of co-occurrences of place names in a text corpus. We innovate by exploiting a so far unparalleled amount of data, namely the billions of web pages contained in the commoncrawl web archive, and by applying the method also to small places that tend to be ignored by other methods. The entire settlement system of the Netherlands is consequently explored. In addition, we innovatively apply machine learning techniques to classify these relations. Much attention is paid to solving biases deriving from place name disambiguation. Gravity modelling is employed to assess the resulting spatial organization of the Netherlands. It turns out that the gravity model fits very well with the pattern of relationships between places as found in digital space, which contributes to our assessment that the toponym co-occurrence method is a solid proxy for relationships in real space. Using the method, it is established that the relationships in the Randstad region, by many considered a coherent metropolitan entity, are actually somewhat less strong than expected. In contrast, historically important, but nowadays small cities in the periphery tend to have maintained their prominent position in the pattern of relationships. Suburban, relatively new places in the shadow of a larger city tend to be weakly related to other places. Several suggestions to further improve the method, in particular the classification of relationships, are discussed.

ARTICLE HISTORY



Received 9 February 2018
Accepted 2 July 2018

KEYWORDS

Urban system; place name disambiguation; city network; gravity model; semantic relatedness; Randstad

1. Introduction

Cities and regions cannot be studied in isolation. Their fate and fortune depends on how they are embedded in flows of goods, people, information and capital, as well as their absorptive capacity to use and exploit these flows. A wide range of literature has been

CONTACT Evert Meijers  e.j.meijers@tudelft.nl  Faculty of Architecture and the Built Environment, Delft University of Technology, Julianalaan 134, Delft 2628 BL, The Netherlands

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

stressing the importance of network embeddedness for urban and regional development (Camagni, 2017; Camagni & Capello, 2004; McCann & Acs, 2011; Meijers, Burger, & Hoogerbrugge, 2016; Meijers, Hoogerbrugge, & Cardoso, 2018; Neal, 2013a; Taylor & Derudder, 2016), discussing the existence of ‘urban network externalities’ (Burger & Meijers, 2016; Capello, 2000) that would complement, or even substitute local factors. Stressing the importance of non-local factors in the development of cities is not new and has been widely discussed by urban historians (Hohenberg & Lees, 1985) and geographers (Bourne & Simmons, 1978). However, the importance of networks between places and regions has always been obscured by the difficulty of obtaining consistent information on these networks or flows between places. This is why much of the research on the competitiveness of cities and their development still uses ‘stock’ data rather than ‘relational’ data.

The lack of evidence on networks between cities has been considered the ‘dirty little secret’ (Short, Kim, Kuu, & Wells, 1996) of research into networks of cities, especially on the global scale. Despite considerable progress over the last 20 years, which we review later, the availability and adequacy of data on relationships between cities still remains a critical issue because of a variety of problems. First of all, this relates to the use of indicators that only sketch a partial picture in that they cannot account for the multiplicity of networks, which refers to the fact that the spatial organization of different types of functional linkages is not necessarily identical (Berroir et al., 2017; Burger, Meijers, & van Oort, 2014; Limtanakool, Schwanen, & Dijst, 2009). Second, inadequate proxies are, or need to be used in the absence of more direct indicators. The most common example is that often accessibility is measured, not actual flows (e.g. Meijers et al., 2016). Resorting to inadequate proxies is partly linked to a third issue, namely that in particular information on flows at higher spatial scales (global, continental) is missing, given the mismatch with the scale covered by the main data-collecting agencies, the national statistical bureaus. And if supranational data is available, it often needs to be matched with national data sources for consistency and disaggregation to the level of cities or city-regions (an assumption-rich process in itself) – trade data being a case in point (Burger, Thissen, van Oort, & Diodato, 2014). However, a fourth concern is that even if a proxy appears successful in approximating flows on higher spatial scales, as the popularity of the interlocking network model for measuring global city networks suggests (Taylor, 2001; Taylor & Derudder, 2016), such a proxy does not necessarily work well for smaller spatial scales (Lambregts, 2009). Fifth, changes in networks between cities are often incremental and demand a considerable time period to become visible, but data often does not span such periods, and definitions and ways of collecting data change, hampering comparisons in time. Exemplary is the discussion on the adequacy of the central place model stressing hierarchy to describe urban systems, which static analyses tend to confirm (Schiff, 2015; van Meeteren & Poorthuis, 2018), whereas dynamic analyses point sometimes more towards the rise of a ‘network model’, which stresses horizontal relationships between similar-sized cities and an increasing disconnection between size and function in more polycentric territories (Batten, 1995; Meijers, 2007).

The possibilities of ‘big data’ have sparked new hopes to disentangle networks between cities and recent explorations based on data derived from social media or the internet have given new impetus to the study of intercity relationships. ‘Big Data’ typically refers to large datasets, mined in bulk from modern electronic devices, building often on social media platforms or on sensor networks, and is often crowd sourced. So far, the vast majority

of applications of big data does not move beyond the scale of individual cities (often under the umbrella term ‘smart cities’), but increasingly, also its potential for studying networks between cities has been recognized. We had already become used to real-time traffic information on major roads linking cities, but increasingly, applications serve to specifically understand the urban system, distinguishing urban nodes and links between them (e.g. Zhong, Arisona, Huang, Batty, & Schmitt, 2014). For instance, migration patterns between cities in China are derived from crowdsourced geotagged posts on Baidu during the Spring Festival (Xu et al., 2017). van Meeteren and Poorthuis (2018) test the micro foundations of central place systems using geo-tagged tweets and venues derived from foursquare, while Yuan and Medel (2016) derive international travel behaviour from geotagged photos on Flickr. Also, the referring link structure of Wikipedia is used to infer central place systems (Keßler, 2017). Other approaches involve the spatial structure of hyperlinks to study networks (Janc, 2012, 2015), the exploration of reciprocal relationships between cities in Google Maps’ data representation to explore how ‘close’ cities are in cyberspace (Zook, Devriendt, & Dodge, 2011) or revisiting the world city network based on geolocated tweets (Lenormand, Gonçalves, Tugores, & Ramasco, 2015), to name but a few.

A promising new avenue in the study of intercity relationships is becoming available with the increasing availability of digital archives. This paper explores the potential of what we call the ‘toponym co-occurrence approach’ that can be applied to such digital archives. The essence of this approach is that it (a) retrieves information on intercity relationships from text corpora in which places are mentioned together¹ (‘semantic relatedness’), and (b) uses machine learning techniques to excavate the context in which these place names co-occur in texts in order to categorize these relationships in a meaningful way. While this method has been successfully employed in a variety of fields such as financial trading (Preis, Moat, & Eugene Stanley, 2013) and public health (Thornton, Handley, Kay-Lambkin, & Baker, 2017), its systematic application to the study of relationships between cities has only just started to develop (e.g. Hu, Ye, & Shaw, 2017), following some initial small-scale explorations of the potential of this method (e.g. Devriendt, Derudder, & Witlox, 2008; Liu, Wang, Kang, Gao, & Lu, 2014; Janc, 2015).

The objective of this paper is to apply the toponym co-occurrence method to identify the pattern of relations between places in a systematic way. The empirical focus will be on the Dutch settlement system. Besides interpreting the results, we primarily focus on an evaluation of the applicability and feasibility of the systematic application of this method to identify and categorize inter-urban relationships. Therefore, we consider our application primarily as an experiment from which we can learn the preconditions for successful implementation, the potential drawbacks and the potential gains of applying the co-occurrence method to identify inter-city relationships.

The paper is structured as follows. First, we provide a brief overview of the different approaches to measuring relationships between cities, which culminates into a discussion of the first applications of the co-occurrence method (section 2). Second, we present our experiment, detailing the steps taken in the process (section 3). Third, we present and map the pattern of relations in the settlement system of the Netherlands (section 4). Finally, we conclude with a discussion of the pros and cons of the co-occurrence method and how this method can be successfully implemented in future studies (section 5).

2. Measuring relationships between cities

2.1. Overview of methods

Data availability has always played a crucial role in the development of the systems of cities research. Population data was the main source of information on urbanization at the national and regional scale during the first boom of this literature in the 1960s and 1970s. Inspired by early contributors such as Auerbach (1913) and Zipf (1949), researchers were using the rank size rule as a proxy to assess the intensity of relations within a system of cities. The underlying assumption was that if the settlement system in a country or region followed a clear rank-size distribution it would be characterized by a high degree of interdependence while the presence of a primate city would reflect a low level of integration (Vapnarsky, 1969). After this initial focus, the literature was soon enriched by studies focusing on migration of people (Simmons, 1979) and data on information circulation and the diffusion of innovation between cities (Pred, 1977, 1980).

More generally, there are two main types of data used in studies on relationships between cities: ‘stock’ data and ‘relational’ data. Stock data refers to information available for each city in the system. This data is useful for comparing cities and analyzing trends within the system of cities. Looking at the employment data of French urban agglomerations over 40 years, Paulus (2004) highlighted processes of co-evolution of cities through a process of spatial diffusion of innovation in the system of cities. Stock data is also used to evaluate to what extent some urban characteristics change with size within a system, which is referred to as ‘scaling laws’, an approach that has been widely used in the past 10 years (Bettencourt, Lobo, Helbing, Kühnert, & West, 2007; Pumain, Paulus, Vacchiani-Marcuzzo, & Lobo, 2006). The most widespread model to measure intercity relations in the last decade has been the ‘Interlocking Network Model’ (INM; Taylor, 2001). This approach is also based on stock data – the presence of advanced producer services (APS) firms in cities – but derives relational information from their location patterns. This method draws an analogy between the corporate organization of firms and intercity relationships. The INM model defines two cities as linked in a network to the extent that they host offices of the same APS firm. The assumptions underlying the INM method have not remained uncontested (Liu & Derudder, 2013; Neal, 2012, 2013b; Nordlund, 2004), and it is not well capable of measuring relationships between (smaller) cities on the regional scale (Lambregts, 2009; Burger, Meijers, et al., 2014). Other data allows studying intercity firms relations with actual relational data on ownership relations between headquarters and subsidiaries of multinational enterprises (Rozenblat, Zaidi, & Bellwald, 2017).

Relational data gives information on actual flows and links between cities and can be obtained from very diverse sources. Transportation data is a great source of information on relationships between cities. It can be obtained by looking at the infrastructure such as a railway, roads or postal road network (Bretagnolle & Franc, 2017; Derudder, Liu, Kunaka, & Roberts, 2014), by looking at the moves of vehicles such as ships (Ducruet, Cuyala, & El Hosni, 2018) or by looking at actual traffic, which covers both goods and people. Numerous studies have looked at flows of people to measure intercity relations at the regional, national or global scales, whether it is air passengers (Derudder & Witlox, 2005), train passengers (Berroir et al., 2017) or commuters (Nelson & Rae, 2016). Recently, flows of people have also been identified through geolocated posts of people on social media (Lenormand

et al., 2015; Zhang, Derudder, Wang, Shen, & Witlox, 2016), which allows to overcome the national dimension of data collection, but is not necessarily without representative bias. Another interesting source of relational are mails and telephone calls (Krings, Calabrese, Ratti, & Blondel, 2009; Zipf, 1946).

Nowadays, the combination of several sources of information to study the different networks and flows connecting cities and their mutual interdependencies are increasingly popular (Berroir et al., 2017; Burger, Meijers, et al., 2014; Choi, Barnett, & Chon, 2006; Ducruet, Ietri, & Rozenblat, 2011), as are approaches that employ ‘big data’, some of which were discussed in the introduction. The toponym co-occurrence method that takes centre stage in this paper has also developed from an initial manual exercise to an example of a big data approach to analyzing systems of cities.

2.2. Using co-occurrences to determine inter-city relationships

The co-occurrence of words in text corpora has long been considered a measure of relatedness. The very first application that we are aware of actually addresses urban systems. This seminal paper by Tobler and Wineburg (1971) explores the co-occurrence of 119 pre-Hittite towns on cuneiform tablets made almost 4000 years ago in Cappadocia to derive an approximation of how the towns were located relative to each other, basing themselves on the assumptions that ‘the mere mention of two town names on the same tablet is taken to define a relation between these towns’ (p. 40) and on what has become known as Tobler’s first law of geography, namely that ‘everything is related to everything else, but near things are more related than distant things.’

Co-occurrence analysis, sometimes referred to as co-word analysis, was taken to a higher level in the field of scientometrics (Callon, Courtial, Turner, & Bauin, 1983), where it is often used to measure relatedness, in this case identifying scientific fields and their development. The basic assumption still being that ‘the greater the probability of two elements co-occurring in the same article, the more strongly they are related’ (Chavalarias & Cointet, 2013, p. 2). So far, these ‘elements’ have included for instance organizations and firms (Vaughan & You, 2010); hyperlinks (Boulton, Devriendt, Brunn, Derudder, & Witlox, 2011; Salvini & Fabrikant, 2016) or even hashtags (Lorenz, Wolf, Braun, Djurdjevac Conrad, & Hövel, 2018) in addition to the key words characterizing scientific fields – see Peris, Meijers, and van Ham (forthcoming) for such a scientometric approach for the field of urban systems research. The increasing availability of crowd-sourced ‘big data’ and technological advances have provided an important impetus to the application of co-occurrence analysis. In particular web data has been considered suitable, because ‘[i]f two organizations are related, their names are likely to be mentioned together on Webpages’ (Vaughan & You, 2010, p. 483), making co-occurrence analysis also an important tool for Webometrics.

In addition to keywords, people, papers, hyperlinks, countries, organizations or hashtags, also place names, or toponyms, can be used. It has been estimated that about 70% of our online documents contain place references (Hill, 2006). In a similar way, we assume that the greater the frequency by which place names co-occur on Web pages (or in any other text corpora), the more they are related. This turns the toponym co-occurrence method into a novel method of identifying relationships between cities.

Several decades after Tobler and Wineburg's initial application, this potential has been re-established by a number of urban scholars. At a time when cyberplace approaches (focusing on physical digital infrastructure) were still dominant, Devriendt et al. (2008) provided a first cyberspace (focusing on virtual connections) approach directed at the content of websites to study inter-city relationships. They queried Google and AltaVista to develop a 40×40 matrix of co-occurrences on web pages of a small sample of 40 large European cities. Liu et al. (2014) perform a similar analysis to detect relatedness between Chinese provinces, focusing on Chinese public media reports accessed through Baidu. In a similar vein, Janc (2015) queried Google News to study the Polish urban system. Also basing themselves on the news, but just from a single source, Zhong et al. (2014) develop what they call a 'toponym co-occurrence network', which moves beyond the co-occurrence of geographic entity names in single documents to build a network of documents on the basis of the appearance of a single toponym in a set of documents. This way it accounts for indirect relationships: if city A is being mentioned together with city B in a document, and city B is mentioned in a document in which also city C is mentioned, then an indirect link is identified between cities A and C. While this allows identifying clusters of cities that are often mentioned together and consequently apply the toolkit of network analysis, it is hard if not impossible to conceptualize the exact nature of an indirect relatedness, such as between cities A and C.

This is probably why most previous work in the field has focused on direct relations between city pairs. Salvini and Fabrikant (2016), extracting co-occurrences through Wikipedia pages that link to two or more Wikipedia city pages, do not just focus on frequencies of these co-occurrences, but also label relations between cities according to the article categories in which they appeared, finding evidence for what Burger, Meijers, et al. (2014) term 'multiplexity': the fact that relations between places vary according to the type of flows or network studied. Hu et al. (2017) take this one step further by applying natural language processing to the texts of news articles rather than relying on classifications by users. The size of the datasets of these recent contributions has expanded substantially compared to early (often manual) approaches. For instance, Hu et al. (2017) exploit the archive of the Guardian newspaper, retrieving a quarter of a million news articles with co-occurrences of the place names of the 100 largest U.S. cities.

Here, we adopt a somewhat similar approach, focusing on co-occurrences of Dutch place names to trace the relatedness between places, and hence to obtain an image of the spatial organization of the Netherlands, and we also try to move beyond simple frequencies in an attempt to categorize relationships employing machine learning techniques. Instead of a focus on newspaper articles from a single source, we use the gigantic archive of websites known as the CommonCrawl to avoid selection bias. We believe that the Web archive provides a less biased data source than websites queried through a particular search engine like google. In addition, we innovate by a focus on both large and small places, essentially including all place names. Exploring whether this leads to relevant and valid results is of importance since reliable existing data on relationships of smaller places hardly exists, and it is precisely in this respect that the co-occurrence method potentially has unique advantages.

3. Research approach

3.1. Geographical focus

We decided to employ the co-occurrence method to explore the settlement system of the Netherlands, one of the reasons simply being familiarity with this country, which we believe is essential in this experimental phase to also tentatively judge the findings. However, the focus on the Netherlands allows to study not just relations between larger cities, which was the focus of the small number of previous studies employing this method, but also to explore the suitability of this approach to study relationships of smaller places. With Janc (2015), we believe that an important merit of the co-occurrence method is exactly the easy inclusion of smaller cities for which reliable sampled data is hard to find.

Our list of cities includes all places with over 750 inhabitants ($N = 1639$).² ‘Place’ can refer to a village, town or city and their immediate rural surroundings (which carries the name of the place in their postal address). For this reason, we do not refer to the ‘urban system’, but rather use ‘settlement system’ below, unless we analyse a subset of just larger places. The entire territory of the Netherlands is assigned to a place. Strict urban planning policies have generally prevented the coalescing of places into larger, contiguous built-up areas, making the places studied spatially distinct and meaningful entities from the cultural, economic and social point of view.

3.2. Data

The World Wide Web or internet has become a very important source of knowledge, and this knowledge tends to be accessed through using search engines such as Google or Bing. The co-occurrence method rests in particular on the counts of co-occurrences of places in text corpora such as texts on websites. Most previous applications of the co-occurrence method have used the Google search engine, entering two place names as search query. However, the counts of results returned are ambiguous at best, since they vary according to the computer one is using, and vary according to the country one is based in and the copy of google being used (Google has multiple copies running and queries will be dispatched to the copy that is least busy), while results also tend to be personalized based on previous search queries (see Janc, 2015, for a discussion of some of these). What is more, the number of results returned is an estimate, not an actual number of pages one can actually click on, which turns out to be far less if one tries. Other have used Wikipedia as source, which may also suffer from potential biases, such as the fact that Wikipedia authors are not representative for the larger society and structural determinism (Neal, 2012) looms (Salvini & Fabrikant, 2016).

Given the difficulties inherent to using search engine results, Wikipedia or a single source of news, we decided to use the Common Crawl as a data source.³ This is an archive of Webpages. Their corpus contains petabytes of raw web page data, extracted metadata and text extractions crawled together over the last 7 years. It essentially provides a snapshot of the web, thus including anything from blogs and personal websites, to informational news sites, e-commerce, community building and social media sites, commercial websites etc. Common Crawl data is freely available, gigantic in size and regularly updated (nowadays released on a monthly basis), making the database a popular source of

information in research (see e.g. Mühleisen & Bizer, 2012). Here, we use the March 2017 data. The Common Crawl data comes in three formats, of which the WET format is most useful for the co-occurrence method as it only contains extracted plain text.

Our focus on the Netherlands allows us to filter the dataset by only considering web pages with the .nl extension, which is the internet country code top-level domain name (and by far the most popular extension for websites in the Netherlands). Roughly 25 million pages out of the close to 3 billion pages available in Common Crawl were filtered out this way. Important to note is that searching for a top-level domain like .nl only includes the first page of every matching domain.

Another way to filter the dataset (which also brings the additional advantage of limiting the requirements for the speed and size of the data storage platform), is to only consider those pages that contain co-occurrences of place names. The obvious lower threshold is that two place names co-occur, but we set also a maximum threshold of 25. A substantial number of pages contain lists of cities, for instance to let users select their place of birth or their home address, although these hardly represent relationships between cities. The maximum of 25 was set after considering the graph below (Figure 1) and having inspected a sample of pages with 20–25 unique co-occurrences, concluding that these should generally be included. Building on Rasool, Tiwari, Singla, and Khare (2012) this filtering was implemented using the Aho-Corasick algorithm, which is a multi-pattern exact string matching algorithm, allowing to match a list of places against the text on a web page.

3.3. The problem of false positives and underestimation when using place names

The frequency of place names in the data may be overestimated due to a number of complications. Below, we list these potential biases, and present our way of solving these.

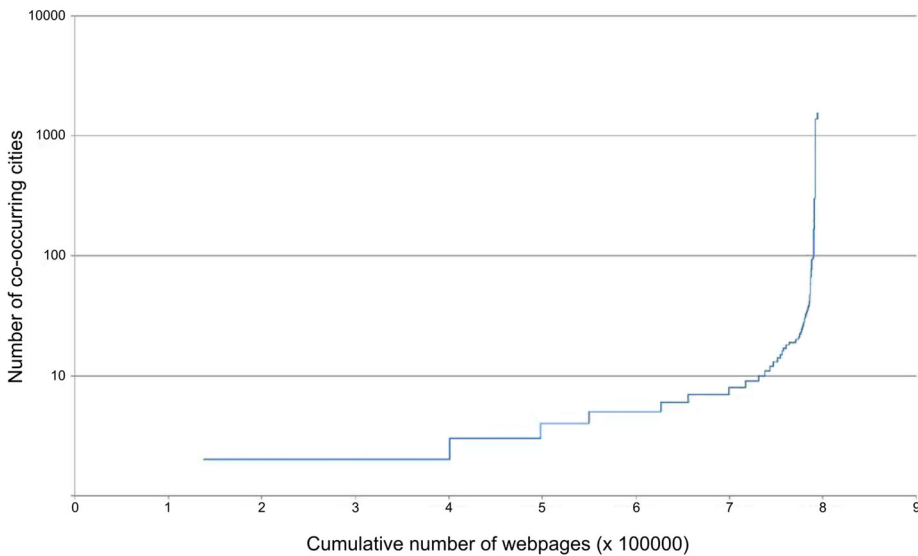


Figure 1. Number of webpages plotted against the number of unique occurrences contained in these pages. Source: Brunner, Mališ, Reichert, and van Agtmaal (2017).

- A place name may have multiple meanings. Many place names are very specific, but a small number of place names also act as nouns (e.g. ‘Assen’ = axles, ‘Hoorn’ = horn, ‘Huizen’ = houses, ‘Houten’ = made of wood) or verbs (e.g. ‘Kampen’ = fight). We have limited this problem by searching only for place names insofar they are written with a capital letter, while these nouns and verbs generally are not (unless appearing at the start of a sentence, but this is not so common in the Dutch language).
- Different places may have the same place name. Some place names occur twice. In our dataset there are about 50 such instances (with ‘Hengelo’, ‘Bergen’, ‘Beek’, ‘Elst’, ‘Heusden’ and ‘Zevenhuizen’ denoting the largest places). Some place names occur three times (e.g. ‘Rijswijk’) and one four times (‘Alteveer’ is the only case) in the Netherlands. Although the vast majority of these doubles concern small hamlets, which are dropped because of our minimum threshold, there still is an overestimation of network embeddedness of the remaining places when this occurs. A particular case is when place names also crop up in other countries – a situation that particularly occurs in formerly colonized territories. In the case of the Netherlands, some place names reappear in Surinam and South Africa.
- Places may also lend their name to the territories of which they are part. It could be that a text refers for instance to the province of Groningen rather than to the city carrying the same name. In the Dutch context, this also is the case for Utrecht, which acts as the capital of the province of Utrecht. In addition, there is a place called Zeeland, which is also the name of a province (of which it is not part).
- Place names also regularly feature in family names. This also leads to overestimation, depending on how often place names occur in family names. Out of the 100 most common family names, there are (parts of) five family names containing also a place name (‘Vries’, ‘Veen’, ‘Beek’, ‘Dongen’ and ‘Doorn’), so this problem should not be exaggerated, but it nevertheless is another source of overestimation if a family name is widespread. In addition, place names may also be used as first name: Brunn, Devriendt, Boulton, Derudder, and Witlox (2010) warn for the ‘Paris Hilton’ effect in this respect. However, none of the Dutch place names tends to be used as a first name.
- Place names may be carried by organizations, institutions or firms. Yet, this generally means that these actors are tied to that place, so this should be a limited problem: e.g. F.C. Utrecht refers to a professional soccer club, but it is associated with Utrecht.
- A place name could be part of another word. For instance the city of ‘Leiden’ could be part of the verb ‘Leidende’. To solve this, we added an additional check in the co-occurrence filtering that skipped words that contain city names.
- Place names in one country may also have a meaning in a different language. This would be particularly problematic in the case a Dutch place name also acts as a word in English, given the penetration of the latter language in the Dutch society. Even though we only focus on websites with the .nl extension, English language texts can often be found on Dutch websites. Examples include the Dutch place names ‘Born’, ‘Son’, ‘Made’, ‘Well’, ‘Thorn’ and ‘Hall’.

While the situations above would normally lead to overestimation of the relatedness of the places concerned, there are also three situations in which there could be underestimation:

- Place names sometimes change. This is a particular problem when doing longitudinal research that goes back to previous centuries. However, this does not need solving here, since our study is not longitudinal.
- Places may be referred to with multiple names (synonyms). Sometimes this is related to place names that changed, in which the older and new names are used simultaneously. Two examples in particular come to mind: The Hague (Den Haag in Dutch) is also (but increasingly less) referred to with the more formal, older ‘s-Gravenhage’ and the same applies to Den Bosch, but its official place name still is ‘s-Hertogenbosch’.
- Places known by multiple names due to the presence of multiple official (regional) languages. A particular subset of synonyms is due to multiple official languages being present in an area. In the Netherlands, this applies to the province of Friesland, where the Frisian language is an official second language; place name signs here tend to be bilingual.

Sometimes combinations of some biases occur, for instance when a synonym for one place (‘Alphen’ for Alphen aan den Rijn) happens to be also the name of two other places. Similarly, hardly anyone refers to what is officially ‘Amsterdam Zuidoost’ (population of over 81k), which essentially is a neighbourhood of Amsterdam (and referred to as such). In the end, over 85% of place names are truly unique and unbiased (see Table 1). As far as we could not yet deal with these potential biases, we will control for them by including dummies for each type of bias in our statistic evaluation of the results.

Our ambition here is not to solve disambiguation, but rather to assess to what extent this disambiguation hampers the potential of toponym co-occurrences to retrieve the relatedness of cities. In addition to our inspection of the list of place names in the Netherlands (checking for multiple occurrences of similar place names, place names that have a meaning in a different language that often surfaces in the Netherlands, place names that also refer to different geographical entities, and whether place names also appear in the top 100 most common family names), we will identify problematic cases also through employing the gravity model and exploring whether the extreme outlying cases can be attributed to the potential problems with place names above.

3.4. Classification of co-occurrences

The filtered dataset allowed counting the co-occurrence of place names, but an attempt was made to also classify co-occurrences according to the type of relationship or flow between places. Given the number of web-pages with co-occurrences, we used machine learning to classify relationships between cities. Traditional travel surveys tend to

Table 1. Potentially biased place names.

Source of bias	Frequency	Percentage ^a
Multiple meanings place name	62	3.8%
Multiple places with same name	46	2.8%
Place names occurring in common family names	6	0.4%
Place name part of English vocabulary	16	1%
Synonyms for same place	6	0.4%
Place names spelled different in Frysian language	99	6.1%
Unbiased place names	1404	85.7%

^aRelative to 1616 different place names (‘unbiased place names’ relative to 1639 places).

distinguish between different travel motives such as ‘commuting’, ‘education’, ‘leisure’, ‘shopping’ etc., so we decided to explore whether it would be possible to classify relationships between places according to similar motives, based on the textual context in which the co-occurrence of place names appears. That means that we employ a so-called supervised algorithm, which requires an input set and a corresponding output set, with which a model is trained to predict the classification of web pages that have not been seen or classified by humans. To train this algorithm, we used labelled data to train the classifier. Several options were considered (e.g. newspaper articles tagged with keywords that correspond to the motives for travel) but in the end we relied on the open data repository of Netherlands Statistics (CBS), who have tagged articles on their websites in a professional way and, not unimportantly, these cover the different travel motives we intend to study – also because they are the source of the more traditional studies into travel behaviour in the Netherlands. In implementing the machine learning algorithm, several steps and decisions were taken. First, the documents were cleaned by getting rid of common, unspecific words like articles (‘de’, ‘het’, ‘een’ in Dutch) and symbols, using NLTK (Bird, Klein, & Loper, 2017). Second, we used ‘Term Frequency over Inverse Document Frequency’ (TF-IDF) to give more weight to words based on their frequency in a document relative to the frequency of these words in the complete document set. With over 65,000 words in the document set, we narrowed down the number of features to the top 10% of words that have the highest TF-IDF weights. Such a dimensionality reduction is needed to prevent a slow process and diminishes over-fitting problems (Sebastiani, 2002), while Yang and Pedersen (1997) have stated that a dimensionality reduction by a factor 10 using this approach does not lead to a loss of accuracy. Even with 6500+ features, we need a machine learning algorithm that works well with feature rich problems, which is why the ‘Support Vector Machines’ (SVM) algorithm was chosen.

4. Results of the co-occurrence method

4.1. Overall pattern of co-occurrences

In this section, we will both visualize our results with maps, as well as explore the reliability of using co-occurrences to measure relationships between cities. For the latter, we compare the pattern of co-occurrences found with the pattern we would expect according to the gravity model. However, this does not mean that we suggest that our data should necessarily obey the rules of gravity, since in particular the role of distance in ‘cyberspace’ can be discussed, as well as whether the digital space formed by websites and ‘real space’ are identical. For instance, Liu et al. (2014, p. 100) found that ‘movements in geographical space experience a stronger distance decay effect than the information flow on the web’. As we interpret toponym co-occurrences on web pages to be a reflection of real interaction patterns on the ground, we will use the gravity model to calibrate our method (see Lenormand, Bassolas, & Ramasco, 2016), detect outliers that may be caused by place name disambiguation, and to move beyond a simple visualization of the strongest flows on maps to indicate to what extent a relationship between places is stronger or weaker than expected (based on the residuals of the gravity model).

Out of the 1,342,341 pairs of places in the Netherlands, 515,658 co-occur at least once (38.4%). Our previous choice to only store web pages with co-occurrences implies that

pairs of places without co-occurrences are not in our database, and these missing zeroes mean that the implementation of a gravity model is biased by not taking these into account. Therefore, we also limit the set of place names to the 100 largest places, which happens to coincide with the threshold above which all places have co-occurrences with the other places. In addition, we will also run analyses for places with 10,000 people and over, in order to be able to compare the applicability of the toponym co-occurrence method to places with different sizes. Table 2 presents the results of two types of models, namely the baseline gravity model (models 1, 3, 5) and the extension of this model with dummies that capture place name ambiguity (models 2, 4, 6).⁴

Place name disambiguation is a problem that needs to be dealt with when applying the toponym co-occurrence method; the accuracy of the gravity model is substantially improved when the dummies capturing the various types of place name disambiguation problems are included, leading to substantially improved fits of the model (compare Adjusted R^2 values). Most prominent problem, at least in the Netherlands, is the fact that multiple places may have the same name, followed by bias caused by place names having a meaning in the English language and the fact that place names can have multiple meanings in Dutch (model 2). The signs of the coefficients are generally as expected, although some differences between the models can be seen. The use of multiple synonyms for one place was expected to lead to underestimation of co-occurrences, but this is only true for larger places. The fact that place names are written differently in the Netherlands' second language (Frisian) was expected to cause underestimation, but the opposite is true, which suggests that those places in the province of Friesland are actually more related than

Table 2. Gravity model, place name disambiguation and toponym co-occurrences (dependent: Ln Total co-occurrences).

	(1) Places > 750	(2) Places > 750	(3) Places > 10,000	(4) Places > 10,000	(5) Places > 31,500	(6) Places > 31,500
Intercept	-4.735 (.020)**	-5.191 (.019)**	-16.281 (.105)**	-17.451 (.099)**	-22.387 (.332)**	-23.885 (.289)**
Pop. A (ln)	.421 (.002)**	.440 (.001)**	1.110 (.007)**	1.171 (.007)**	1.391 (.023)**	1.522 (.020)**
Pop. B (ln)	.567 (.002)**	.589 (.002)**	1.060 (.009)**	1.104 (.008)**	1.266 (.022)**	1.289 (.018)**
Distance (ln)	-.516 (.002)**	-.550 (.002)**	-.515 (.008)**	-.540 (.007)**	-.305 (.019)**	-.376 (.016)**
Place name with multiple meanings		.772 (.006)**		.711 (.016)**		.709 (.030)**
Place name part of English vocabulary		.919 (.010)**		.970 (.031)**		n.a.
Place name occurs frequently as family name		.670 (.015)**		.755 (.036)**		n.a.
Multiple places with same name		1.193 (.005)**		.836 (.017)**		.308 (.038)**
Synonyms for place name		.190 (.014)**		-1.166 (.032)**		-1.388 (.043)**
Frisian/Dutch place name different		.107 (.006)**		.368 (.021)**		.472 (.057)**
<i>N</i> city pairs	515,658	515,658	47,533	47,533	4,950	4,950
<i>N</i> places	1,639	1,639	319	319	100	100
Adjusted R^2	.332	.426	.555	.623	.648	.747
F	85308.069**	42545.800**	19737.941**	8731.034**	3036.960**	2091.307**

** $p < 0.01$.

others. The other dummies for place name disambiguation are invariably causing overestimation.

Interestingly, the fit of the gravity model with the co-occurrences found increases with population size. The size of the places and the distance between them explains almost two-thirds of the variety in co-occurrences found for the largest 100 places in the Netherlands (model 5), versus just one-third when taking all 1,639 places into account (model 1). Part of the explanation is that the dataset for the 100 largest cities does not contain any 'zeroes' (non-existing co-occurrences between places).

This may also partly explain the decreased importance of the role of distance when comparing the results for the 100 largest Dutch places to the results for datasets containing smaller places. A 1% increase in distance, diminishes the number of co-occurrences with 0.38% (model 6), whereas for the other datasets this elasticity is -0.52% . The standardized Beta coefficients of model 6 (not reported) suggest that both population variables are about three times more important in explaining co-occurrences than distance.

4.2. The spatial organization of the Netherlands

While Figure 2 presents the pattern of absolute flows between the 100 largest places in the Netherlands, Figure 3 provides a normative interpretation of these flows by indicating whether they are stronger or weaker than expected given the gravity model and place name disambiguation (using the standardized residuals of model 6).

As could be expected, the strongest relationships in absolute terms between places in the Netherlands can be found in the Randstad region where the country's four largest cities (Amsterdam, Rotterdam, The Hague and Utrecht) form the anchors of a polycentric urban region (Figure 2). Quite strongly connected to this region are places like Eindhoven, Breda and Arnhem, forming a kind of larger urban field in the central area of the Netherlands. Outside that area, the more distant city of Groningen stands out as being strongly

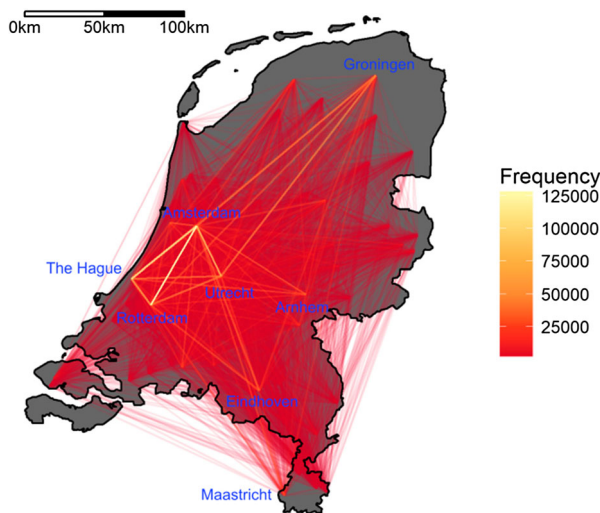


Figure 2. Observed spatial organization of the Netherlands based on the pattern of toponym co-occurrences.

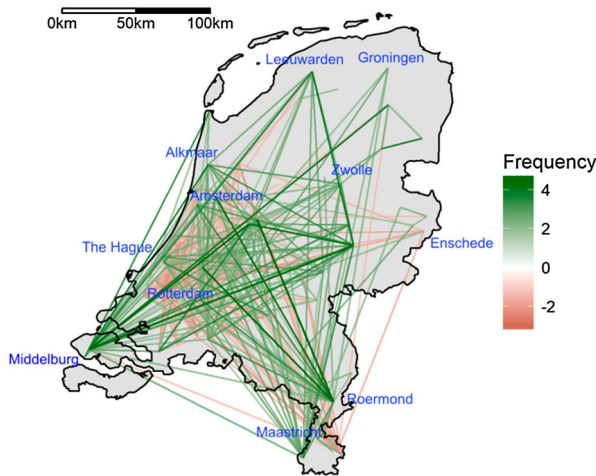


Figure 3. Observed versus expected relations between Dutch places, based on toponym co-occurrences.

Note: For clarity of the visualization, only standardized residuals ≥ 2 and ≤ -1 are displayed.

related to the main Randstad cities. However, the comparison of Figures 2 and 3 is of interest. Whereas the relation between Rotterdam and Amsterdam is the strongest in absolute terms, it happens to be somewhat less strong than expected (-2.8% to be precise). Within the Randstad region, The Hague stands out as a city that is more related to the other main Randstad cities (The Hague – Amsterdam: $+10\%$; The Hague – Rotterdam: $+7\%$; The Hague – Utrecht: $+10\%$). The relations Amsterdam-Utrecht (-3%) and Utrecht-Rotterdam (-4%) are less strong than expected. More generally, the Randstad area does not turn out to be more strongly related than expected. Rather, longer distance relations among cities in the periphery and between them and the seat of national government The Hague stand out, although there are also some peripheral cities that are clearly less well related.

This can be further explored by calculating the sum of all unstandardized predicted and residual values and comparing these. Table 3 presents the 10 relatively most strongly related cities in the Netherlands, as well as those 10 that are least related (considering again only the 100 largest places in the Netherlands and leaving aside some names that suffer from place name disambiguation). These figures were calculated by aggregating all unstandardized predicted and residual values and comparing these. Those that are

Table 3. Places that are relatively more strongly and more weakly related to other places.

Relatively more related places	%	Relatively less related places	%
Roermond	20.11	Capelle aan den IJssel	-15.62
Middelburg	16.71	Spijkenisse	-14.10
Zutphen	13.32	IJsselstein	-10.74
Maastricht	12.01	Landgraaf	-10.72
Zwolle	10.40	Hellevoetsluis	-9.70
Hoogeveen	10.00	Vlaardingen	-9.65
Gorinchem	9.82	Zwijndrecht	-9.11
Wageningen	9.75	Almere	-8.92
Vlissingen	9.34	Etten-Leur	-8.80
Alkmaar	8.84	Kerkrade	-8.40

more related tend to be historically important cities located in the periphery, whereas those that are less related to other cities than expected tend to be either relatively new, suburban places near the main Randstad cities (Capelle aan den IJssel, Spijkenisse, IJsselstein, Hellevoetsluis, Almere), or older places that have always been in the ‘agglomeration shadow’ (see Meijers & Burger, 2017) of a larger close-by city (Vlaardingen near Rotterdam, Zwijndrecht next to Dordrecht, Etten-Leur next to Breda) or former mining towns (Landgraaf, Kerkrade).

One of the potentials of the co-occurrence method is that it can also be applied to very small places. Therefore, we map a rural province in the southwestern delta area of the Netherlands (Zeeland). Again, we show absolute flows (Figure 4) and relative flows (Figure 5).

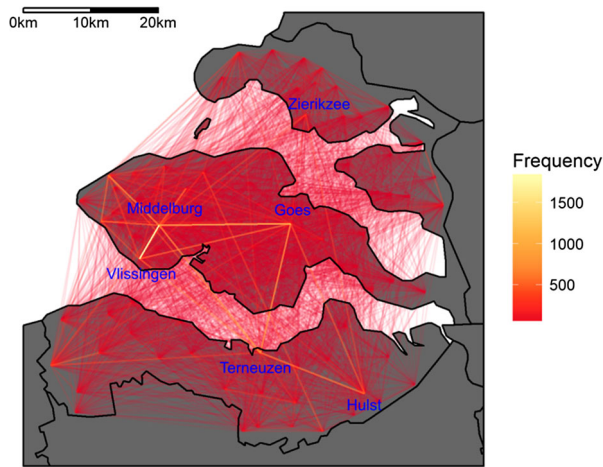


Figure 4. Observed spatial organization of Zeeland based on the pattern of toponym co-occurrences.

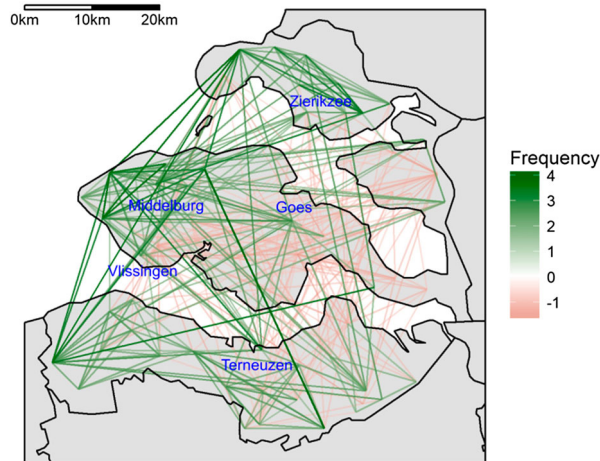


Figure 5. Observed versus expected relations between places in Zeeland, based on toponym co-occurrences.

Note: Given the absence of ‘zeroes’ in our data and the problem of overdispersion when applying the nowadays increasingly used Poisson regression, we opt for conventional OLS which also leads to better model fits.

Many of the villages in Zeeland count just about 1,000 inhabitants, but the toponym co-occurrence method also appears to deliver relevant and accurate information on relationships between these small places in the sense that the patterns could be tentatively expected and logically explained. The region appears rather well integrated, with a dominance of relationships that are stronger over relationships that are weaker. [Figure 5](#) also seems to show that the more touristic places along the coast are more related among each other than the more agrarian villages to the east of the province. We also chose this province of Zeeland because sea arms clearly divide the region, and we would expect that these hamper the development of relations between places on both sides of the different estuaries. Even though relationships with places located on the same peninsula seem stronger, we also see quite some well-established relations with places in other parts of the province. The west-east divide seems more prominent, which could be explained by the fact that places in the eastern part are perhaps more oriented to cities in the neighbouring province Noord-Brabant.

4.3. *Classifying co-occurrences*

The spatial organization of a territory differs according to which type of relationship or flow is being taken into account ('multiplexity'), and even the pattern for a particular type of flow differs for different types of persons ('individual-level heterogeneity'; see Burger, Meijers, et al., 2014). To account for the former we applied machine learning to interpret relationships, using a supervised algorithm to apply pre-defined categories that are common types of flows (commuting; shopping; leisure; education, collaboration, transportation). For each page our trained classifier estimates the probabilities of a document belonging to each available category. Depending on these probabilities, we can decide which type of flow is assigned to the webpage in question. Using different thresholds leads to different results. [Table 4](#) presents the number of city pair relationships categorized into a particular category. The number of relationships identified is substantially lower when applying a probability level of 0.75, which should however be judged superior over the lower probability threshold of 0.25. Again, we use the gravity model to calibrate and judge the results obtained.

Out of 515,658 co-occurrences, our trained classifier managed to label between 11% (commuting) and 61% (collaboration) using the lower probability threshold (0.25). Using this threshold regularly implies that webpages are classified as reflecting multiple types of flows. It is hard to believe that 61% of the co-occurrences do indeed reflect cooperative relationships, so the stricter probability threshold of 0.75 appears better. However, this threshold implies that 0.22% of all co-occurrences are categorized as 'shopping', up to 6.8% for 'education'. On average, 1.75% of all co-occurrences are categorized, which seems a low number. Yet, the gravity model (in its basic form) is significant for all types of flows at both probability levels. Note that distance is not significant at the .75 probability level. Remarkable is also that population has a negative coefficient for shopping, but this pattern is not well captured by the gravity model (adjusted R^2 is just .019). An explanation could be the rise of online shopping that seems not much hampered by geographical distances or a limited urban mass. All in all, the low number of city pairs classified at this desired probability level and the limited fit of the model for especially commuting, shopping and leisure flows is somewhat disappointing. The classification method shows promise, but needs to be improved to be truly useful.

Table 4. Classified flows between places versus the gravity model.

	Commuting	Shopping	Leisure	Education	Collaboration	Transportation
Probability level 0.25						
Significant factors	Pop A** (+), Pop B** (+), Distance** (-)	Pop A** (+), Pop B** (+), Distance** (-)	Pop A** (+), Pop B** (+), Distance** (-)	Pop A** (+), Pop B** (+), Distance** (-)	Pop A** (+), Pop B** (+), Distance** (-)	Pop A** (+), Pop B** (+), Distance** (-)
<i>N</i> (city pairs)	56411	180570	213073	104328	313445	63376
F	6380.198**	14743.782**	25398.494**	10553.011**	48185.436**	5933.531**
Adjusted <i>R</i> ²	.253	.197	.248	.233	.316	.219
Probability level 0.75						
Significant factors	Pop A** (+), Pop B** (+)	Pop A** (-)	Pop A** (+), Pop B** (+)	Pop A** (+), Pop B** (+)	Pop A** (+), Pop B** (+)	Pop A** (+), Pop B** (+)
<i>N</i> (city pairs)	2501	1135	9987	34826	11671	3655
F	92.596**	8.140**	159.694**	3020.210**	821.088**	151.846**
Adjusted <i>R</i> ²	.099	.019	.046	.206	.174	.110

***p* < 0.01. OLS regression. All variables have been log-transformed.

5. Conclusion

This paper further pioneered the toponym co-occurrence method to establish relationships between places. This method captures relationships between places in digital space. The widely accepted gravity model has often shown a good fit with relationships in real, physical space. Since the gravity model also fits well with our results, we believe that the co-occurrence method is a good proxy for relationships between places in the real world, and as such allows to construct the spatial organization of a territory. Next to information on the strength of relationships obtained through the frequency of co-occurrences, it also delivers a classification of these relationships. In this paper, we applied this method to a so far unseen amount of data, namely the billions of pages available in the not for profit web archive CommonCrawl, which stores websites from all over the world and as such provides a snapshot of the Web at a particular moment in time. In addition, we applied machine learning techniques to the Web texts containing place name co-occurrences, in order to classify the type of relationships. Whereas previous contributions have all focused on detecting networks between large cities, we applied the method to the entire settlement system of the Netherlands, including all settlements of 750 people and over. Several sources of place name disambiguation were identified and dealt with in applying our method.

In fact, the applicability of the method to places of any size makes the toponym co-occurrence method suitable for many types of analyses, e.g. novel ways of identifying functional urban areas, detecting infrastructural needs, or studying the importance of network embeddedness for development. However, if good quality detailed data on for instance commuting flows or transport flows is available, the results of the co-occurrence method should be considered a complement rather than a substitute. The method could, however, be of particular importance in situations where such data is lacking, and one of its strengths is the ability to carry out analyses on supranational level (e.g. Europe) following a single, uniform and harmonized method.

Our analyses show that the strongest connections may be with nearby places, but that longer distance relationships between places also frequently exist, and are often stronger than expected. Given our focus on applying and evaluating this novel toponym co-occurrence method, our analysis of the spatial organization of the Netherlands was reasonably limited, but nevertheless showed for instance that the coherence in the Randstad region was less strong than expected, even though it is by many considered to be a single metropolitan entity. It also put forward several suggestions why some places are strongly or weakly positioned in networks of relationships. This obviously demands further research.

The toponym co-occurrence method is widely applicable to many types of ('big') data, basically any archive with textual data lends itself. The accuracy of the results of the method, however, is also much determined by the quality of the underlying data. While we used a gigantic Web archive and considered this source better than using the strongly varying results of a search engine like google (see also Devriendt et al., 2008; Hu et al., 2017) or a single source of information like an individual newspaper archive, we are at the same time aware that the web contains a substantial amount of 'noise'. In training our classifier, it was often not possible to give a particular label to texts on website, or at least not one that was related to a type of flow. Something that requires checking is whether pages mentioning larger cities contain more noise than pages mentioning

smaller places, potentially causing overestimation. In addition, Web pages relating to for instance 'leisure' are much more abundant than pages where people report about their daily commute. This particularly has consequences for the interpretation of different types of flows, in that the patterns can be compared, but not necessarily the strengths of relationships. The classification exercise in this paper delivered reasonable, but not yet satisfying results. One way to improve this could be the adoption of an unsupervised classification algorithm, rather than departing from a number of pre-defined categories of flows as we did here, which would allow to categorize more webpages than the algorithm was able to do now. Alternatively, the classifier may need to be trained more extensively than we were able to do. An issue to take into account is that categorizing a website is not necessarily the same as categorizing the exact flow between places (see also Janc, 2015). For instance, a retail website listing the place names of shops of the same shoe selling firm will be labelled as 'shopping', but the flows between the locations of this firm are not shopping flows, but rather flows of information, goods and possibly people working for that firm. Following this, we believe that the main challenge in improving this method lies in the classification part, which requires the application of more sophisticated machine learning tools.

Perhaps the use of digital archives of multiple newspapers is a convenient way out too, since one can use the logical classification derived from the different sections and columns that newspapers generally use ('economy', 'sports' etc.), a potential that has been identified already by Hu et al. (2017), while Salvini and Fabrikant (2016) exploit a similar potential of Wikipedia. Possibly interesting in this regard is the specific news dataset of the Common-Crawl and the efforts to digitalize newspaper archives that are going on in many countries, the potential of which seems to have been predominantly identified by digital humanities researchers but not yet by social science scholars.

Another challenge is to solve place name disambiguation in a more automated way than we did here. Luckily, the issue of place name disambiguation is an important concern in (geographic) information retrieval and computational linguistics, and 'named entity recognition' procedures are becoming increasingly accurate and precise.

Despite these challenges still ahead, we are convinced that the toponym co-occurrence method could break new ground in studying urban systems and networks between places. After all, it is not accidental that the method was invented in this domain (Tobler & Wineburg, 1971), and, their analysis points us to what should be one of the most exciting possibilities of this method: a longitudinal analysis of the development of urban systems over time.

Notes

1. Hu et al. (2017) refer to this as 'semantic relatedness'.
2. This threshold was somewhat pragmatically chosen as smaller places tend to be less well identifiable as 'villages' proper, but may for instance be more a grouping of scattered buildings in hamlets, while place name disambiguation was also a greater concern with these smaller places.
3. commoncrawl.org
4. Given the absence of 'zeroes' in our data and the problem of overdispersion when applying the nowadays increasingly used Poisson regression, we opt for conventional OLS which also leads to better model fits.

Acknowledgement

The authors would like to thank Piet van Agtmaal, Tom Brunner, Marko Mališ and Gijs Reichert, all computer science students at TU Delft, for executing the technical part of the research, and Dr. Claudia Hauff of TU Delft's Data Science Centre for her co-mentoring. The student's technical report is available online (Brunner et al., 2017).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Netherlands Organisation for Scientific Research (NWO) under Grant number VIDI 452-14-004.

References

- Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59, 74–76.
- Batten, D. (1995). Network cities: Creative urban agglomerations for the 21st century. *Urban Studies*, 32, 313–327.
- Berroir, S., Cattan, N., Dobruszkes, F., Guérois, M., Paulus, F., & Vacchiani-Marcuzzo, C. (2017, February 6). Les systèmes urbains français: une approche relationnelle. *Cybergeog: European Journal of Geography*. doi:10.4000/cybergeog.27945
- Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104, 7301–7306.
- Bird, S., Klein, E., & Loper, E. (2017). *Natural Language Toolkit 3.2.5 documentation*. Retrieved from <http://www.nltk.org/api/nltk.stem.html>
- Boulton, A., Devriendt, L., Brunn, S., Derudder, B., & Witlox, F. (2011). City networks in cyberspace and time: Using google hyperlinks to measure global economic and environmental crises. In R. J. Firmino, F. Durate, & C. Ultramari (Eds.), *ICTs for mobile and ubiquitous urban infrastructures: Surveillance, locative media and global networks* (pp. 67–87). Hershey, PA: IGA Global.
- Bourne, L., & Simmons, J. (Eds.). (1978). *Systems of cities: Readings on structure, growth and policy*. New York: Oxford University Press.
- Bretagnolle, A., & Franc, A. (2017). Emergence of an integrated city-system in France (XVIIth–XIXth centuries): Evidence from toolset in graph theory. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50, 49–65.
- Brunn, S. D., Devriendt, L., Boulton, A., Derudder, B., & Witlox, F. (2010). Networks of European cities in worlds of global economic and environmental change. *Fennia*, 1881, 37–49.
- Brunner, T., Mališ, M., Reichert, G., & van Agtmaal, P. (2017). *UrbanSearch. Final report graduation project*. TU Delft. Delft: Delft University of Technology.
- Burger, M. J., Meijers, E. J., & van Oort, F. G. (2014). Multiple perspectives on functional coherence: Heterogeneity and multiplexity in the Randstad. *Tijdschrift voor economische en sociale geografie*, 105, 444–464.
- Burger, M., & Meijers, E. (2016). Agglomerations and the rise of urban network externalities. *Papers in Regional Science*, 95, 5–15.
- Burger, M., Thissen, M., van Oort, F., & Diodato, D. (2014). The magnitude and distance decay of trade in goods and services: New evidence for European countries. *Spatial Economic Analysis*, 9, 231–259.
- Callon, M., Courtial, L.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22, 191–235.

- Camagni, R. (2017). The city of business: The functional, the relational-cognitive and the hierarchical distributive approach. *Quality Innovation Prosperity*, 21, 31–48.
- Camagni, R., & Capello, R. (2004). The city network paradigm: Theory and empirical evidence. In R. Capello & P. Nijkamp (Eds.), *Urban dynamics and growth* (pp. 495–529). Amsterdam: Elsevier.
- Capello, R. (2000). The city network paradigm: Measuring urban network externalities. *Urban Studies*, 37, 1925–1945.
- Chavalarias, D., & Cointet, J. (2013). Phylomemetic patterns in science evolution—The rise and fall of scientific fields. *PLoS ONE*, 8, e54847.
- Choi, J. H., Barnett, G. A., & Chon, B.-S. (2006). Comparing world city networks: A network analysis of internet backbone and air transport intercity linkages. *Global Networks*, 6, 81–99.
- Derudder, B., Liu, X., Kunaka, C., & Roberts, M. (2014). The connectivity of South Asian cities in infrastructure network. *Journal of Maps*, 10, 47–52.
- Derudder, B., & Witlox, F. (2005). An appraisal of the use of airline data in assessing the world city network: A research note on data. *Urban Studies*, 42, 2371–2388.
- Devriendt, L., Derudder, B., & Witlox, F. (2008). Cyberplace and cyberspace: Two approaches to analyzing digital intercity linkages. *Journal of Urban Technology*, 15, 5–32.
- Ducruet, C., Cuyala, S., & El Hosni, A. (2018). Maritime networks as systems of cities: The long-term interdependencies between global shipping flows and urban development (1890–2010). *Journal of Transport Geography*, 66, 340–355. doi:10.1016/j.jtrangeo.2017.10.019
- Ducruet, C., Ietri, D., & Rozenblat, C. (2011). Cities in worldwide air and sea flows: A multiple networks analysis. *Cybergeog: European Journal of Geography*. doi:10.4000/cybergeog.23603
- Hill, L. L. (2006). *Georeferencing: The geographical associations of information*. Cambridge, MA: MIT Press.
- Hohenberg, P. M., & Lees, L. H. (1985). *The making of urban Europe, 1000–1950*. Cambridge, MA: Harvard University Press.
- Hu, Y., Ye, X., & Shaw, S.-L. (2017). Extracting and analyzing semantic relatedness between cities using news articles. *International Journal of Geographical Information Science*, 31, 2427–2451.
- Janc, K. (2012). Possibilities of hyperlink application in spatial research. *Bulletin of Geography*, 17, 57–65.
- Janc, K. (2015). Geography of hyperlinks—Spatial dimensions of local government websites. *European Planning Studies*, 23, 1019–1037.
- Kefler, C. (2017). Extracting central places from the link structure in wikipedia. *Transactions in GIS*, 21(3), 488–502.
- Krings, G., Calabrese, F., Ratti, C., & Blondel, V. D. (2009). Urban gravity: A model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009, L07003.
- Lambregts, B. (2009). *The polycentric metropolis unpacked. Concepts, trends and policy in the Randstad Holland*. Amsterdam: AMIDST.
- Lenormand, M., Bassolas, A., & Ramasco, J. J. (2016). Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51, 158–169.
- Lenormand, M., Gonçalves, B., Tugores, A., & Ramasco, J. J. (2015). Human diffusion and city influence. *Journal of The Royal Society Interface*, 12, 1–9.
- Limtanakool, N., Schwanen, T., & Dijst, M. (2009). Developments in the Dutch urban system on the basis of flows. *Regional Studies*, 43, 179–196.
- Liu, X., & Derudder, B. (2013). Analyzing urban networks through the lens of corporate networks: A critical review. *Cities*, 31, 430–437.
- Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014). Analyzing relatedness by toponym co-occurrences on webpages. *Transactions in GIS*, 18, 89–107.
- Lorenz, P., Wolf, F., Braun, J., Djurdjevac Conrad, N., & Hövel, P. (2018). Capturing the dynamics of hashtag-communities. In C. Cherifi, H. Cherifi, M. Karsai, & M. Musolesi (Eds.), *Complex networks & their applications VI. Complex networks 2017*. Studies in Computational Intelligence. (Vol. 689, pp. 401–413). Cham: Springer.

- McCann, P., & Acs, Z. J. (2011). Globalization: Countries, cities and multinationals. *Regional Studies*, 45, 17–32.
- van Meeteren, M., & Poorthuis, A. (2018). Christaller and “big data”: Recalibrating central place theory via the geoweb. *Urban Geography*, 39, 122–148.
- Meijers, E. (2007). From central place to network model: Theory and evidence of a paradigm change. *Tijdschrift voor Economische en Sociale Geografie*, 98, 245–259.
- Meijers, E., & Burger, M. (2017). Stretching the concept of ‘borrowed size’. *Urban Studies*, 54, 269–291.
- Meijers, E., Burger, M., & Hoogerbrugge, M. (2016). Borrowing size in networks of cities: City size, network connectivity and metropolitan functions in Europe. *Papers in Regional Science*, 95, 181–198.
- Meijers, E., Hoogerbrugge, M., & Cardoso, R. (2018). Beyond polycentricity: Does stronger integration between cities in polycentric urban regions improve performance? *Tijdschrift voor Economische en Sociale Geografie*, 109, 1–21.
- Mühleisen, H., & Bizer, C. (2012). *Web data commons - Extracting structured data from two large web corpora*. LDOW 2012, April 16, Lyon, France.
- Neal, Z. (2012). Structural determinism in the interlocking world city network. *Geographical Analysis*, 44, 162–170.
- Neal, Z. (2013b). Brute force and sorting processes: Two perspectives on world city network formation. *Urban Studies*, 50, 1277–1291.
- Neal, Z. P. (2013a). *The connected city: How networks are shaping the modern metropolis*. New York: Routledge.
- Nelson, G. D., & Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLOS ONE*, 11, e0166083.
- Nordlund, C. (2004). A critical comment on the Taylor approach for measuring world city interlock linkages. *Geographical Analysis*, 36, 290–296.
- Paulus, F. (2004). Coévolution dans les systèmes de villes : croissance et spécialisation des aires urbaines françaises de 1950 à 2000. Université Panthéon-Sorbonne - Paris I. Retrieved from <https://tel.archives-ouvertes.fr/tel-00008053/document>
- Peris, A., Meijers, E., & van Ham, M. (forthcoming). The evolution of the systems of cities literature: Schools of thought and their interaction. *Networks and Spatial Economics*. <https://doi.org/10.1007/s11067-018-9410-5>
- Pred, A. (1977). *City systems in advanced economies: Past growth, present processes, and future development options*. New York: Wiley, 256 p.
- Pred, A. (1980). *Urban-growth and city-systems in the United States, 1840–1860*. Cambridge, MA: Harvard University Press.
- Preis, T., Moat, H., & Eugene Stanley, H. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, 3, article number 1684.
- Pumain, D., Paulus, F., Vacchiani-Marcuzzo, C., & Lobo, J. (2006). An evolutionary theory for interpreting urban scaling laws. *Cybergeo: European Journal of Geography*. doi:10.4000/cybergeo.2519
- Rasool, A., Tiwari, A., Singla, G., & Khare, N. (2012). String matching methodologies: A comparative analysis. *International Journal of Computer Science and Information Technologies*, 3, 3394–3397.
- Rozenblat, C., Zaidi, F., & Bellwald, A. (2017). The multipolar regionalization of cities in multinational firms’ networks. *Global Networks*, 17, 171–194.
- Salvini, M., & Fabrikant, S. (2016). Spatialization of user-generated content to uncover the multi-relational world city network. *Environment and Planning B: Planning and Design*, 43, 228–248.
- Schiff, N. (2015). Cities and product variety: Evidence from restaurants. *Journal of Economic Geography*, 15, 1085–1123.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1–47.

- Short, J. R., Kim, Y., Kuu, M., & Wells, H. (1996). The dirty little secret of world cities research: Data problems in comparative analysis. *International Journal of Urban and Regional Research*, 20, 697–717.
- Simmons, J. W. (1979). *The Canadian urban system: An overview*. University of Toronto, 53 p.
- Taylor, P. J. (2001). Specification of the world city network. *Geographical Analysis*, 33, 181–194.
- Taylor, P. J., & Derudder, B. (2016). *World city network: A global urban analysis*. Abingdon, UK: Routledge.
- Thornton, L., Handley, T., Kay-Lambkin, F., & Baker, A. (2017). Is a person thinking about suicide likely to find help on the internet? An evaluation of Google search results. *Suicide and Life-Threatening Behavior*, 47, 48–53.
- Tobler, W., & Wineburg, S. (1971). A cappadocian speculation. *Nature*, 231(5297), 39–41.
- Vapnarsky, C. A. (1969). On rank-size distributions of cities: An ecological approach. *Economic Development and Cultural Change*, 17, 584–595.
- Vaughan, L., & You, J. (2010). Word co-occurrences on webpages as a measure of the relatedness of organizations: A new Webometrics concept. *Journal of Informetrics*, 4, 483–491.
- Xu, J., Li, A., Li, D., Liu, Y., Du, Y., Pei, T., ... Zhou, C. (2017). Difference of urban development in china from the perspective of passenger transport around spring festival. *Applied Geography*, 87, 85–96.
- Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, 97, 412–420.
- Yuan, Y., & Medel, M. (2016). Characterizing international travel behavior from geotagged photos: A case study of flickr. *PLoS ONE*, 11, e0154885.
- Zhang, W., Derudder, B., Wang, J., Shen, W. & Witlox F. (2016). Using location-based social media to chart the patterns of people moving between cities: The case of weibo-users in the Yangtze River Delta. *Journal of Urban Technology*, 23, 91–111.
- Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28, 2178–2199.
- Zipf, G. K. (1946). Some determinants of the circulation of information. *The American Journal of Psychology*, 59, 401–421.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley Press, 596 p.
- Zook, M., Devriendt, L., & Dodge, M. (2011). Cyberspatial proximity metrics: Reconceptualizing distance in the global urban system. *Journal of Urban Technology*, 18, 93–114.