Cities in interaction

Analysing the Dutch system of cities with computational methods

Peris, A.F.T.

DOI
10.7480/abe.2021.07

Publication date
2021

Document Version
Final published version

Citation (APA)
Peris, A. F. T. (2021). *Cities in interaction: Analysing the Dutch system of cities with computational methods*. [Dissertation (TU Delft), Delft University of Technology]. A+BE | Architecture and the Built Environment. https://doi.org/10.7480/abe.2021.07

Cities in interaction

Analysing the Dutch system of cities with computational methods

Antoine Peris

Cities in interaction

Analysing the Dutch system of cities with computational methods

Antoine Peris

TOC

A+BE | Architecture and the Built Environment | TU Delft BK

21#07

Design | Sirene Ontwerpers, Véro Crickx

Cover image | Alicia Ramon

ISBN 978-94-6366-400-4
ISSN 2212-3202

This dissertation is open access at https://doi.org/10.7480/abe.2021.07

TOC

Cities in interaction

Analysing the Dutch system of cities with computational methods

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Monday, 10 May 2021 at 12:30 hours

by

Antoine Ferdinand Théophile PERIS
Master Sciences Humaines et Sociales, à finalité Recherche,
Mention Géographie, spécialité Sciences des territoires-GEOPRISME,
Université Paris 1 Panthéon-Sorbonne, France
born in Prades, France

TOC

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. M. van Ham	Delft University of Technology, promotor
Dr. E.J. Meijers	Delft University of Technology, promotor

Independent members:

Prof. dr. B. Derudder	Katholieke Universiteit Leuven, Belgium
Prof. Dr. -Ing C.M Hein	Delft University of Technology
Prof. dr. F.G. van Oort	Erasmus University Rotterdam
Prof. dr. J.E. Stoter	Delft University of Technology
Dr. M. van Meeteren	Loughborough University, United Kingdom

The research in this thesis was done at the Faculty of Architecture and the Built Environment, Delft University of Technology. The project was funded by the Netherlands Organization for Scientific Research (NWO), grant number 452-14-004 (VIDI Project DISPERSAL: Beyond Agglomerations: Mapping Externality Fields and Network Externalities).

Acknowledgements

Doing PhD research sometimes feel like a lonely exercise, but looking back at these past years, I realise that it is actually not. For completing this dissertation, I have been supervised, helped, advised, motivated, inspired and supported by many people without whom none of these pages would exist.

The first person I would like to thank is Evert Meijers. I am very grateful to him for having offered me a PhD position at TU Delft, which has allowed me to work on a subject I am passionate about, and to travel to nice places for sharing my research and learn from others. I would like to thank him for sharing his passion for research, his enthusiasm, and for cheering me up during the downs and encouraging me during the ups. I am also very thankful to Maarten van Ham who also supervised me. His constant support has been determinant in this PhD process and I have learned a lot from him, especially when it comes to better frame my papers and to know when to slow down and play banjo.

Thank you also to all the people that contributed scientifically to this work through discussions, advice, or lines of code: Willem Jan Faber, Bijan Ranjbar-Sahraei, Clémentine Cottineau, Romain Leconte, developers from the open-source geospatial community, creators of tutorials, and forum contributors.

My thanks also go to all the great colleagues and friends that made working in the western wing of the Bouwkunde such a nice experience: Duco, Igor, André, Ali, Rodrigo, Ruta, Karel, Arie, Ana, people from the former SRO section and from the Urban Studies section of the Department of Urbanism.

I am also very thankful to my Dutch brother Douwe, to all the friends who came to Rotterdam during these four years, to my father, mother, sister and brothers, to the Ramon-Deilhes Family, and to Alicia, who has been by my side throughout this PhD and gave me so much support and joy during these years.

TOC

Contents

List of Tables

List of Figures

Summary

This PhD research develops alternative methods of measuring patterns of interrelationships between cities, and uses them to advance our understanding of the spatial organization and the evolution of the Dutch system of cities. The underlying knowledge gap that motivated this work was that for many years, urban system research has focused on aspects such as the concentration of populations, economic activities and urban functions, but had a limited focus on the actual networks connecting cities and the impact they have on urban dynamics. During the first wave of urban systems research in the 1960s–1980s, only a limited number of studies were actually looking at relational data and the mentions of flows and networks were mostly metaphorical. This slow development of a network approache to urban systems was mainly due to limitations in computational power and data availability. More recently, these issues were partly addressed by the development of 'big data'. The availability of big geospatial databases related to the development of location-aware technologies has been the cause of a new surge of interest for urban systems research. Cross-fertilization with the emerging fields of network sciences and complex systems analyses have led to new frontiers in urban research; notably the possibility to work on very large population groups at very fine spatial and temporal scales. However, these approaches are not without limitations. Beyond common statistical problems related to 'big data' sources such as ecological fallacy, self-selection bias and non-representativeness, the fact that this data is generated by fast changing technologies with fluid categories makes it difficult to create datasets that are consistent over time to analyse cities from an urban systems perspective.

Research approach

For this dissertation, I was inspired by a recent stream in urban systems research that builds on assembling data sources, such as archives and registers, and analysing these with modern computational methods to explore patterns of interurban relationships. In my PhD research, I proposed to further explore these approaches in the context of the Netherlands. My aim was to develop new methods to measure relations between cities that are consistent through time and over the entire Dutch urban system. In order to achieve this goal, I explored the potential of three datasets: two text archives (the Common Crawl Web Archive and Delpher) and micro-level register data of Statistics Netherlands. In a first experiment to fulfil this aim, I looked

at the co-mention of city names in an internet archive. This was an opportunity to upscale the 'co-occurrence method', to approach text as data and to identify the challenges of dealing with such sources. The research question I answered was*: to what extent can the toponym co-occurrence method be used to identify the patterns of relations between cities in a systematic way?* (*RQ1).* To respond to the issue of time consistency, my strategy was to use another database containing text data: the Delpher archive that includes millions of digitally searchable historical newspapers. My second research question was about dealing with the unstructured aspect of these sources: *how can we extract data on relations between cities for a long period of time from a massive archive of historical newspapers? (RQ2).* The third question was about the usefulness of such data sources. I wanted to know *to what extent does data extracted from historical newspaper archives reflect the long-term evolution of the territorial organisation of a system of cities (RQ3).* Finally, in the final empirical chapter of this dissertation, I questioned the scale of analysis of urban system research because intra- and interurban processes are increasingly difficult to differentiate as urbanisation intensifies, especially in the Dutch context where numerous big cities are located very close to each other. I then tried to answer the following question: *to what extent can longitudinal individual data be used to build bottom-up definitions of cities and urban systems? (RQ4).*

Summary of the research results

Chapter 2 of this thesis, published in *Networks and Spatial Economics*, presents an overview of the system of cities literature since the mid-1990s. This study of the evolution of the urban systems literature does not take the form of a classical literature review study. Rather it adopts a bibliometric approach to analyse a set of 1,491 papers on intercity relationships from 1995 onwards. The methodology I use combines an analysis of semantic and citation networks to give a 'bird's eye view' of the scientific field. This analysis results in the identification of 5 different schools of thought that manifest strong differences in terms of methods, scale of analysis, thematic focus, data, and influences from other disciplines. The first school of thought identified is the research on world cities (WCN) which focuses on the top of the urban hierarchy by looking at global networks (mainly corporate and transportation networks). The second group of publications that appeared clearly from this analysis is the cluster about regional urban systems (REG), centred around the analysis of polycentric urban regions. A third cluster focused on the study of systems of cities as complex systems also appears (SIM). This cluster is characterised by terms related to modelling and simulation. A fourth cluster gathers the vocabulary of studies dealing with aspects associated with city size (CSD). It contains all the publications dealing with the descriptions and explanations behind

the regularity in city size distributions. Finally, a last cluster characterized by the vocabulary of economic geography also appeared (ECON). In terms of citation patterns, three out of five schools of thought (REG, WCN and CSD) appeared as very coherent internally. However, recently, an increase in inter-cluster citations can be observed, indicating increasing interactions between the subfields.

Chapter 3, published in the *International Journal of Urban Sciences*, investigates the feasibility of using the toponym co-occurrence method on web pages containing city names to identify patterns of relations between cities in a systematic way (RQ1). This approach builds urban systems on the basis of co-mentions of place names in documents of a text corpus. In this research, millions of web pages contained in the Common Crawl web archive are analysed. Because the city names were retrieved with simple string queries, attention is also paid to evaluating the impact of ambiguous place names in the resulting frequency of relations. In order to benchmark the method, gravity modelling is employed to assess the resulting spatial organization of the Netherlands. It turns out that the gravity model fits the pattern of relationships between places as found in the corpus of web pages rather well, which contributes to the assessment that the toponym co-occurrence method is an interesting proxy for relationships in real physical space. The usability of the toponym co-occurrence method therefore hinges on a good quality dataset to which it is applied. Ideally, the dataset would allow a better characterization the meaning of a co-occurrence of place names as they appear within a document, a paragraph, or a sentence. Rather than closing the discussion, this chapter leads to new challenges for improving the retrieving of geographical information from unstructured text data. Challenges that I try to meet in the two next chapters.

Chapter 4, published as a data paper in *Cybergeo, European Journal of Geography*, answers the second research question about the extraction of meaningful data on intercity relations from an archive of digitized historical newspapers (RQ2). Such extraction relies on four crucial steps: the careful selection of a sub-corpus and of relevant geographical objects, the creation of a meaningful ontology to go from a collection of news items to an origin-destination matrix, the identification of problems in automatic identification of the city names and finally, the design of an algorithm in a trade-off between accuracy and computing time. Once applied, the algorithm, which is based on string queries and performs Named Entity Recognition (NER) for ambiguous place names, produces very good results in terms of accuracy. The result of this research is DIGGER, a dataset that allows for a study of the spatial diffusion of

information on and between Dutch cities from a corpus of 81 newspapers published in 29 different cities between 1869 and 1994, which has been published as Open data[1].

Chapter 5, published in *Computers, Environment and Urban Systems*, aims to answer my third question about the usefulness of such data to study the evolution of the spatial organisation of a territory (RQ3). The overarching goal of the chapter is to test the potential of this novel data source for reconstructing the evolving spatial organisation of an entire country. In order to achieve this goal, I look for regularities in how flows of information develop over time and test different hypotheses related to the evolution of systems of cities previously identified by the literature, that could not, however, always be empirically tested yet. The findings confirm previous research that shows that space matters greatly in the process of information diffusion. During the period 1869–1930, I observed a decrease of the distance decay in the spread of information, with almost all newspapers increasingly reporting news about distant cities. However, this process is not homogeneous through space. In many cases, increasing attention is paid to places in the immediate proximity of where a newspaper was published and at the same time, to the four largest cities of the Randstad, most of the time at the expense of closer-by medium-sized cities. The main driving factors behind this increase of long-distance interactions are indeed related to a polarization around the main economic, political, and demographic core of the Netherlands, that from 1938 onwards would be referred to as the "Randstad". Based on these empirical results, I conclude that it is feasible to use a computational social science approach to construct completely novel, geographically relevant data sets on interurban relations, which allow the reconstruction of the evolution of the spatial organisation of a territory over time (RQ3).

Chapter 6 answers my last research question about the potential of using microdata to build bottom-up definitions of cities and urban systems (RQ4). In order to create these definitions of cities on a functional basis, I present a method that is applicable in the Dutch context where there is no exhaustive commuting data. This method departs from the extraction of data on the job and residential careers of the full Dutch population to identify people moving without changing job ('stable workers'). Because these people continue working at the same place but commute from two different locations, we can deduct that these two locations belong to the same labour and housing market. Using a hierarchical clustering algorithm, I then extract the nested hierarchy of functional regions in the Netherlands based on these flows. The different steps of this procedure reveal meaningful functional regions,

1 https://data.4tu.nl/articles/DIGGER_a_dataset_built_on_Delpher_the_digital_archive_of_historical_newspapers_of_the_National_Library_of_the_Netherlands/12709190

their nested hierarchy, and their integration into wider functional regions at a higher scale. Among these multiple definitions of functional regions that were created, one provides an alternative to the existing division of labour market areas that is currently based on the perception of local stakeholders. One of the main advantages of our method is that it is based on individual data and allows for an exploration of the movement patterns of different population groups. To explore such possibilities, I investigate which age group of workers is the most inclined to migrate between two functional regions while keeping their job. These flows indicate an integration of two regions at a higher scale. This brief analysis shows that workers between 25 and 34 years old are the ones that have a higher chance of moving between two regions while keeping their job compared with older workers. From this exploration I conclude that using microdata on job and residential careers of individuals provides a good alternative to commuting data for delimiting functional regions and patterns of interconnections between them, and presents a great opportunity to understand the relative functional integration of cities and urban systems depending on population subgroups (RQ4).

Synthesis of the results and scientific contribution

I started this PhD thesis with the aim of developing alternative methods for measuring patterns of interrelationships between cities and evaluating their potential to advance our understanding of the organisation and evolution of the Dutch system of cities. The exploratory dimension of this aim, and my strategy of exploiting three different databases, have led me to a situation where I have opened several doors without necessarily closing them. However, these explorations were characterised by successes and identifications of limitations that further advance the debate on the data issue in urban system research.

An important contribution of this thesis has been to give an overview of the research on the systems of cities. Because of its multidisciplinary dimensions, urban system research is not very well integrated, and one often comes across studies that ignore important parts of the literature or ‘rediscover’ existing approaches. Providing such an overview has been made possible by doing a bibliometric analysis of the research field. The procedure that was designed to delineate the scientific landscape has proven to be efficient and resulted in a set of publications highly relevant for the subject of my dissertation. I believe that this robust way of delineating a scientific field can help researchers beyond geographic research willing to conduct a bibliometric analysis.

With a focus on two text archives, this thesis also contributed to an emerging trend of research, which explores ways of automating the analysis of cities through their mentions in media or textual materials. An important part of this thesis has been dedicated to the use of computational methods, and notably Named Entity Recognition (NER), to extract geographical information from raw text sources. Beyond the data I collected, this approach can be adapted and extended by researchers wanting to process other types of text documents mentioning cities or geographical entities.

I have also shown that such sources can be used to analyse actual spatial dynamics. Chapter 5 of this thesis demonstrates the possibility of developing a theory-driven experiment based on data collected from the archive and answers geographical questions related to the evolution of a system of cities. This work gives an example of the possibilities offered by computational social sciences. It departed from the assemblage of a massive number of historical sources that was analysed with Natural Language Processing (NLP) methods, and a modelling approach rooted in quantitative geography was applied in order to reveal hidden patterns in the corpus related to the evolution of the territorial organisation of the country.

The exploration of the potential of using microdata for analysing the Dutch urban system has led to two main contributions. The first one is that using residential relocations of stable workers provides a good alternative to commuting data for delimiting functional regions and their interdependence from a job and housing market point of view. The second one is that it opens new possibilities for considering individual level heterogeneity and for analysing the functional space of different population subgroups, aspects that are important to better understand the organisation of cities and urban systems.

This work has contributed to the line of research that seeks to approach cities in a relational way. It has highlighted processes and mechanisms that link cities together or make them evolve interdependently. These elements show that cities cannot and should not be seen as closed, bounded, coherent entities, but rather as open, complex, and interconnected to ranges of other spaces and places.

Samenvatting

In dit promotieonderzoek worden alternatieve methoden ontwikkeld om patronen van onderlinge relaties tussen steden te meten en worden deze methoden gebruikt om ons inzicht in de ruimtelijke indeling en de evolutie van het Nederlandse stedenstelsel te vergroten. De motivatie voor dit proefschrift was dat het onderzoek naar stedenstelsels zich al jarenlang toespitst op aspecten als de concentratie van bevolkingsgroepen, economische activiteiten en stedelijke functies, maar dat kennis over – en aandacht voor – de eigenlijke netwerken die steden met elkaar verbinden en de invloed ervan op de stedelijke dynamiek beperkt is. In de eerste golf van onderzoek naar stedenstelsels, in de jaren zestig tot en met tachtig, waren er maar weinig studies waarin echt naar relationele data werd gekeken. De begrippen 'stromen' en 'netwerken' werden vooral metaforisch gebruikt. Dat het gebruik van netwerkmethodes bij onderzoek naar stedenstelsels zo langzaam op gang kwam, lag hoofdzakelijk aan beperkingen in rekenkracht en beschikbaarheid van data. Door de ontwikkeling van 'big data' zijn deze beperkingen de laatste jaren gedeeltelijk verdwenen. De beschikbaarheid van grote geospatiale databases als gevolg van de ontwikkeling van locatietechnologie heeft geleid tot een nieuwe golf van belangstelling voor onderzoek naar stedenstelsels. Door kruisbestuiving met nieuwe vakgebieden – netwerkwetenschappen en analyse van complexe systemen – zijn er nieuwe mogelijkheden ontstaan voor het onderzoek naar steden, met name in het werken met zeer grote bevolkingsgroepen op zeer fijnmazige schalen van ruimte en tijd. Deze benaderingen kennen echter hun beperkingen. Naast de gewone statistische problemen met bronnen van big data, zoals 'ecologische fouten', vertekening door zelfselectie en niet-representativiteit, is het probleem ook dat deze data worden gegenereerd door snel veranderende technologieën met categorieën die niet vastliggen. Daardoor is het moeilijk om tijdsconsistente datasets te bouwen om steden te analyseren vanuit het perspectief van stedenstelsels.

Onderzoeksaanpak

Voor dit proefschrift ben ik geïnspireerd door recente onderzoeken naar stedenstelsels op basis van gegevensbronnen zoals archieven en registers, waarbinnen met moderne rekenmethoden patronen van interstedelijke relaties worden geanalyseerd. In mijn promotieonderzoek wilde ik deze benaderingen verder onderzoeken in de context van Nederland. Mijn doel was nieuwe methoden

te ontwikkelen om relaties tussen steden te meten die consistent zijn in de tijd en binnen het gehele Nederlandse stedenstelsel. Hiervoor heb ik gekeken naar de mogelijkheden die worden geboden door drie datasets: twee tekstarchieven (het Common Crawl Web Archive en Delpher) en de microdata van het CBS. In een eerste experiment heb ik gekeken naar de gelijktijdige vermelding van stadsnamen in een internetarchief. Dit was een gelegenheid om de *co-occurrence*-methode ('gezamenlijk voorkomen') op te schalen, door tekst als data te benaderen en te kijken welke problemen er zijn in het werken met dergelijke bronnen. Mijn eerste onderzoeksvraag is: *In hoeverre kan de* toponym co-occurrence-*methode worden gebruikt om de relatiepatronen tussen steden op een systematische manier te benoemen? (OV1)*. Om consistentie in de tijd te bewerkstelligen heb ik een andere database met tekstgegevens gebruikt: het Delpher-archief, dat miljoenen digitaal doorzoekbare historische kranten bevat. Mijn tweede onderzoeksvraag betreft het ongestructureerde karakter van deze bronnen: *Hoe kunnen we uit een enorm archief van historische kranten gegevens extraheren over de relaties tussen steden gedurende een lange periode? (OV2).* De derde vraag gaat over het nut van dergelijke gegevensbronnen. Ik wilde weten: *In hoeverre weerspiegelen data uit historische krantenarchieven de langetermijnevolutie van de territoriale indeling van een stedenstelsel? (OV3).* In het laatste empirische hoofdstuk van dit proefschrift breng ik de schaal van de analyse in het onderzoek naar stedenstelsels ter sprake, omdat intra- en interstedelijke processen door de toenemende verstedelijking steeds moeilijker van elkaar te onderscheiden zijn. Dat geldt des te meer in de Nederlandse context waar talrijke grote steden zeer dicht bij elkaar liggen. Ik heb hiervoor de volgende vraag proberen te beantwoorden: *In hoeverre kunnen longitudinale individuele data worden gebruikt om bottom-up definities van steden en stedenstelsels op te stellen? (OV4).*

Samenvatting van de onderzoeksresultaten

Hoofdstuk 2 van dit proefschrift, dat is gepubliceerd in *Networks and Spatial Economics*, geeft een overzicht van de literatuur over stedenstelsels sinds het midden van de jaren negentig. Dit onderzoek naar de evolutie van de literatuur over stedenstelsels heeft niet de vorm van een klassieke literatuurstudie, maar is een bibliometrische analyse van een verzameling van 1491 artikelen over interstedelijke relaties vanaf 1995. In mijn analyse combineer ik semantische en citatienetwerken om een breed perspectief van het vakgebied te geven. Dit resulteert in de identificatie van vijf verschillende stromingen die sterk verschillen in methode, schaal van de analyse, thematische focus, gebruikte data en invloeden uit andere disciplines. De eerste van deze vijf stromingen is die van het onderzoek naar wereldsteden (WCN) met de focus op de top van de stedelijke hiërarchie, waarbij naar wereldwijde

netwerken (voornamelijk bedrijfs- en vervoersnetwerken) wordt gekeken. De tweede groep publicaties die duidelijk uit deze analyse naar voren kwam, is het cluster dat zich bezighoudt met regionale stedenstelsels (REG), met de focus op analyse van polycentrische stedelijke regio's. Het derde cluster (SIM) bestaat uit publicaties over onderzoek naar stedenstelsels als complexe systemen. In dit cluster worden termen gebruikt die verband houden met modellering en simulatie. Een vierde cluster is gegroepeerd rondom het vocabulaire van onderzoeken die betrekking hebben op bepaalde aspecten van de stadsgrootte (CSD). Dit cluster bevat alle publicaties die betrekking hebben op de beschrijvingen en verklaringen achter de regelmatigheid in stadsgrootteverdelingen. En het vijfde cluster dat ik vond kenmerkt zich door het vocabulaire van de economische geografie (ECON). In termen van citatiepatronen bleken drie van de vijf stromingen (REG, WCN en CSD) intern zeer coherent te zijn. Sinds kort is er echter een toename van citaties tussen clusters te zien, wat wijst op meer interactie tussen de deelgebieden.

In **hoofdstuk 3**, gepubliceerd in het *International Journal of Urban Sciences*, wordt onderzocht in hoeverre de *toponym co-occurrence*-methode geschikt is om op systematische wijze relatiepatronen tussen steden vast te stellen voor webpagina's met stadsnamen (OV1). Bij deze methode worden stedenstelsels gedefinieerd op basis van gelijktijdige vermeldingen van plaatsnamen in documenten van een tekstcorpus. In dit onderzoek zijn miljoenen webpagina's uit het webarchief Common Crawl geanalyseerd. Aangezien de plaatsnamen zijn opgehaald met eenvoudige string-query's, wordt ook gekeken wat de invloed is van ambigue plaatsnamen op de resulterende frequentie van relaties. Om de methode te benchmarken wordt met behulp van een zwaartekrachtmodel de resulterende ruimtelijke indeling van Nederland beoordeeld. Het zwaartekrachtmodel blijkt behoorlijk goed te passen bij het patroon van relaties tussen plaatsen zoals gevonden in het corpus van webpagina's, hetgeen bijdraagt aan het oordeel dat de *toponym co-occurrence*-methode een interessante indicator is voor relaties in de echte fysieke ruimte. De *toponym co-occurrence*-methode is alleen bruikbaar als zij wordt toegepast op een dataset van goede kwaliteit. Idealiter maakt de dataset het mogelijk om beter te verklaren wat het betekent dat plaatsnamen gezamenlijk voorkomen in een document, een alinea of een zin. Dit hoofdstuk biedt geen afsluiting van de discussie maar leidt tot nieuwe uitdagingen om op een betere manier geografische informatie uit ongestructureerde tekstgegevens te extraheren. Deze uitdagingen komen in de volgende twee hoofdstukken aan de orde.

Hoofdstuk 4, gepubliceerd als datapaper in *Cybergeo, European Journal of Geography*, beantwoordt de tweede onderzoeksvraag over het extraheren van gegevens over interstedelijke relaties uit een archief van gedigitaliseerde historische kranten (OV2). Voor dit proces zijn vier cruciale stappen nodig: een zorgvuldige

keuze van een subcorpus en van relevante geografische objecten; het creëren van een zinvolle ontologie om van een verzameling nieuwsberichten tot een oorsprong-bestemmingsmatrix te komen; het benoemen van problemen bij de automatische identificatie van de plaatsnamen; en het ontwerp van een algoritme, waarbij een afweging wordt gemaakt tussen nauwkeurigheid en rekentijd. Dit algoritme, dat gebaseerd is op string-query's en dat *Named Entity Recognition* (NER) alleen uitvoert voor ambigue plaatsnamen, geeft zeer nauwkeurige resultaten. Dit onderzoek heeft geleid tot DIGGER, een dataset waarmee de ruimtelijke verspreiding van informatie over en tussen de Nederlandse steden kan worden bestudeerd vanuit een corpus van 81 kranten die tussen 1869 en 1994 in 29 verschillende steden zijn gepubliceerd, en die als open data is gepubliceerd.[2]

Hoofdstuk 5, gepubliceerd in *Computers, Environment and Urban Systems*, is gewijd aan mijn derde vraag, over de bruikbaarheid van dergelijke gegevens voor het bestuderen van de evolutie van de ruimtelijke indeling van een gebied (OV3). Het algehele doel van dit hoofdstuk is om te testen welk potentieel deze nieuwe gegevensbron heeft als we willen reconstrueren hoe de ruimtelijke indeling van een heel land zich heeft ontwikkeld. Hiertoe kijk ik naar regelmatige patronen in de wijze waarop informatiestromen zich in de tijd ontwikkelen, en test ik verschillende hypothesen over de evolutie van stedenstelsels die eerder in de literatuur zijn benoemd, maar die nog niet altijd empirisch konden worden getest. De bevindingen bevestigen eerder onderzoek waaruit blijkt dat ruimte een belangrijke rol speelt in het proces van informatieverspreiding. Ik constateer in de loop van de onderzochte periode een afname van het afstandsverval bij de verspreiding van informatie; bijna alle kranten brengen steeds vaker nieuws over verafgelegen steden. Dit proces is echter niet homogeen in de ruimte. In veel gevallen wordt er steeds meer aandacht besteed aan plaatsen in de onmiddellijke nabijheid van waar een krant wordt gepubliceerd, en tegelijkertijd aan de vier grootste steden van de Randstad, meestal ten koste van de middelgrote steden die dichterbij liggen. De belangrijkste drijvende factoren achter deze toename van lange-afstandsinteracties houden verband met een polarisatie rondom de economische, politieke en demografische kern van Nederland, die sinds 1938 de 'Randstad' wordt genoemd. Op basis van deze empirische resultaten concludeer ik dat het mogelijk is om met behulp van een methode uit de computationele sociale wetenschap volledig nieuwe geografisch relevante datasets over interstedelijke relaties op te bouwen, waarmee de evolutie van de ruimtelijke indeling van een gebied in de tijd kan worden gereconstrueerd (OV3).

2 https://data.4tu.nl/articles/DIGGER_a_dataset_built_on_Delpher_the_digital_archive_of_historical_newspapers_of_the_National_Library_of_the_Netherlands/12709190

Hoofdstuk 6 beantwoordt mijn laatste onderzoeksvraag over de mogelijkheid om met behulp van microdata bottom-up definities van steden en stedenstelsels op te stellen (OV4). Om dit op functionele basis te doen, presenteer ik een methode die toepasbaar is in de Nederlandse context, waar geen uitputtende gegevens over woon-werkverkeer voorhanden zijn. Bij deze methode worden data over de woon- en werkgeschiedenis van de volledige Nederlandse bevolking geëxtraheerd om mensen te identificeren die verhuizen zonder van baan te veranderen ('stabiele werknemers'). Aangezien deze mensen op dezelfde plaats blijven werken, maar vanuit een andere plaats gaan forensen, kunnen we concluderen dat de huidige en de vorige woonlocatie qua arbeidsmarkt en woningmarkt tot hetzelfde gebied behoren. Met een algoritme voor hiërarchische clustervorming extraheer ik vervolgens op basis van deze stromen de geneste hiërarchie van functionele regio's in Nederland. De verschillende stappen van deze procedure geven een beeld van betekenisvolle functionele regio's, hun geneste hiërarchie en hun integratie in bredere functionele regio's op een hoger schaalniveau. Van de vele definities van functionele regio's biedt er één een alternatief voor de bestaande verdeling van de arbeidsmarktgebieden, die momenteel gebaseerd is op de perceptie van lokale stakeholders. Een belangrijk voordeel van onze methode is dat deze gebaseerd is op individuele gegevens, en het mogelijk maakt om verplaatsingspatronen van verschillende bevolkingsgroepen te onderzoeken. Hiertoe kijk ik welke leeftijdsgroep van werknemers het meest geneigd is om met behoud van baan naar een andere functionele regio te verhuizen. Deze stromen wijzen op een integratie van twee regio's op een hoger schaalniveau. Uit deze korte analyse blijkt dat werknemers tussen 25 en 34 jaar vaker dan oudere werknemers van de ene regio naar de andere verhuizen terwijl ze hun baan behouden. Als we functionele regio's en patronen van onderlinge verbindingen daartussen willen afbakenen, is het gebruik van microdata over woon- en werkgeschiedenis van individuen dus een goed alternatief voor data over woon-werkverkeer; ik concludeer dat dit een veelbelovende aanpak is voor het verkrijgen van inzicht in de relatieve functionele integratie van steden en stedenstelsels, afhankelijk van de subgroepen van de bevolking (OV4).

Synthese van de resultaten en wetenschappelijke bijdrage

Ik ben aan dit promotieonderzoek begonnen om alternatieve methoden te ontwikkelen voor het meten van interrelationele patronen tussen steden en om te evalueren in hoeverre dergelijke methoden kunnen bijdragen aan ons begrip van de indeling en de evolutie van het Nederlandse stedenstelsel. Het verkennende karakter van deze doelstelling en mijn strategie om drie verschillende databases te gebruiken hebben ertoe geleid dat ik verschillende deuren heb geopend zonder ze per se ook

te sluiten. Deze verkenningen hebben geleid tot zowel successen als het constateren van beperkingen, en helpen de discussie over het gebruik van data in onderzoek naar stedenstelsels dus vooruit.

Een belangrijke bijdrage van dit proefschrift is dat er een overzicht wordt gegeven van het onderzoek naar stedenstelsels. Onderzoek naar stedenstelsels is vanwege het multidisciplinaire karakter niet erg goed geïntegreerd en men stuit vaak op onderzoek waarin belangrijke delen van de literatuur worden genegeerd of waarin bestaande benaderingen worden 'herontdekt'. Ik ben erin geslaagd een literatuuroverzicht te maken door een bibliometrische analyse van het onderzoeksgebied uit te voeren. De procedure die ik heb ontworpen om het wetenschappelijke landschap af te bakenen, is efficiënt gebleken en heeft geresulteerd in een reeks publicaties die zeer relevant zijn voor het onderwerp van mijn proefschrift. Naar mijn mening kan deze robuuste methode om een wetenschappelijk gebied af te bakenen behulpzaam zijn voor onderzoekers die bereid zijn om naast geografisch onderzoek ook een bibliometrische analyse uit te voeren.

Met mijn focus op twee tekstarchieven heb ik ook bijgedragen aan een opkomende trend in het onderzoek naar manieren om de analyse van steden te automatiseren op basis van hun vermeldingen in media of tekstmateriaal. In dit promotieonderzoek is een belangrijke plaats weggelegd voor computationele methoden – met name *Named Entity Recognition* (NER) – om geografische informatie uit ruwe tekst-bronnen te extraheren. Los van de data die ik hiermee heb verzameld, kan deze methode ook worden aangepast en uitgebreid door onderzoekers die andere soorten tekstdocumenten willen verwerken waarin steden of geografische entiteiten worden genoemd.

Ik heb ook aangetoond dat dergelijke bronnen kunnen worden gebruikt om daad-werkelijke ruimtelijke dynamiek te analyseren. Uit hoofdstuk 5 van dit proefschrift blijkt dat het mogelijk is om op basis van de data uit het archief een theoriegestuurd experiment te ontwikkelen en om geografische vragen over de evolutie van een stedenstelsel te beantwoorden. Dit werk geeft een voorbeeld van wat er mogelijk is met computationele sociale wetenschap. Het uitgangspunt was het verzamelen van een groot aantal historische bronnen die zijn geanalyseerd met behulp van *Natural Language Processing* (NLP) en ik heb een modelleermethode uit de kwantitatieve geografie toegepast om verborgen patronen in het corpus aan het licht te brengen die verband houden met de evolutie van de territoriale indeling van het land.

De verkenning van de mogelijkheid om met microdata het Nederlandse stedenstelsel te analyseren heeft tot twee belangrijke bijdragen geleid. Ten eerste blijken gegevens over verhuizingen van stabiele werknemers een goed alternatief voor gegevens over

woon-werkverkeer, als we functionele regio's en hun onderlinge afhankelijkheid willen afbakenen vanuit het perspectief van arbeidsmarkt en woningmarkt. Ten tweede zijn er nieuwe mogelijkheden aan het licht gekomen om rekening te houden met heterogeniteit op individueel niveau en om de functionele ruimte van verschillende subgroepen van de bevolking te analyseren, aspecten die van belang zijn om de indeling van steden en stadsstelsels beter te begrijpen.

TOC

1 Introduction

Research on urban dynamics has highlighted the importance of the spatial scale of the *system of cities*. Cities never function in isolation but as nodes in overarching systems characterised by flows of goods, people, and information. To fully understand the evolution of cities, a relational approach is needed, which investigates cities in relation to other cities and urban regions. While a significant part of urban system research has focused on aspects such as the concentration of populations, economic activities, and urban functions, the understanding of the actual networks connecting cities and their impact is still limited and needs to be further developed. However, the required data is notoriously difficult to obtain. This dissertation contributes to knowledge on the relationship between cities in the Netherlands by exploiting – in novel ways – three data sources: web pages mentioning cities, local historical newspapers, and administrative registers.

1.1 Overall context of the research

In the early days of quantitative geography research, the main principle of urbanisation has been described by Torsten Hägerstrand as "the art of moving material, people and information"[3]. This formula adequately illustrates a revolution in the understanding of cities during the middle of the 20th century, when geographers started to be interested not only in the places themselves but also in the interactions happening between them. Nowadays, the idea that cities can be conceived as the result of interactions between agents and between agents and their environment is widely acknowledged (Neal, 2013; Pumain, 2011). Studying them as a set of interactions rather than locations is the first principle of the "new science of cities" introduced by Batty (2013) and is at the core of the research agenda of quantitative studies of cities (Lobo et al., 2020).

3 Translated and cited by Törnqvist (1968)

Networks and flows are not bounded to the intra-urban scale. Exchanges and relations also take place between cities. Interurban exchanges are even among the first characteristics described by urban theorists and historians to describe the emergence and maintenance of cities in history (Braudel, 1979; Mumford, 1961). As explained by Scott and Storper (2015), "cities have *always* functioned as nodes in systems of long-distance trade". But trade is only one of the many forms of interurban relations identified by urban scholars. Cities are indeed connected by multiple social and economic networks that translate into flows of material goods, exchanges of information, transfers of capital, and human mobility.

Today, more than ever, cities across regions, countries, continents and even across the world are tied by networks of interdependencies. A recently developed line of research on 'borrowed size' in systems of cities shows that the embeddedness in national and international networks plays an important role in explaining the level of performance of cities (Meijers et al., 2016; Meijers and Burger, 2017). According to Hesse (2014), "borrowing size or significance no longer relies on physical proximity between the cities, but on embeddedness in overarching networks between and within polycentric city-regions, via corporate relations, market pervasion and, last but not least, information and communication networks." The rapidity at which the COVID-19 pandemic spread across the world and the huge impact that limiting the circulation of people and goods had on the economy is a striking illustration that networks and relations between distant places are crucial in current societies. Understanding these patterns of interrelationships is therefore of critical importance.

An important body of scientific literature on these interurban relations has grown since the middle of the 20th century. However, according to Pumain (2003), only a limited number of studies[4] on urban systems were actually looking at relational data; the mention of flows and networks was mostly metaphorical in the first wave of urban system research. According to Ducruet and Beauguitte (2013), this slow development of network analysis in geography is partly due to limitations in computational power and data availability. Similar conclusions were made by researchers interested in the world city network. They were pinpointing the data scarcity issue as the "dirty little secret" of the field (Short et al., 1996) that was characterised by "theoretical sophistication and empirical poverty" (Taylor, 2004).

4 See for example Board et al., 1970; Cattan, 1990; Dietvorst and Wever, 1977; Rozenblat and Pumain, 1993

To solve this empirical poverty, at the end of the 1990s and the beginning of the 2000s, more and more geographers started to look at new ways to proxy relations between cities, for instance deriving them from the spatial distribution of multinational corporations (Taylor, 2001) or by doing large surveys on a limited number of firms (Rozenblat and Pumain, 2007). However, the development of network approaches in urban systems research really took off over the last decade. This surge in interest can be largely attributed to the availability of big geospatial databases related to the development of locational-aware technologies such as the mobile phone, social media, and navigation systems (Arribas-Bel and Reades, 2018; Miller and Goodchild, 2015). This period has been characterised by the cross-fertilisation with the emerging field of network sciences and contributions from physicists and computer scientists. Such datasets have been used to precisely map human communication between cities (Krings et al., 2009) or to map human mobility and city attractiveness with data from user-generated content on social media (Lenormand et al., 2015; Zhang et al., 2016). 'Big data' has opened up new frontiers in urban research, and notably the possibility to work on local events with very fine time granularity.

However, over the last few years, important limitations of such datasets were discussed. For instance, the difficulty of knowing the true representativeness of this data and the quick development, and changes in technology make this data hard to use for longitudinal research (Kitchin, 2013). However, urban system research needs longitudinal analysis because having a certain time depth is essential for understanding cities (Pumain, 1997). More traditional sources, that do not fall under the definition of 'big data', because, for example, they lack velocity, might have been overlooked. Recently, teams of geographers have shown that more traditional data such as archives and registers, processed and analysed through modern computational methods have a great potential for the analysis of urban systems at multiple spatial scale (Breschi and Lenzi, 2016; Ducruet et al., 2018; Rozenblat, 2018).

This PhD research will further explore such approaches in the context of the Netherlands. The aim is to develop new methods to measure relations between cities that are consistent through time and over the entire urban system. In order to achieve this goal, this research explores the potential of three datasets than can be analysed with modern computational methods: two text archives and register data. Section 2 of this chapter elaborates on the way in which interurban relations are approached in urban system literature. Section 3 deepens the question of the intricate dimension of data availability and progress in urban systems research. Section 4 presents the research approach. Finally Section 5 details the outline of this thesis.

1.2 The relational dimension of urban systems

It is quite paradoxical that the study of relations between cities has received so much attention without any consensus on the terminology to characterise its object. In scientific literature, sets of interdependent cities are referred to with several terms such as 'urban system', 'system of cities', 'urban network', 'city-system', 'settlement system', *etc*. Some of these terms are more ambiguous than others. As already argued by Pred (1977), 'urban system' is often used by geographers to refer to individual cities. The same holds true for 'urban network', which can characterise an intra-urban transport network for example. However, while there are some inconsistencies due to the fact that these terms are used loosely and interchangeably, they refer to a precise body of scientific literature and methods of analysis (Derudder, 2019; Pumain, 2003). This body of literature, that will be refer to as 'urban system research', has grown over the 20th century, drawing mainly on contributions from geographers, but also from other social sciences such as history, economics, sociology, and archaeology. More recently, the field has regained a lot of attention due to the intervention of physicists and computer scientists trying to describe the quantitative properties of cities with massive geospatial databases that appeared with the rise of information and communication technologies (ICT). In this section I will briefly retrace the different phases in urban system research with an explicit focus on the way relations were studied.

The first systematic studies on set of cities were focused on the regularity in the size distribution of cities (Auerbach, 1913; Zipf, 1949). Most of the time, in these studies, the actual relations were implicit (Pumain, 2003) but some studies, such as the one of Vapnarsky (1969), interpret the shape of the distribution with considerations on the "integration" of the system. For this author, if the settlement system in a country or region followed a clear rank-size distribution, it would be characterised by a high degree of interdependence, while the presence of a primate city would reflect a low level of integration. Drawing on a pioneering contribution from Christaller (1933), sets of cities were also analysed through Central Place Theory, a model describing the size, spatial distribution and number of cities by focusing on their nested hierarchy and linking economic functions with size of the centre and its catchment

area. This approach received a lot of attention in the 1950s-1960s[5]. But despite the presence of a 'transport principle' in the initial formulation of the theory (which results from the search for economy in travel between central places), the analysis of actual exchanges and flows between cities was largely ignored or implicit.

In the 1970s, urban systems were at the core of the research agenda of the fast growing discipline of quantitative geography. Törnqvist (1968) provides an explanation for this increasing interest in the study of relations between cities. For him, the post-Second World War economy witnessed important changes in the way the economy was organised. The agrarian population declined and 'urban activities' such as manufacturing, storage, trade, transport, communications, and services rose rapidly. These types of activities were characterised by specialised production units that necessitated exchanges of goods and information between them. Based on these observations, he developed a theory on interurban exchanges where information flows play the most determinant role. He was among the first ones to conduct an actual study on the flows of information between cities with data from a survey about contacts between business holders and managers (Pred and Törnqvist, 1973). Similar conceptions of information flows in systems of cities were present in the work of Pred, who analysed how much time it took for information to travel between cities, revealing the structure of the American urban system in the 19th and early 20th century (Pred, 1977, 1971). But aside from these empirical studies, the vast majority of urban system research remained focused on population data and urban functions. This might be due to data availability. Indeed, there have been some studies on the patterns of interactions between territories at the regional scale (Berry, 1966; Simmons, 1970), but it is hard to find similar attempts at the interurban scale. For this reason, Pumain (2003) concludes that the use of notions such as 'system' and 'network' in urban system research at this time was mostly metaphorical.

An hypothesis that can explain the late development of actual network approaches to urban systems is the almost exclusive focus on planar graphs[6] by quantitative geographers between the 1960s and 1980s (Ducruet and Beauguitte, 2013). This focus is very clear in the classical book by Haggett and Chorley (1969) where non-planar graphs receive little attention. However, the networks characterising systems of cities do not necessarily follow the definition of planarity. According to Ducruet and Beauguitte (2013), an important turning point can be identified in the 1990s, when

5 see the annotated bibliography by Berry and Pred (1961) for an overview of the field at the beginning of the 1960s

6 Planar graphs are graphs that can be drawn on a plane without crossing edges.

geographers moved away from these classical network approaches and started to be interested in non-planar graphs: “Studies of the European urban system progressively integrated a network dimension with the works of Cattan (1995a, b) on airlines and railways, and Rozenblat and Pumain (1993) on multinational firms, thereby opening new ways considering systems of cities”. It is also during this period that a first typology of interurban relationships was drawn (Smith and Timberlake, 1995).

From the 1990s onward, urban system research evolved in separate paths. The analysis of the different schools of thoughts in this field of research is the focus of the Chapter 2 of this dissertation so I will not spend too much time on this period here. However, it is still interesting to note that most of the different schools have progressively paid more attention to the relations between cities. In the 2000s, an important body of literature on networks of multinational companies between world cities emerged. Although not based on actual relational data (the relations are derived from the presence of sites of a company in different cities), these study incorporated some methodological aspect of graph theory (Taylor, 2001). The same holds for studies of urban systems at the regional scale that were based on interurban commuting and firm networks (De Goei et al., 2010; Green, 2007; Taylor et al., 2008). However, actual engagement with the emerging interdisciplinary field of network science remained rather limited during this decade (Derudder and Neal, 2018; Ducruet and Beauguitte, 2013).

It is truly over the last decade that network approaches to urban systems skyrocketed. This can be related to the fact that the formalisation of network analysis has been chosen by quantitative urban scholars as the main language to describe and analyse urban interactions at multiple spatial scales (Batty, 2013; Lobo et al., 2020; Neal, 2013). Complex network analysis has spread to already existing areas of research such as the analysis of firm ownership networks (Rozenblat et al., 2016; Zhao et al., 2017), transportation between cities and urban regions (Ducruet et al., 2018, 2011) and commuting and travel data (Nelson and Rae, 2016; Poorthuis and Meeteren, 2019). It has also been used to study interactions captured by new urban ‘Big data’ (generated by social media, mobile devices, crowd-sourcing, and sensors) allowing to map communications between cities in an entire country (Krings et al., 2009), or individual human mobility at the regional or global scale (Lenormand et al., 2015; Zhang et al., 2016). Finally, concepts from the interdisciplinary field of network science such as “multiplexity”[7] are increasingly being used to simultaneously study the multiple kinds of interactions connecting cities (Berroir et al., 2020; Burger et al., 2014).

7 A multiplex network, or multilayer network, is a network with multiple types of relations.

1.3 Data availability and progress in urban system research

Data availability and technological developments have always played a crucial role in the progress of urban system research. In the 1960s and 1970s, researchers benefited from the improvement of census methods by gaining access to detailed geographical data (Wheeler, 2001). These new datasets were used to produce novel analyses on city size distributions and urban functions and specialisation at the scale of the entire American urban system. Such analyses were also made possible thanks to the diffusion of mainframe computers in university campuses and the development of computer-assisted cartography.

Similarly, the revitalisation of the field that happened over the last two decades has to be put in a similar context of technological revolution. The rapid evolution of information and communication technology (ICT) during the recent period has opened up new opportunities for researchers. This revolution is characterised by increasing computing power, possibilities to automatically collect data at a very large scale, and the development of capabilities for storing large amounts of data and processing it. The datasets produced via these technologies are usually referred to as "big data". Their main characteristics are that they are big in *volume*, high in *velocity* (in real-time or near real-time) and big in *variety*, meaning the data comes from various decentralised sources and is usually semi-structured (Katal et al., 2013). Many of these datasets have geographical (and temporal) components, notably the ones generated by locational aware technologies such as sensors attached to mobile phones and vehicles, credit and transport cards, geo-referenced social media and web-based services (Miller and Goodchild, 2015).

With the skyrocketing of data availability, disciplines that were not especially engaged with geographical research started to be more and more interested in urban studies, including urban system analysis. According to Pumain (2020), debates on urban theories that were usually restricted to the social sciences have been revitalised by scholars from physics, applied mathematics and computer sciences. Researchers from these disciplines were among the first to work with 'big data', and notably with the large network datasets generated by the ICT revolution that has made the emergence of the interdisciplinary field of network science possible (Barabási and Pósfai, 2016). During the 2000s, these researchers have progressively integrated space in the analysis of networks (Barthelemy, 2011), but actual collaborations between natural and social scientists remained limited until very recently (Ducruet and Beauguitte, 2013; O'Sullivan and Manson, 2015).

The use of 'big data' in geographical research has led to considerable progress both from a methodological and thematic point of view. Looking back at the statements on the data scarcity issue made in the 1990s (i.e. in Short et al., 1996), one could argue that some of them have been at least partially solved with the development of big data. A problem often mentioned in the case of studies on relations between cities at a transnational scale was the lack of comparative data on cities because data was mostly provided by national statistical agencies using their own categories. In their paper, Short et al., (1996) listed a series of relevant indicators that could not be fulfilled because of the lack of empirical material. One example is the identification of transport nodes in the global urban system. Nowadays, thanks to the availability of affordable geolocated ICT services, large amounts of data on how people move are generated in real time. These digital traces generated by individuals can be used to reconstruct mobility networks at the global scale. Such an effort has been done by Lenormand et al. (2015) using Twitter data. In their paper, they have analysed millions of geolocated social media posts in order to map human mobility and measure the influence of major cities at different spatial scales. Thanks to the nature of Twitter data, they could develop an indicator that was not focused on a specific mode of transport and available beyond national boundaries. According to them "data (…) generated from online social media as Twitter, Flickr or Foursquare can refer to the whole globe". Although this statement has to be nuanced because some states block some of these platforms, such analysis would not have been possible with traditional sources.

Another important novelty is the possibility of working on very localised real time events (Adam, 2019), something that was impossible before the advent of big data (Shearmur, 2015). Very recently, such approaches have demonstrated their usefulness to map the tremendous population changes in the French and Indian urban systems related to the announcement of the lockdown measures to slow the spread of the Covid-19 pandemic (Denis et al., 2020; Pullano et al., 2020). Using data from mobile phone operators for France, and from Facebook logins for India, research has highlighted a flight from big cities to less dense areas and pinpointed potential new clusters of the pandemic.

However, these new sources of data are not without limitations. Big data generally refers to entire groups of populations, however these populations are subject to self-selection issues. Many studies on the demographics of social media users have highlighted important biases in terms of representativeness. For example, in the case of geo-referenced tweets in the United-States, it has been shown that more twitter users are more likely to be found in urban areas, areas with young populations (18-29) and ethnic minorities (Malik, 2018). Some other studies have highlighted biases in terms of gender and level of education of social media users (Greenwood et al.,

2016; Sloan et al., 2015) and there are regional variations of these biases (Jiang et al., 2019). This means that one has to be very careful before making inferences about larger populations. This issue of representativeness is largely due to the fact that these datasets are not meant to study the society but markets and users (Shearmur, 2015).

Another major issue with big data is their lack of consistency through time. As stated by Kitchin (2013): "What data are captured is shaped by the technology used, the context in which data are generated and the data ontology employed." This has two implications for longitudinal studies. First, big data is usually generated by technologies that are adopted and abandoned at a very high pace. In the second half of the 2000s, researchers working on urban systems have measured the level of contact between cities using data on messages and phone calls between cities (Krings et al., 2009) or redrawn functional regions based on this data (Ratti et al., 2010). However, one decade later, this data would not have the same level of exhaustively because a huge part of these contacts are now going through the internet via encrypted instant messaging applications. Second, because this data is usually collected for short-term operational use, its objects and categories are often very fuzzy (Shearmur, 2015). In opposition with census categories that are developed for the long-term purpose.

In the field of urban system research, where both exhaustiveness and longitudinal approaches are central, it is very unlikely that big data will provide all possible answers. Recent studies have shown that more traditional sources processed and analysed with modern computational techniques show great promises for the analysis of systems of cities at multiple spatial scales. Researchers have, for example, explored the impact of 'external linkages' in the innovation capacity of cities with a large-scale analysis of patent databases (Breschi and Lenzi, 2016; Capone et al., 2019), changes in the interurban patterns of scientific collaboration with bibliometric data (Maisonobe et al., 2016), and the recent reconfiguration of interurban relations though multinational firms' ownership linkages with company registers (Rozenblat, 2018). Another example of a computational approach using traditional data is the work of Ducruet et al. (2018) that digitised the archives of a quasi-monopolistic shipping insurance company to reconstruct networks of maritime trade between cities from 1890 to 2010. All these studies have shown the possibility of assembling large amounts of data on interurban relationships with stable categories over time and with a more careful attention for the representativeness of the data.

Progress in urban system research have been often connected to increase in data availability and new methodological developments. Over the recent period, one of the biggest innovations has been to work on the "real time city" thanks to the very

fine spatial and temporal granularity of big data. Although data-driven studies of cities have led to the creation of new knowledge on urban systems, they also present important limitations, notably regarding representativeness issues and consistency over medium- and long-term periods. Traditional sources such as archives, censuses, and administrative registers should not be overlooked. This is especially true since computational methods allow researchers to analyse them in novel ways.

1.4 Research approach

1.4.1 Aim and research questions

The starting point of this PhD research was the observation that the analysis of urban dynamics, and, in particular, the role of networks in the explaining growth and decline of cities, is limited by a lack of consistent data on the relations between cities. By consistent, I mean that the data is available in a stable form for a long period of time and across the entire territory of a system of cities. Computational approaches of archives and registers have shown great promise to fulfil this requirement. In this PhD research, these approaches are explored further by focusing on the Netherlands. The aim is to develop new ways of measuring relations between cities that are consistent through time and over the whole Dutch territory. These explorations are done with a focus on three datasets: two text archives (the CommonCrawl and Delpher) and the register data from the Social Statistics Database provided by Statistics Netherlands (CBS).

In a first experiment, the potential of mining text corpora is explored to identify patterns of interurban relationships in a contemporary setting by looking at co-occurrences of city names in web pages. The objective is to upscale the method pioneered by Tobler and Wineburg (1971) to the entire Dutch urban system with a contemporary text corpus.

RQ1: To what extent can the toponym co-occurrence method be used to identify the patterns of relations between cities in a systematic way?

To respond to the issue about time consistency, this research was inspired by classical urban system research that looked at the circulation of news between cities (Pred, 1976; Zipf, 1946). Such data sources were considered as potential answers to the data scarcity issue in urban system research (Beaverstock et al., 2000). However, because of the cost of data collection, these attempts remained limited to small samples of cities or short period of times. In order to develop this approach further, the following questions were posited:

RQ2: How can we extract data on information circulation between cities for a long period of time from a massive archive of historical newspapers?

RQ3: To what extent is data extracted from historical newspaper archives reflecting the long-term evolution of the territorial organisation of a system of cities?

Regarding contemporary urban dynamics, it has been argued that urban boundaries tend to blur as urbanisation intensifies, which implies that intra- and interurban processes are sometimes difficult to differentiate. Such questions are even more critical in the context polycentric urban regions (PURs), the dominant form of urbanisation in the Netherlands. For this reason, I explore the potential of using individual level data on individual job and residential career histories to better identify the different scales at which urban systems form:

RQ4: To what extent can longitudinal individual data be used to build bottom-up definitions of cities and urban systems?

1.4.2 Geographical focus

The geographical focus of this PhD research is on the Netherlands. This focus has been chosen for three main reasons: access to data, the intrinsic structure of the Dutch urban system and affinity with the study area.

The choice for the Netherlands is quite opportunistic, and based on the fact that there was an opportunity to access unique datasets covering the entire Dutch territory, which seemed like good possible candidates to derive consistent longitudinal measures of interurban relations from. One of these datasets is the digital archive of historical newspapers assembled by the National Library of the Netherlands (KB) which is accessible through the Delpher platform. The availability of a large quantity of searchable news items from the past makes it possible to extract information flows between cities. Beyond Delpher, this thesis also benefited

from the work of quantitative historians that digitised the Dutch historical census to confront the data extracted from newspapers with dynamics of urban populations. The access to data and interactions with collection specialists have also been enhanced by the fact that I have been selected for the researcher-in-residence program of the KB for a period of 6 months. For contemporary dynamics, there was an opportunity to access microdata from the Social Statistics Database provided by Statistics Netherlands, that is built from the crossing of several administrative registers and allows for work on highly detailed longitudinal and geocoded data on population dynamics.

Second, the settlement structure of the Netherlands presents interesting features for the study of relations between cities. The Netherlands is a highly urbanised country and the majority of its territory is located in the European megalopolis. According to the last OECD/UE definition of cities, it contains 37 core cities of more than 50,000 inhabitants that are integrated within 25 functional urban areas with the biggest one (Rotterdam-The Hague) being close to 3.5 million inhabitants. One particular aspect of this urban system is the presence of big population centres close to each other in the central part of the country. This polycentric urban region, at which level integration is often discussed, gathers Amsterdam, Rotterdam, The Hague, Utrecht, Leiden, Delft and Dordrecht among other urban centres. The reflection on relations between cities has a long tradition in the Netherlands, not only in academic debates. The reflection on urban networks and relations between cities is indeed an important focus of the planning debate for several decades, as it is assumed that network externalities and processes of borrowed size between the cities can substitute a single large metropolis.

The last reason is that geographical research cannot be done without concrete, tangible connections with its research object. As this work was done in the Netherlands, it was much easier to confront the results of the approaches developed throughout this research with local expertise and a sense of the past and present urban dynamics.

1.5 Outline of the thesis

Chapter 2 of this dissertation (published in *Networks and Spatial Economics*) presents the evolution of urban systems literature over the last two decades. In order to give a bird's eye view of the field, it draws on an innovative methodology coupling citation network analysis and text mining.

Chapter 3 (published in *International Journal of Urban Sciences*) answers *RQ3* by applying the toponym co-occurrence method on a corpus of websites mentioning city names in order to identify the pattern of relations between places in a systematic way.

Chapter 4 (published in *Cybergeo: European journal of Geography*), answers *RQ1* and presents the method developed to build a dataset on information circulation between cities from the historical newspapers contained in the Delpher database. It details the identification of a relevant corpus of news on places, the Natural Language Processing (NLP) methods used to identify place names in raw text data, and reports the steps for the validation of the data collection.

Chapter 5 (published in *Computers, Environment and Urban Systems*), answers *RQ2*, by testing the potential of the novel data on information flows between cities to reconstruct the evolution of the Dutch system of cities over a period of 60 years. For a search in regularities in how flows of information develop over time and a testing of different hypotheses related to the evolution of urban systems is completed here.

Chapter 6 answers *RQ4*. It presents an exploration of register data on individual migration patterns and job history in order to derive measures of relations between places at multiple spatial scales by analysing the residential relocation of people that kept the same job. This data is used to question the scales of cities and urban systems in a contemporary setting.

Finally, **Chapter 7** summarises and synthetises the research findings. It outlines the main innovations and limitations of this PhD research and presents ways for further development.

Bibliography

Adam, A., 2019. Exploring new geographies of interactions in and around the metropolitan area of Brussels (phdthesis). Université catholique de Louvain.

Arribas-Bel, D., Reades, J., 2018. Geography and computers: Past, present, and future. Geography Compass 12, e12403. https://doi.org/10.1111/gec3.12403

Auerbach, F., 1913. Das Gesetz der Bevölkerungskonzentration. Petermanns Geographische Mitteilungen 59, 74–76.

Barabási, A.-L., Pósfai, M., 2016. Network Science. Cambridge University Press.

Barthelemy, M., 2011. Spatial Networks. Physics Reports 499, 1–101. https://doi.org/10.1016/j.physrep.2010.11.002

Batty, M., 2013. The New Science of Cities. MIT Press.

Beaverstock, J.V., Smith, R.G., Taylor, P.J., Walker, D.R.F., Lorimer, H., 2000. Globalization and world cities: some measurement methodologies. Applied Geography 20, 43–63. https://doi.org/10.1016/S0143-6228(99)00016-8

Berroir, S., Cattan, N., Dobruszkes, F., Guérois, M., Paulus, F., Vacchiani-Marcuzzo, C., 2020. French urban systems: a relational approach. Traduction. Cybergeo : European Journal of Geography. https://doi.org/10.4000/cybergeo.35587

Berry, B.J.L., 1966. Essays on Commodity Flows and the Spatial Structure of the Indian Economy. Department of Geography, University of Chicago.

Berry, B.J.L., Pred, A., 1961. Central place studies. A bibliography of theory and applications. Central place studies. A bibliography of theory and applications.

Board, C., Davies, R. j., Fair, T. j. d., 1970. The structure of the South African space economy: An integrated approach. Regional Studies 4, 367–392. https://doi.org/10.1080/09595237000185361

Braudel, F., 1979. Civilisation matérielle, économie et capitalisme: XVe-XVIIIe siècle. Librairie générale française.

Breschi, S., Lenzi, C., 2016. Co-invention networks and inventive productivity in US cities. Journal of Urban Economics 92, 66–75. https://doi.org/10.1016/j.jue.2015.12.003

Burger, M.J., Meijers, E.J., van Oort, F.G., 2014. Multiple Perspectives on Functional Coherence: Heterogeneity and Multiplexity in the Randstad. Tijdschr Econ Soc Geogr 105, 444–464. https://doi.org/10.1111/tesg.12061

Capone, F., Lazzeretti, L., Innocenti, N., 2019. Innovation and diversity: the role of knowledge networks in the inventive capacity of cities. Small Bus Econ. https://doi.org/10.1007/s11187-019-00268-0

Cattan, N., 1995a. Attractivity and Internationalisation of Major European Cities: The Example of Air Traffic. Urban Stud 32, 303–312. https://doi.org/10.1080/00420989550013095

Cattan, N., 1995b. Barrier Effects: The Case of Air and Rail Flows: International Political Science Review. https://doi.org/10.1177/019251219501600304

Cattan, N., 1990. Une image du réseau des métropoles européennes par le trafic aérien. Espace géographique 19, 105–116. https://doi.org/10.3406/spgeo.1990.2957

Christaller, W., 1933. Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen. University Microfilms.

De Goei, B., Burger, M.J., Van Oort, F.G., Kitson, M., 2010. Functional Polycentrism and Urban Network Development in the Greater South East, United Kingdom: Evidence from Commuting Patterns, 1981–2001. Regional Studies 44, 1149–1170. https://doi.org/10.1080/00343400903365102

Denis, E., Telle, O., Benkimoun, S., Mukhopadhyay, P., Nath, S., 2020. Mapping the lockdown effects in India: how geographers can contribute to tackle Covid-19 diffusion. The Conversation 7.

Derudder, B., 2019. Network Analysis of 'Urban Systems': Potential, Challenges, and Pitfalls. Tijdschrift voor economische en sociale geografie. https://doi.org/10.1111/tesg.12392

Derudder, B., Neal, Z., 2018. Uncovering Links Between Urban Studies and Network Science. Netw Spat Econ 18, 441–446. https://doi.org/10.1007/s11067-019-09453-w

Dietvorst, A.G.J., Wever, E., 1977. Changes in the Pattern of Information Exchange in the Netherlands, 1967-1974. Tijdschrift voor economische en sociale geografie 68, 72–82. https://doi.org/10.1111/j.1467-9663.1977.tb01397.x

Ducruet, C., Beauguitte, L., 2013. Spatial Science and Network Science: Review and Outcomes of a Complex Relationship. Netw Spat Econ 14, 297–316. https://doi.org/10.1007/s11067-013-9222-6
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. https://doi.org/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. Cybergeo : European Journal of Geography. https://doi.org/10.4000/cybergeo.23603
Green, N., 2007. Functional Polycentricity: A Formal Definition in Terms of Social Network Analysis. Urban Studies 44, 2077–2103. https://doi.org/10.1080/00420980701518941
Greenwood, S., Perrin, A., Duggan, M., 2016. Social media update 2016. Pew Research Center 11, 1–18.
Haggett, P., Chorley, R.J., 1969. Network Analysis in Geography. Edward Arnold.
Hesse, M., 2014. On borrowed size, flawed urbanisation and emerging enclave spaces: The exceptional urbanism of Luxembourg, Luxembourg: European Urban and Regional Studies. https://doi.org/10.1177/0969776414528723
Jiang, Y., Li, Z., Ye, X., 2019. Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. Cartography and Geographic Information Science 46, 228–242. https://doi.org/10.1080/15230406.2018.1434834
Katal, A., Wazid, M., Goudar, R.H., 2013. Big data: Issues, challenges, tools and Good practices, in: 2013 Sixth International Conference on Contemporary Computing (IC3). Presented at the 2013 Sixth International Conference on Contemporary Computing (IC3), pp. 404–409. https://doi.org/10.1109/IC3.2013.6612229
Kitchin, R., 2013. Big data and human geography: Opportunities, challenges and risks. Dialogues in Human Geography. https://doi.org/10.1177/2043820613513388
Krings, G., Calabrese, F., Ratti, C., Blondel, V.D., 2009. Urban gravity: a model for inter-city telecommunication flows. J. Stat. Mech. 2009, L07003. https://doi.org/10.1088/1742-5468/2009/07/L07003
Lenormand, M., Gonçalves, B., Tugores, A., Ramasco, J.J., 2015. Human diffusion and city influence. Journal of The Royal Society Interface 12, 20150473. https://doi.org/10.1098/rsif.2015.0473
Lobo, J., Alberti, M., Allen-Dumas, M., Arcaute, E., Barthelemy, M., Bojorquez Tapia, L.A., Brail, S., Bettencourt, L., Beukes, A., Chen, W.-Q., Florida, R., Gonzalez, M., Grimm, N., Hamilton, M., Kempes, C., Kontokosta, C.E., Mellander, C., Neal, Z.P., Ortman, S., Pfeiffer, D., Price, M., Revi, A., Rozenblat, C., Rybski, D., Siemiatycki, M., Shutters, S.T., Smith, M.E., Stokes, E.C., Strumsky, D., West, G., White, D., Wu, J., Yang, V.C., York, A., Youn, H., 2020. Urban Science: Integrated Theory from the First Cities to Sustainable Metropolises (SSRN Scholarly Paper No. ID 3526940). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3526940
Maisonobe, M., Eckert, D., Grossetti, M., Jégou, L., Milard, B., 2016. The world network of scientific collaborations between cities: domestic or international dynamics? Journal of Informetrics 10, 1025–1036. https://doi.org/10.1016/j.joi.2016.06.002
Malik, M.M., 2018. Bias and beyond in digital trace data (PhD Thesis). Doctoral dissertation, Carnegie Mellon University). Retrieved from http ….
Meijers, E.J., Burger, M.J., 2017. Stretching the concept of 'borrowed size.' Urban Studies 54, 269–291. https://doi.org/10.1177/0042098015597642
Meijers, E.J., Burger, M.J., Hoogerbrugge, M.M., 2016. Borrowing size in networks of cities: City size, network connectivity and metropolitan functions in Europe. Papers in Regional Science 95, 181–198. https://doi.org/10.1111/pirs.12181
Miller, H.J., Goodchild, M.F., 2015. Data-driven geography. GeoJournal 80, 449–461. https://doi.org/10.1007/s10708-014-9602-6
Mumford, L., 1961. The City in History: Its Origins, Its Transformations, and Its Prospects. Harcourt Brace Jovanovich.
Neal, Z.P., 2013. The Connected City: How Networks are Shaping the Modern Metropolis. Routledge.
Nelson, G.D., Rae, A., 2016. An Economic Geography of the United States: From Commutes to Megaregions. PLOS ONE 11, e0166083. https://doi.org/10.1371/journal.pone.0166083
O'Sullivan, D., Manson, S.M., 2015. Do Physicists Have Geography Envy? And What Can Geographers Learn from It? Annals of the Association of American Geographers 105, 704–722. https://doi.org/10.1080/00045608.2015.1039105

Poorthuis, A., Meeteren, M. van, 2019. Containment and Connectivity in Dutch Urban Systems: A Network-Analytical Operationalisation of the Three-Systems Model. Tijdschrift voor economische en sociale geografie n/a. https://doi.org/10.1111/tesg.12391
Pred, A., 1977. City Systems in Advanced Economies: Past Growth, Present Processes, and Future Development Options. Wiley.
Pred, A., 1976. The interurban transmission of growth in advanced economies: Empirical findings versus regional-planning assumptions. Regional Studies 10, 151–171. https://doi.org/10.1080/09595237600185161
Pred, A., Törnqvist, G., 1973. Systems of cities and information flows: Two essays. Royal University of Lund, Sweden, Department of Geography : C.W.K. Gleerup, Lund.
Pred, A.R., 1971. Urban Systems Development and the Long-Distance Flow of Information Through Preelectronic U.S. Newspapers. Economic Geography 47, 498–524. https://doi.org/10.2307/142641
Pullano, G., Valdano, E., Scarpa, N., Rubrichi, S., Colizza, V., 2020. Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study. The Lancet Digital Health S2589750020302430. https://doi.org/10.1016/S2589-7500(20)30243-0
Pumain, D., 2020. Introduction, in: Pumain, D. (Ed.), Theories and Models of Urbanization: Geography, Economics and Computing Sciences, Lecture Notes in Morphogenesis. Springer International Publishing, Cham, pp. 1–9. https://doi.org/10.1007/978-3-030-36656-8_1
Pumain, D., 2011. Systems of Cities and Levels of Organisation, in: Bourgine, P., Lesne, A. (Eds.), Morphogenesis. Springer Berlin Heidelberg, pp. 225–249. https://doi.org/10.1007/978-3-642-13174-5_13
Pumain, D., 2003. La modélisation des réseaux urbains.
Pumain, D., 1997. Pour une théorie évolutive des villes. Espace géographique 26, 119–134. https://doi.org/10.3406/spgeo.1997.1063
Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., Strogatz, S.H., 2010. Redrawing the Map of Great Britain from a Network of Human Interactions. PLOS ONE 5, e14248. https://doi.org/10.1371/journal.pone.0014248
Rozenblat, C., 2018. Urban Systems Between National and Global: Recent Reconfiguration Through Transnational Networks, in: Rozenblat, C., Pumain, D., Velasquez, E. (Eds.), International and Transnational Perspectives on Urban Systems, Advances in Geographical and Environmental Sciences. Springer, Singapore, pp. 19–49. https://doi.org/10.1007/978-981-10-7799-9_2
Rozenblat, C., Pumain, D., 2007. Firm linkages, innovation and the evolution of urban systems. Cities in globalization: Practices, policies, theories 130–156.
Rozenblat, C., Pumain, D., 1993. The Location of Multinational Firms in the European Urban System. Urban Studies 30, 1691–1709. https://doi.org/10.1080/00420989320081671
Rozenblat, C., Zaidi, F., Bellwald, A., 2016. The multipolar regionalization of cities in multinational firms' networks. Global Networks n/a-n/a. https://doi.org/10.1111/glob.12130
Scott, A.J., Storper, M., 2015. The Nature of Cities: The Scope and Limits of Urban Theory. Int J Urban Regional 39, 1–15. https://doi.org/10.1111/1468-2427.12134
Shearmur, R., 2015. Dazzled by data: Big Data, the census and urban geography. Urban Geography 36, 965–968. https://doi.org/10.1080/02723638.2015.1050922
Short, J.R., Kim, Y., Kuus, 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. https://doi.org/10.1111/j.1468-2427.1996.tb00343.x
Simmons, J.W., 1970. Interprovincial Interaction Patters in Canada. Research Paper.
Sloan, L., Morgan, J., Burnap, P., Williams, M., 2015. Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10, e0115545. https://doi.org/10.1371/journal.pone.0115545
Smith, D.A., Timberlake, M., 1995. Cities in global matrices: toward mapping the world-system's city system. World cities in a world-system 79–97.
Taylor, P.J., 2004. World City Network: A Global Urban Analysis. Psychology Press.
Taylor, P.J., 2001. Specification of the world city network. Geographical analysis 33, 181–194.

Taylor, P.J., Evans, D.M., Pain, K., 2008. Application of the Interlocking Network Model to Mega-City-Regions: Measuring Polycentricity Within and Beyond City-Regions. Regional Studies 42, 1079–1093. https://doi.org/10.1080/00343400701874214

Tobler, W., Wineburg, S., 1971. A Cappadocian Speculation. Nature 231, 39–41. https://doi.org/10.1038/231039a0

Törnqvist, G., 1968. Flows of Information and the Location of Economic Activities. Geografiska Annaler. Series B, Human Geography 50, 99–107. https://doi.org/10.2307/490320

Vapnarsky, C.A., 1969. On Rank-Size Distributions of Cities: An Ecological Approach. Economic Development and Cultural Change 17, 584–595.

Wheeler, J.O., 2001. Assessing the role of spatial analysis in urban geography in the 1960s. Urban Geography 22, 549–558. https://doi.org/10.2747/0272-3638.22.6.549

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 0, 1–21. https://doi.org/10.1080/10630732.2016.1177259

Zhao, M., Derudder, B., Huang, J., 2017. Examining the transition processes in the Pearl River Delta polycentric mega-city region through the lens of corporate networks. Cities 60, 147–155. https://doi.org/10.1016/j.cities.2016.08.015

Zipf, G.K., 1949. Human behavior and the principle of least effort: an introduction to human ecology. Addison-Wesley Press.

Zipf, G.K., 1946. Some Determinants of the Circulation of Information. The American Journal of Psychology 59, 401–421. https://doi.org/10.2307/1417611

TOC

2 The evolution of the systems of cities literature since 1995

Schools of thought and their interaction

This is the Author's Accepted Manuscript version (final draft post-refereeing as accepted for publication by the journal). The definitive, peer-reviewed and edited version of this article is published as: Peris A., Meijers E. & van Ham M. (2018) The Evolution of the Systems of Cities Literature Since 1995: Schools of Thought and their Interaction. Network and Spatial Economics, DOI: https://doi.org/10.1007/s11067-018-9410-5

ABSTRACT The study of relations between cities has long been a major focus in urban research. For decades, this field has grown integrating contributions from many disciplines. But today, the field appears rather fragmented. This study analyses the body of literature that has developed over the last 23 years to identify schools of thought on interurban relationships and to see to what extent these interact with each other. It does so by innovatively employing bibliometric analysis to the study of systems of cities, which allows a bottom-up identification of five schools of thought: one predominantly focusing on the regional or intra-metropolitan scale and centred on concepts of polycentricity; one addressing the global scale with a focus on world city networks; one employing simulation and complexity theories to understand behaviour of agents building the urban system bottom-up; one rooted in (new) economic geography and focusing on growth and decline in the urban system; and, one seeking regularities with respect to city size distributions. The conceptual, methodological and empirical aspects of these different schools are discussed by

means of a ‘semantic map’ derived from the vocabulary of titles and abstracts of papers. The coupling of the semantic map with the citation networks of these schools of thought confirms the increasing fragmentation of the field over the last decades. However, in the most recent years, the different schools of thought start to interact slightly more. The desirability and feasibility of a move from multidisciplinarity to interdisciplinarity in urban systems research needs further exploration.

KEYWORDS Urban system, System of cities, Urban network, Bibliometric analysis, Text mining

2.1 Introduction

Cities do not function in isolation, but are organised in systems of cities characterised by strong interdependencies that develop at the scale of a large region, a nation, a continent or even at the global scale (Pumain 2011). A large literature on interrelated cities has developed since the end of the 19th century. Early contributions include work observing the regularities in the size distribution of cities in countries (Auerbach 1913; Gibrat 1931; Zipf 1949) as well as the formulation of central place theory (Christaller 1933). These contributions provided the basis for an upsurge of work on intercity relationships in the 1960s and 1970s, addressing many aspects of a system of cities such as the size, location and specialisation of cities as well as the uneven circulation of people, goods and information among them (Berry, 1964; Bourne and Simmons, 1978; Pred, 1977). The definition of a ‘system of cities’ by Allan Pred (1977, p.13) is still valid today: “a national or regional set of cities that are interdependent in such a way that any significant change in the economic activities, occupational structure, total income or population of one member city will directly or indirectly bring about some modification in the economic activities, occupational structure, total income or population of one or more other set members”. Nowadays, this definition can also be extended to global urban systems because of long-distance interrelationships between cities, particularly those at the top of national urban hierarchies becoming more common.

Since the 1990s the literature on systems of cities has developed further and expanded, but the current landscape of research appears rather fragmented. Increasingly the term ‘paradigm change’ is used by researchers willing to position themselves in opposition with ‘classical’ approaches. For example, Friedmann (1995) talked about a ‘world city paradigm’, by which he meant an encompassing approach of different aspects of intercity relations at the global scale, which tended to be

studied separately. Also Capello (2000) and Meijers (2007) suggested a 'paradigm change', claiming that the classical Central Place model was unable to describe contemporary trends in the pattern of intercity relations. More recently, Batty (2013) wrote about the rise of a 'new paradigm' in the conception of cities. Building on previous works that consider cities as emerging from the multitude of interactions between individuals, he underlines that processes of centralised decision-making such as planning and governing have a limited influence on cities. These different theoretical positions have an important impact on the research approach. While some studies focus more on stakeholders (Alderson and Beckfield, 2004; Sassen, 1991), others look at the emergent properties of a system of cities by considering the basic interactions between urban agents, for instance in a simulation framework (Sanders et al, 1997) or in the methodological individualism of economics (Fujita et al., 1999). There are also different positions regarding the scale at which the most important urban processes take place. For some researchers, in the context of globalisation, the global scale has become most determinant (Taylor and Derudder, 2015). For others, the erosion of national borders in this context puts the regional scale at the centre of economic processes (Kloosterman and Musterd, 2001; Parr, 2014). Somewhere in between these scales is a research stream stressing the importance of the national scale given its determining influence on many structures and parameters that experience strong path dependencies (Pumain 1997; Bretagnolle and Franc 2017). Differences in ontological and epistemological perspectives translate into wildly varying objectives of research, ranging from identifying universal laws of urbanisation (Bettencourt et al., 2007) to much more policy oriented studies (Meijers and Romein, 2003). This variety in objectives partly relates to the disciplinary background and sources of influence of the researchers. While theoretical and quantitative geographers and physicists will look for basic mechanisms, planners will aim to give policy recommendations. As the field exploring relationships between cities has received contributions and influences from many different disciplines such as geography, regional science, sociology, economics, physics and the interdisciplinary movement of complexity theories, it seems that increasingly separate approaches or subfields have emerged.

This paper aims to answer the following questions: How did the system of cities literature evolve over the last two decades? Which different schools of thought can be distinguished and what are their defining elements? And, to what extent do these schools interact? Assessing interdisciplinarity in this research field is all the more important given the frequent calls for interdisciplinarity in urban systems research (Pflieger and Rozenblat, 2010), and because there is clear evidence that innovation in geography – still the main discipline addressing systems of cities - is fostered by collaborations among disciplines (Ducruet and Beauguitte, 2013).

This study of the evolution of the urban systems literature does not take the form of a classical literature review paper. Rather it adopts a bibliometric approach to analyse a set of 1,491 papers on intercity relationships from 1995 onward. The main advantage of this method is that certain bias in reviewing the literature can be avoided, such as a too narrow disciplinary point of departure. Indeed, during our readings, we noticed that certain studies were ignoring entire parts of the field. By limiting human intervention on the collection of the set of papers, we intend to overcome this bias and show the diversity of the field and its complex internal structure. This does not mean that our approach can replace extensive readings, at the contrary, but it allows to get a 'bird's eye view' for further exploration. Our approach is inspired by the hyper-network approach, which combines the analysis of semantic and citation networks. This approach has for instance been applied recently to the papers of a journal (Raimbault, 2017), or to the classification of a large set of patents (Bergeaud et al., 2017). The approach entails a two-step approach. First, following Chavalarias and Cointet (2013) who define scientific fields "as sets of 'keywords' delineating a research area", a semantic network based on co-occurrences of words in the titles and abstracts of the set of papers is extracted. This allows then to identify the different subfields or schools of thought on systems of cities. Second, the pattern of citations of the papers developed within these schools are analysed to understand the connections between the different subfields.

The following section presents and discusses the bibliometric method as a way to undertake a literature review, as well as the delineation procedure that is needed to select a relevant corpus of texts (section 2). In the subsequent section, we present the content-analysis based on the vocabulary of the papers and how differences in vocabulary allow to identify different schools of thought. In addition, we explore the relations between these different schools of thought by analysing the evolution of citation patterns (section 3). The last section concludes and discusses the implications of our findings (section 4).

2.2 Bibliometric analysis and corpus delineation

2.2.1 Bibliometric analysis in social science

Citation networks do not only reveal intellectual connections, but also the social organization of science (Leydesdorff 1998). This dimension of science stands central when studying the formation of concepts but can be enriched by the analysis of the text related to the production of knowledge (Callon et al., 1983). Words have a central place in science because scientists are first of all readers and writers (Latour and Woolgar, 2013). Consequently, to study the evolution of a notion such as systems of cities, we can, in addition to exploring citation patterns, also study the semantic network extracted from the set of papers. Using a mixed approach combining citation and text analysis such as the recently developed hyper-network framework allows characterising a corpus more precisely (Raimbault 2017). The semantic network can be obtained through text mining techniques applied to titles and abstracts of the studied papers. Hence, the first step of this research is to extract the vocabulary of scientific publications on system of cities to see which concepts are generally associated with each other. Chavalarias and Cointet (2013) have shown that the vocabulary of publications can be used to study the evolution of a scientific field. Rather than using predefined categories, this approach allows a "bottom-up" reconstruction of science. The basic hypothesis behind co-word analysis is that two words co-occurring often within individual papers will have a great probability of being strongly related (Chavalarias and Cointet, 2013). After this semantic analysis, a citation network analysis can be used to see whether papers with a similar approach are embedded in very homogeneous clusters or not, and/or whether there are exchanges between different schools of thought.

The Scopus database indexes most social science journals and was therefore chosen as point of departure for defining the body of literature to analyse. This implies that we focus on papers published after 1994, because the Scopus database does not systematically provide information preceding this date. The content analysis has been processed and visualised with VOS-viewer (Van Eck and Waltman, 2011), a computer program that was developed at the Centre for Science and Technology Studies of Leiden University and that is freely available[8]. For the creation of the citation network and the creation of a hybrid citation-semantic network we relied on R and notably the *igraph* package (Csardi and Nepusz, 2006).

Although our approach can lead to new insights on the systems of cities literature, the approach also has a number of limitations. First of all, we cannot claim that we are dealing with an all-encompassing, exhaustive set of papers on systems of cities for several reasons other than the absence of pre-1995 papers. As this field is predominantly enriched by social scientists, many contributions, notably some of them published in books are missing. According to Hicks (1999) social scientists publish more in books than natural scientists, resulting in a smaller coverage of the outputs of social science disciplines in journal-based bibliometric databases. However, according to a more recent case study on research outputs in Flanders (Engels et al., 2012) the number of book publications remains rather stable in social science, but their share is diminishing because of the increase in journal publications. Even if Scopus sometimes includes book chapters, we have to accept that some contributions are missing. Moreover, there is a big chance that papers published in English are overrepresented in our corpus because we did the query in this language. Somewhat alleviating the problem is the fact that the Scopus database indexes most of the non-English literature with an English title and abstract. In addition, the focus on just one language avoids the tricky issue of the translation of scientific concepts. Nonetheless, we do think that analysing a large corpus of publications can bring new insights to the field because it covers a very significant part of the scientific production in the given period. The fact that the set of publications was collected by predominantly using a key-word strategy rather than by just climbing up or going down the chain of citations allows to avoid the teleological bias of classical literature reviews. This approach, of course, does not replace the fundamental work of extensive readings but allows framing the literature in a novel way.

8 http://www.vosviewer.com/

2.2.2 Delineating the corpus

The collection of the corpus of relevant publications on systems of cities is a very important and sensitive step because it potentially has a strong influence on the outcomes of the analysis. There is no consensus on the best approach to delineate a scientific field and collect related publications. In practice, three main strategies tend to be used: the key-words strategy (Meeteren et al., 2015) where the set is obtained by collecting all the papers mentioning some chosen key-words; the journal-level strategy (Leydesdorff and Zhou, 2007; Liu, 2005) that supposes that specific scientific areas are covered by a limited number of journals; and, the citation-based strategy (Waltman and van Eck, 2012) which supposes that scientific fields can be conceived as clusters of individual publications citing each other. According to (Zitt 2015) "mixed strategies with learning processes, adaptive queries and multistep protocols, with possible combination of supervised and automatic stages" are welcomed in bibliometric studies and information retrieval. This is especially the case with interdisciplinary fields that are not necessarily institutionalised such as the system of cities literature. For example, previous analysis of the 'urban studies' literature have been based on mixed journal-level/key-words strategies (Kamalski, Kirby, 2012; Wang et al., 2012).

Social sciences and humanities often deal with complex notions that can have multiple meanings and expressions. This is also the case here: the most common expressions used in the studies of a set of interdependent cities are "system of cities", "urban system", "city-system", "urban network" and "city network". Some research included in this study, particularly those dealing with city size distributions, has no explicit mention of intercity relationships but focuses on stock data. However, we do not draw a clear line between those studies and the studies that adopted a more relational approach, as both approaches overlap, intersect and complement each other. This comes forward in the original formulation of the 'Zipf law' that is supposed to describe "the relative population sizes of the communities of the total system", which would be the result of the two opposite forces of diversification and unification: "the actual location of the population will depend upon the extent to which persons are moved to materials and materials to persons in a given system" (Zipf, 1949, p. 352). This model supposes an underlying relational conception of cities that exchange materials but also migrants (p. 359). In fact, many classics have associated the form of the size-distribution of cities with the level of integration of the system (Vapnarsky 1969; Pred 1977). This applies also to work employing scaling laws, as most of them compare cities belonging to a coherent geographical entity that can be considered a system.

A particularly ambiguous expression of intercity relationships is 'urban system'. In the late 1970s, Allan Pred (1977, p. 219) was already underlining the "inconsistent connotation" of the 'urban system' term that was used in connection to both an individual city and to a set of cities. This explains why during our first attempts of corpus delineation we were confronted with a vast array of papers addressing intra-urban infrastructure networks (water, electricity, roads) and the urban metabolism (a research stream focusing on the material flow analysis of a city). Since these generally had nothing to say about inter-urban relations, we adopted a multistep process with adaptive queries, refined after each iteration and mixed with the analysis of citation patterns. The different steps taken are schematised in Figure 2.1 and are described below.

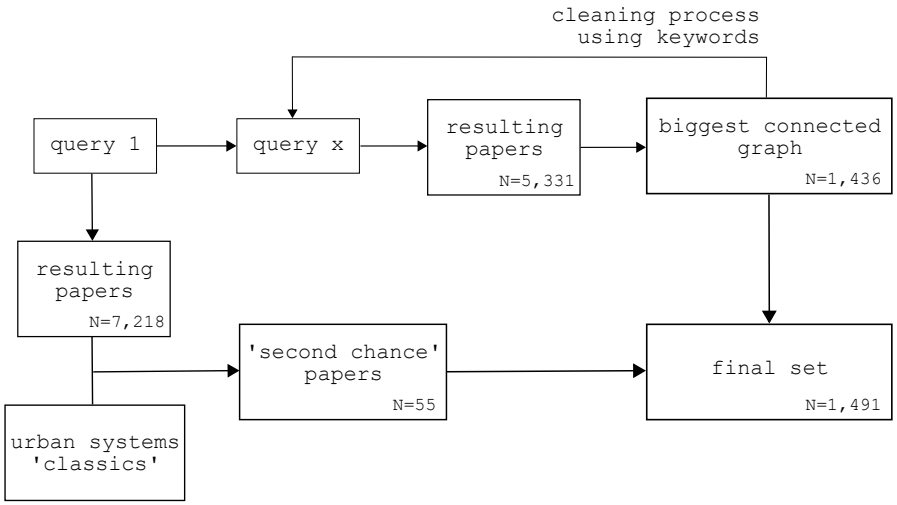

FIG. 2.1 The algorithm of the delineation procedure.

We first mined the references of existing literature reviews (Ducruet and Lugo, 2013; van Meeteren, 2016; Pumain, 2006) to select relevant keywords on the subject. This first step allowed identifying a large spectrum of the field because they cover different aspects of the literature. We then added other terms based on our knowledge of the subject.

TOC

1 *TITLE-ABS-KEY("city system*" OR "urban system*" OR "system* of cities" OR "city network*" OR "urban network*" OR "network* of cities" OR "settlement system*" OR "system* of settlements" OR "central place system*" OR "megaregion*" OR "polycentric urban region*" OR "urban hierarch*") AND PUBYEAR > 1994*

The first query resulted in a set of 7,218 documents. We visualised them as citation networks where publications are nodes and citations are edges. It returned a 'Saturn' type of configuration with a big connected graph of publications in the centre and a ring of unconnected publications surrounding this centre. We took a random sample of 100 publications from the outer-ring to read their title and abstract. The vast majority of them were not about our subject so we decided to keep only the biggest connected graph in the centre. We then took a random sample of 100 publications in this graph and read their title and abstract. The first counting resulted in 36 relevant publications and 64 irrelevant ones. Following (Milanez et al., 2016) we decided to exclude some publications with an iterative cleaning procedure that excluded papers containing certain keywords that are not relevant in for our field, by adding them in the query with the "AND NOT" operator. After each iteration when new words were added to the query, we checked random samples of 100 papers that were part of the biggest connected graph to ensure that they were rightly kept and a similar sample of the papers that were now excluded (ending up in the 'ring of Saturn') to check whether this was fair. When, after several iterations, we established that 95% of the papers in these samples were correctly included or dismissed, we considered our set as sufficiently adequate to scrutinize further. The final query is the following:

2 *TITLE-ABS-KEY("city system*" OR "urban system*" OR "system* of cities" OR "city network*" OR "urban network*" OR "network* of cities" OR "settlement system*" OR "system* of settlements" OR "central place system*" OR "megaregion*" OR "polycentric urban region*" OR "urban hierarch*" AND NOT "dispute settlement system*" AND NOT "traffic control" AND NOT "urban metabolism*" AND NOT "urban ecosystem*" AND NOT "parking*" AND NOT "smart cit*" AND NOT "urban traffic" AND NOT "space syntax" AND NOT "flood*" AND NOT "land use change*" AND NOT "urban ecology" AND NOT "hazard*" AND NOT "emergy" AND NOT "sewage" AND NOT "nitrogen" AND NOT "sensors" AND NOT "mobile landscapes" AND NOT "radial major roads" AND NOT "carbon metabolic network" AND NOT "route perception" AND NOT "waste" AND NOT "healthy cities") AND PUBYEAR > 1994*

This query returns a set of 5,331 papers. After extracting the biggest connected graph in the centre and excluding the 'ring of Saturn', we ended up with a set of 1,436 publications. During the manual check of the excluded publications, we observed that some of them should have been included based on their title and abstract. Most of them were publications citing the classical urban systems literature

but not the contemporary one. We then designed a 'security net' for these cases. We based it on the 'referencing structure function' (Zitt and Bassecoulard, 2006), which is the "fraction of the literature which can be retrieved under two interplaying constraints: a minimum threshold on citation scores for the cited repertoire *Y*, and a minimum closeness of the article with this repertoire, measured by the number *X* of references in common with this repertoire." We used $Y = 15$ for English literature and $Y = 8$ for non-English literature, considering that a non-English paper cited 8 times in the corpus is as least as important as an English paper cited 15 times. We kept only the references published before 1995 as 'classical urban systems literature'. For the number of references in common, we set $X > 1$ because these very central studies are often cited by works outside or at the border of our field. We assume that a text citing at least two of them will be relevant for our study. These publications cited by the biggest connected graph can be considered as 'urban systems classics'.

In the Scopus database, the same reference can reappear with several slightly different spellings, making the citation score of some references artificially low because they are separated in several entries. To give an example, Allan Pred's *City-systems in advanced economies* (1977) appears in 11 different forms, all of them with a citation score below 12. For each single entry in our data, we extracted a sequence made up of the Author's names, the title of the reference, the year of publication, the Journal, the issue and the page numbers and measured the Levenshtein distance between them. This string metric gives the number of characters that have to be deleted, added or substituted to go from a sequence to another. This operation allowed us to identify when the same publication was listed several times with a slightly different spelling, and aggregated their citation score together. We manually sorted them to keep only texts about cities and not general social science or statistical books (Table 2.1).

TABLE 2.1 The 'urban systems classics'.

Author(s)	Year	Title
Auerbach, F.	1913	Das Gesetz der Bevolkerungskonzentration
Zipf, G.K.	1949	Human Behavior and the Principle of Least Effort
Lösch, A.	1954	The Economics of Location
Gottmann, J.	1961	Megalopolis, The Urbanized Northeastern Seaboard of the United States
Berry, B.J.	1964	Cities as systems within systems of cities
Christaller, W.	1966 (1933)	Central Places in Southern Germany
Jacobs, J.	1969	The Economy of Cities
Tobler W.	1970	A computer movie simulating urban growth in the Detroit region
Alonso, W.	1973	Urban Zero Population Growth
Henderson, J.V.	1974	The sizes and types of cities
Pred, A.,	1977	City Systems in Advanced Economies
Rosen, K., Resnick, M.	1980	The size distribution of cities: an examination of the Pareto law and primacy
Friedmann, J., Wolff, G.	1982	World city formation: an agenda for research and action
Pumain, D.	1982	La Dynamique des Villes
Friedmann, J.	1986	The world city hypothesis
Krugman, P.	1991	Geography and Trade
Krugman, P.	1991	Increasing returns and economic geography
Sassen, S.	1991	The Global City
Glaeser, E., Kallal, H.D., Scheinkman, J., Shleifer, A., S	1992	Growth in cities
Camagni, R., Salone, C.,K	1993	Network urban structures in northern Italy: elements for a theoretical framework
Batty, M., Longley, P.	1994	Fractal cities: a geometry of form and function
Sassen, S.	1994	Cities in a World Economy

Then we extracted from the excluded papers all the texts citing at least two of these 'urban system classics' to add them to the final set. This operation allowed reintroducing 55 relevant texts into the final set, leading to a total of 1,491 publications. Figure 2.2 describes the set of papers and book chapters. The production of publications related to systems of cities increased globally over the period, especially from 2010 onwards, with a peak in 2014. The data collection has been realised in October 2017, which explain the low score for the year 2017. The indexation of publications can sometimes take several months. But we decided to keep 2017 incomplete to grasp very contemporaneous trends. In terms of languages, English is largely dominant; recall that its importance was probably increased by the use of an English query. French, Chinese, Spanish and German are also important languages in the literature.

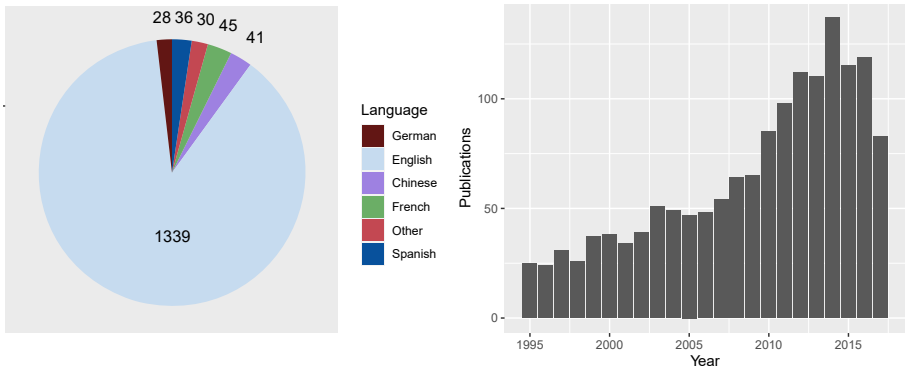

FIG. 2.2 Languages and year repartition of the set of papers and book chapters

2.3 Analysing the system of cities literature

2.3.1 The vocabulary of the urban system literature

Having finalised our set of relevant papers, we now turn to the analysis of the vocabulary of these papers to identify schools of thought with text mining techniques. We first did a graph of co-occurrences of keywords using VOSviewer based on the title and abstract of the 1,491 publications in our set. This software allows to work on noun phrases rather than simple words. It means that the words that are used systematically together will be a single node in the graph (i.e. “cellular automata”, “world city”). Phrases need to have an occurrence score $O \geq 10$ to be in the graph (to keep the Figure readable). The software computes a relevance score T, which corresponds to a tf-idf score for each noun phrase (van Eck and Waltman 2014). A tf-idf score is a statistic that signals the importance of a term in a collection of documents. It returns a high score for the very specific terms, and a low score for the stop-words and also the standard wording of the scientific literature (i.e. “this paper analyses”, “interesting result”). We kept the 60% of the noun phrases with the highest score T. We deleted terms with $O \geq 40$ and $T < 0.7$ because these two thresholds allowed us to remove the ‘hubs’ in the co-occurrence network that are not specific to the different subfields and break the community structure

(i.e. “size”, “urban growth”, “actor”, etc.). VOS-viewer assigns the nodes to a cluster using a variant of the modularity function with a resolution parameter γ that allow to play with the level of detail of the clustering (Waltman, et al., 2010). A high value of γ will result in a large number of clusters. Choosing the level of resolution when performing a clustering algorithm often requires taking an *ad hoc* decision in a trade-off between level of precision and necessity of simplification. This is the stage where the knowledge of the field plays an important role. With $\gamma = 1$, we identified five meaningful clusters that we describe below. The number of terms in different clusters varies from 45 for the smallest to 116 for the largest. The result of the text mining and the cluster analysis can be seen in Figure 2.3. In this visualization, the size of the nodes corresponds to the number of occurrences O of the terms in the corpus. The placing of these nodes is based on their co-occurrences, with terms that co-occur generally being located closer. The 20 most representative terms for each cluster can be found in Table 2.2, and gives a first description of what these clusters are about. Their relative importance have been calculated with a score I, which is the product of their number of occurrences O and their relevance score T.

Below, we will refer to the clusters with the following labels: REG or ‘regional systems’ stands for the first cluster due to the importance of the regional scale in its vocabulary, WCN for the second cluster as acronym for ‘world city network’, SIM for the third cluster with a vocabulary around the notions of ‘simulation and complexity’, ECON for the fourth cluster, which deals with ‘economic geography’ and the branch of urban economics known as ‘new economic geography’, and finally CSD for the fifth cluster refers to ‘city size distribution’, the main focus of interest.

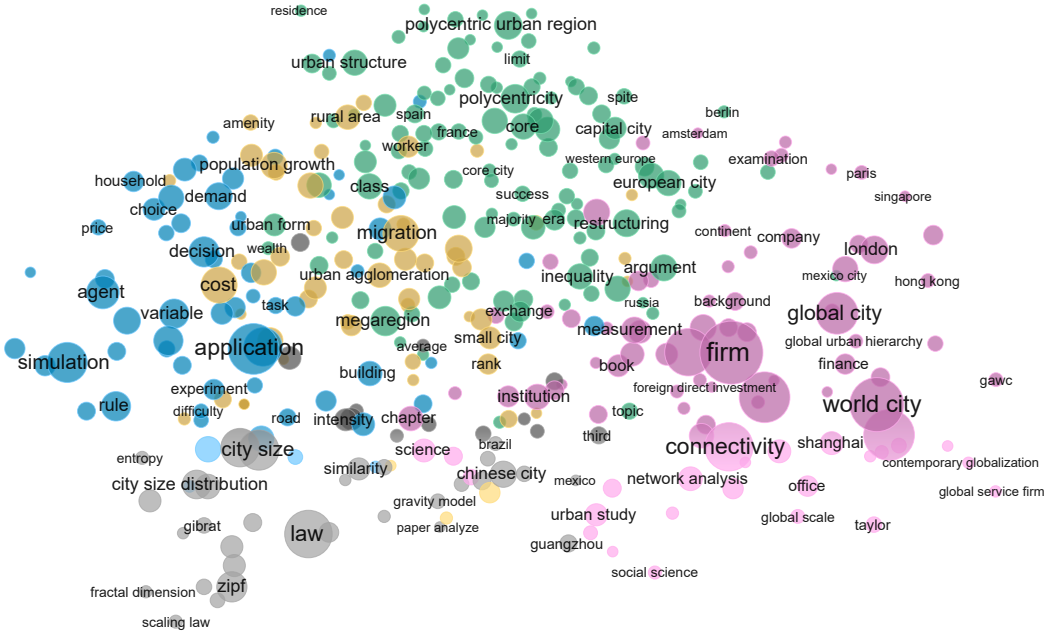

FIG. 2.3 'Semantic map' of co-occurrences of the vocabulary of urban system theory

TABLE 2.2 The vocabulary of the five schools of thought in urban systems research (only the 20 most prominent words included)

	Regional system	World city research	Simulation and complexity	Economic geography	City size research
Urban agents	commuting, planner	firm, producer service, office, global service firm, advanced producer service	agent, individual, household	migrant, urban population, worker	population size
Geographical dimension	polycentric urban region, Netherlands, Randstad, regional study, Mexico City, Spain, PUR, metropolitan region	world city network, globalization, London, Shanghai, New York, Tokyo		county, United-States, large metropolitan area, Canada	Guangzhou

>>>

TOC

TABLE 2.2 The vocabulary of the five schools of thought in urban systems research (only the 20 most prominent words included)

	Regional system	World city research	Simulation and complexity	Economic geography	City size research
Methodology		connectivity	simulation, agent, cellular automata, rule, application, modelling, scenario, behaviour, variable, parameter, decision, choice, estimation, GIS		law, Zipf, exponent, Gibrat, power law, scaling law, rank-size rule, rank-size distribution, fractal dimension
Thematic aspect	polycentric urban region, polycentricity, polycentrism, urban structure, deconcentration, competitiveness, restructuring, complementarity, PUR, metropolitan region, spatial development	world city network, world city, global city, globalization, city network, contemporary globalization, finance	complex urban system, land use, complexity, urban dynamics	migration, population growth, rural area, urban population, economic growth, urban planning	city size distribution, city size, size distribution, rank-size distribution, urban evolution, resilience
Inter-disciplinary vocabulary			complex urban system, complexity, cellular automata	cost, amenity, wage, worker, human capital	self, complex system, fractal dimension, entropy
Other	spatial scale, class	GaWC, Taylor		census, proximity, determinant, census data	property

The vocabulary used in studies of intercity relationships reveals many aspects of the different research approaches. We decided to focus especially on five aspects: the urban agents identified, the geographical scope, the methodology, some thematic aspects and the terms that reveal influence from other disciplines. We placed the 20 most prominent words for each cluster into these categories in Table 2.2.

Here, we describe the five clusters in more detail. REG is articulated around the notions of “polycentricity” and “polycentric urban regions”. Words referring to agents include “policy maker” and the “planner”, revealing a top-down approach, but also “commuting”, that suggests the opposite. The paper by Verhetsel et al. (this issue) fits in this tradition. Indeed, several studies on polycentricity use commuting data (De Goei et al. 2010; Burger et al. 2014) to study the patterns of relations between cities, revealing a bottom-up approach. The words that illustrate the topics of this field seem fairly similar, with many concepts referring to the ambition of urban policies and strategies: “competitiveness”, “complementarity”, “spatial development”, “deconcentration”, “implementation”, “cooperation”, “urban sustainability”. The regional spatial scale is at the centre of the research agenda with “regional development”, “regional economy”, “province” and “metropolitan region” and also the high occurrence of “Randstad”, the classic example of a polycentric urban region in the Netherlands, is consistent with this observation. Yet, also “national urban system” is present in this cluster. Except for “Russia” and “Japan”, most of the geographical toponyms that can be found refer to Europe: “Netherlands”, “Italy”, “European city”, “Barcelona”, “western Europe”. Finally, the co-word analysis does not reveal a shared methodological basis of this field and no interdisciplinary dimension.

WCN corresponds to the research on world cities and world city networks. The urban agents at the centre of this approach are private transnational firms (“firm”, “advanced producer services”, “company”, “office network”). These agents are associated with modern capitalism terms (“globalization”, “finance”, “foreign direct investment”). The geographical scope of most research in this cluster is either the global scale indicated by terms such as “global scale”, “global urban hierarchy”, “global urban network” and “world economy”, or the local scale of the cities making up this world city network, as indicated by the numerous city names that correspond to major centres of the global economy (“London”, “New York”, “Shanghai”, “Tokyo”, “Hong Kong”, “Beijing”, “Paris”, “Moscow” and “Amsterdam”). The profusion of place names revealed by the co-word analysis indicates clearly the strong empirical nature of this cluster of research. From a methodological point of view, the presence of “ranking” indicates the prevalence of benchmarking studies as an outcome of this approach. One can also note the presence of “interlocking network model”, which is the model allowing to build a network from the observation of multiple office locations of firms, which is extensively used by researchers working on world cities (many of them gathered in the Globalisation and World Cities research network) (Derudder et al. 2010; Taylor and Derudder 2015). “Network analysis” also appears in this cluster, strongly related to the notion of “connectivity”. These words refer to graph theory, an area of mathematics use to study graphs, which are models of relations between objects. This type of abstraction has been used frequently by

researchers working on world cities to analyse corporate networks (Neal 2008) or airplane networks (Zhang et al. 2016). Graph theory has been imported by geographers and regional scientists since the 1960s (Kansky 1963; Haggett and Chorley 1969), but recently, spatial scientists have started to use models popular in the interdisciplinary field of network sciences such as scale-free or small-words networks (see Neal, forthcoming, for a meta-analysis of studies using this last model). However, according to our analysis, the WCN cluster does not manifest strong interdisciplinarity.

The SIM cluster is organised around the terms "simulation" and "complex urban systems". The basic entities studied by those employing this approach are clearly identifiable ("agent", "household", "individual"). The interest for parameters set at the micro scale and elementary interactions between individuals is very visible in the vocabulary with terms such as "behaviour", "choice" and "decision". Among all the words of the cluster, only one geographic name appears: "South Africa". There are no other mentions of cities or regions which indicates that this approach is theoretical rather than empirical. The presence of "central place theory" in the graph, one of the most widespread theoretical models of a system of cities is also consistent with this orientation. Methods and tools are at the centre of this approach as the profusion of terms related to modelling and simulation shows. One can see two different methods of simulation: "multi agent system" on the one hand, and "cellular automata" on the other. These methods are associated with a particular terminology: "rule", "scenario" and "prediction". All these terms show clearly the interdisciplinary background of this research line, parallel to the computational turn in social science. According to Sanders (2014) this simulation field draws its main inspiration from physics, with the works on dissipative structures and synergetic, and from computer science and artificial intelligence that notably created the tools such as cellular automata and agent based models. Studies using these two types of simulation are both represented in our corpus (see for example Batty 2001; Bretagnolle and Pumain 2010). But simulation is not the only methodological framework represented in this vocabulary. References to network science are also visible in the cluster given the "complex network" and "graph" terms. The words that reveal the thematic interest are "urban dynamic", "spatial interaction" and "transportation network", which confirm that this cluster covers a significant part of the research program of the theoretical and quantitative geography.

ECON manifests the vocabulary of (new) economic geography. Words related to agents are "migrant", "worker", "human capital", showing an interest for individuals considered as economic agents. This focus on the micro-scale is corroborated by the frequent co-occurrence of "census" and "census data" with vocabulary of this cluster. In terms of scope, the cluster seems to study mainly Northern America with

the inclusion of “United States” and “Canada” as geographical keywords. The very high I score of “county”, a common administrative and political division used in the US as well as in some provinces in Canada confirms this territorial focus. The words that reveal the thematic dimension of the cluster are “population growth”, “urban population”, “economic growth” and “urban function”, referring to the economic dynamics of cities. The data co-word analysis also reveals mention of several levels of the urban hierarchy from “primate city” and “large metropolitan area” to “small city” and “rural areas”. The vocabulary of this cluster shows clearly that some papers of the corpus are characterised by the vocabulary of economics with “cost”, “wage”, “amenity” and “human capital”. This is further confirmed by the presence of “new economic geography” in the cluster.

The fifth cluster (CSD) gathers the vocabulary of studies dealing with aspects associated with city size. As expected, cities are the main unit of analysis, as aggregate of urban dwellers (“population size”, “city size”). The fact that the word “law” is the most representative of the cluster indicates the search for common laws or regularities with respect to city size distributions. This cluster presents numerous terms related to mathematical formalization such as “Zipf” (for the Zipf’s law), “Gibrat” (for the Gibrat’s law), “power law”, “scaling law”, “exponent” and “growth rate”. This interest for regularities is also visible with the mention of the “gravity model”, one of the most widespread models in spatial analysis. From a thematic point of view, the “hierarchical structure” and “size distribution” seems at the core of the research agenda, but also the evolution of these structures (“urban evolution”, “population growth”). In terms of territorial scope, several toponyms referring to China as well as “Brazil” and “Mexico” appear, showing that the field is not only theoretic but also deals with case studies. In terms of interdisciplinarity, along with the influence of physics (“entropy”), one can see the influence of the field of complex systems with “complex system”, “fractal dimension” and “self” (probably reflecting self-organisation and self-similarity).

This analysis allows to distinguish five different lexical fields in urban systems research. Strong differences in terms of methods, scope, thematic focus, main agents identified and influences from other disciplines have become manifest, which warrants to talk about five schools of thought. But some links can already be discerned. Both the SIM and CSD clusters refer to the interdisciplinary field of complexity theories. Moreover, the CSD and ECON clusters seem to focus both on demographic features of cities. There is a clear geographical spread in attention for these different subfields as shown in Table 2.3.

TABLE 2.3 Main contributing countries to the different subfields

Regional system (REG)		World city research (WCN)		Simulation and complexity (SIM)		Economic geography (ECON)		City size research (CSD)	
Country	%	Country	%	Country	%	Country	%	Country	%
Netherlands	16	UK	18	China	18	USA	23	USA	32
USA	12	China	16	USA	18	China	14	China	21
China	10	USA	14	France	13	Canada	8	UK	9
Spain	10	Belgium	13	UK	7	UK	8	France	9
UK	9	Germany	5	Netherlands	4	France	5	Israel	3

Sizable academic communities like those of the U.S.A. and China contribute to all subfields, but they are particularly dominant in the CSD and ECON clusters, followed by SIM where also France is a significant contributor. More than 60% of the WCN contributions derive from just four countries: the U.K., U.S.A., China and comparatively tiny Belgium. The Netherlands is the largest contributor to the REG subfield, which probably reflects its polycentric urban system and the fact that its main metropolitan area, the Randstad has been a classic research laboratory for polycentricity. Also Spain is an unusual suspect for the REG subfield.

2.3.2 Vocabulary and citation patterns

We now turn to exploring the relationships between the five schools of thought as they were identified through the co-word analysis. In order to analyse the evolution of these relations, we assign the papers (and occasional book chapter) to one of these schools, using a simple scoring method. Given the 5 lists of words (Table 2.2) corresponding to the schools of thought described above, we calculated the number of occurrences of words from each category in the title and abstract of individual texts. As the number of words for each school was unequal, we took the 40 most representative words for each school (based on the calculation of the I score as explained in 3.1) to make sure that papers have an equal chance of being assigned to the different schools. For example, if a paper mentioned 20 terms belonging to the SIM school, and 5 terms from the WCN school, it will be considered as belonging to the former. The result of this tagging process can be found in Figure 2.4, and it also informs about the citation patterns between individual publications. The fact that the WCN, REG, and CSD schools are clearly clustered together shows the presence of dense citation networks within these schools. The more dispersed distribution of the SIM and especially the ECON schools suggests that these schools are less internally

coherent, at least in terms of citations. In the case of ECON, we use the term “school of thought” because we identified a community in terms of disciplinary influence and thematic focus, however, the publications belonging to this school are so scattered that they do not seem to be united by a single research approach. It probably shows that the approaches of urban economists (the ‘new economic geography’) and the economic geographers approach have not been integrated much.

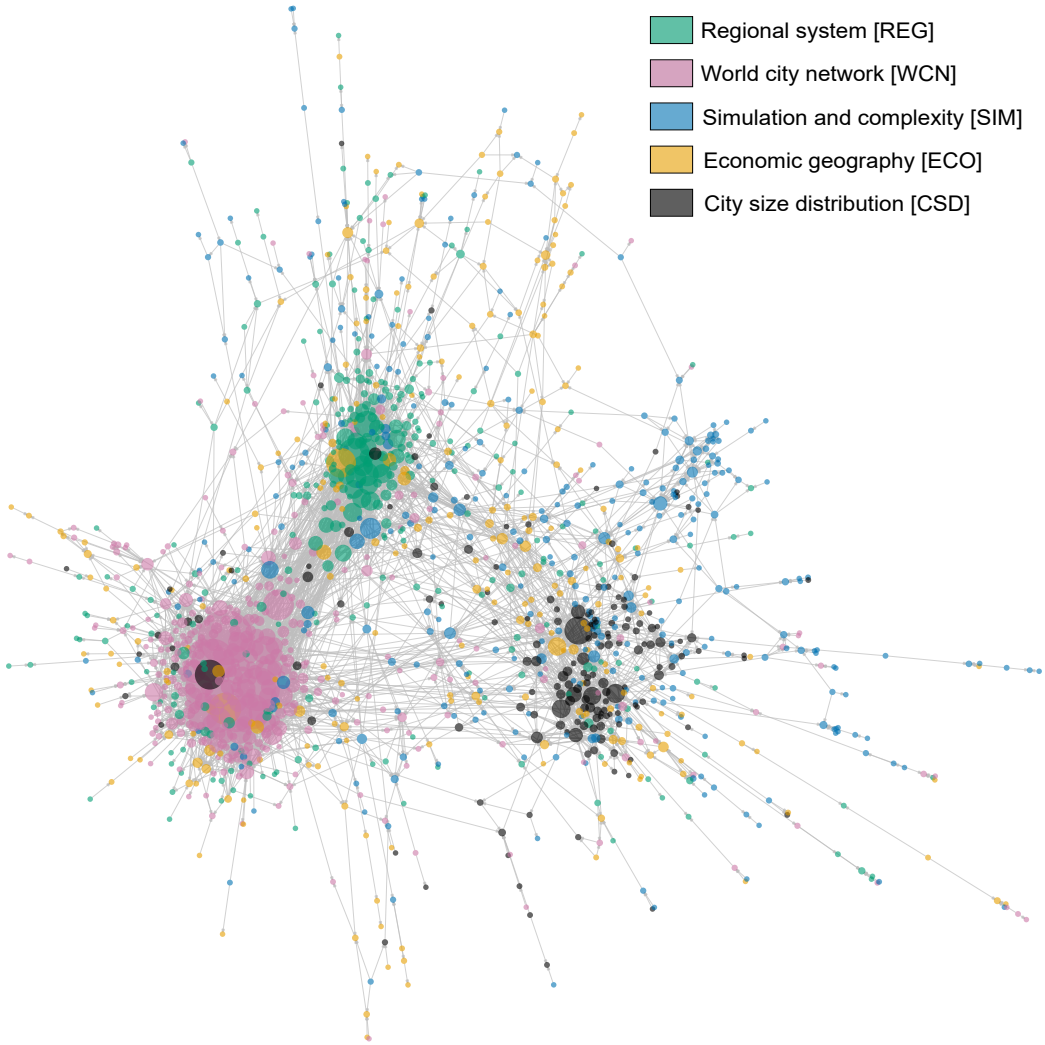

FIG. 2.4 Citation network of publications, labelled according to the schools of thought in urban systems research

TOC

In a next step, we filtered our network in order to obtain snapshots for 5 different periods (1995-1999, 2000-2004, 2005-2009, 2010-2014 and 2015-2017). For each period we computed an insularity index S (1) to analyse whether the schools are self-referencing or drawing inspiration from others in a particular period (after van Meeteren et al., 2015). This index corresponds to the share of citations within the school or subfield divided by the total number of citations. Given the fact that the literature does not disappear from one year to another, we take into account cited publications from previous periods.

$$S_{C_a}{}^{t} = \frac{\sum_{v_i \in C_a{}^{t}, v_j \in C_a} v_i \rightarrow v_j}{\sum_{v_i \in C_a{}^{t}} v_i \rightarrow v_j}$$

Where $S_{C_a}{}^{t}$ is the insularity index S of the subfield C_a at time t and v being a publication. The value of this score ranges between 0 and 1. A score close to 0 means that the papers of a subfield do not cite each other but cite other subfields, while a score close to 1 means that the subfield is very strongly inward-looking. Table 2.4 shows the evolution of this index and of the size of the different clusters (measured by the number of papers).

TABLE 2.4 Size and insularity index of the subfields for each period

		1995-1999	2000-2004	2005-2009	2010-2014	2015-2017
Regional system (REG)	size	38	37	53	109	71
	insularity	0.38	0.59	0.54	0.51	0.55
World city research WCN)	size	20	72	75	142	67
	insularity	0.20	0.74	0.78	0.80	0.75
Simulation and complexity (SIM)	size	31	39	56	128	80
	insularity	0.62	0.40	0.52	0.42	0.23
Economic geography (ECON)	size	34	46	46	76	56
	insularity	0.15	0.43	0.25	0.41	0.24
City size research (CSD)	size	11	13	39	67	36
	insularity	0	0.21	0.55	0.61	0.51

Two of the schools of thought manifest a grossly similar pattern: first a sharp increase of the insularity score, after which it remains relatively stable at a high insularity level. This holds for the REG and WCN schools addressing polycentric regions and world city networks respectively (and grossly speaking). REG scores 0.38 in the beginning of our study period (0.38), peaks during the period 2000-2004, and remains rather stable (between 0.51 and 0.55) the following periods. It is the only school that draws less inspiration from other fields and becomes more internally coherent in the most recent period of analysis. The insularity score of the WCN school is initially quite low for the period 1995-1999, probably reflecting that this school was in its initial stage of development, but its insularity score skyrockets in the following five years. It reaches its maximum between 2010 and 2014, with 80% of the citations remaining within the same school of thought. This suggests that a very coherent body of work emerges, but also that it does not draw a lot of inspiration from other schools of thought. It seems quite 'closed'. The pattern for the CSD school of thought (with a particular focus on city size distributions) is to some extent similar given the rapid increase (even from zero in the first period) and relatively high insularity score, but the increase happens later, and the insularity score drops quite substantially in the most recent period. The latter suggests that it is drawing more inspiration from other fields than before. We interpret these profiles of sharp increase followed by relative stability of the REG, WCN and also CSD schools of thought as the structuration and perpetuation of a clear research program organised after the publication of important seminal studies formulating a research agenda, presenting new models or methodology or giving new empirical results leading to a new approach. In the case of the WCN literature, these papers include those of Beaverstock et al. (2000) and Taylor (2001). For the REG subfield, the most central papers are those of Kloosterman and Musterd (2001) who sketched a research agenda for Polycentric Urban Regions (PURs), and the papers of Parr (2004) and Davoudi (2003) offering a critical reflection on the concept of PUR. In the case of CSD, the study of the size distribution of cities and of the link between urban size and functions is definitely not new in geography and economics. These questions have been debate for many years (see for example Vapnarsky, 1969) and was often featured in urban geography readers (Bourne and Simmons, 1978; Pred, 1977). However, in the mid-2000s, these research questions experienced a revival of scholarly interest with the import of scaling laws from natural science into urban studies following the publication of the seminal studies by Pumain and Guerois (2004) and Bettencourt et al. (2007). In our set, the number of papers related to this subfield increased a lot from one period to another. This has led to numerous studies on how the diverse properties of cities are changing with population, for instance addressing patent numbers, the selling of gas, road length, occupational structure, congestion, crime, etc. Nowadays, this subfield is still very dynamic and animated by a debate about the statistical validity and explanatory power of these

relations (Arcaute et al., 2015; Cottineau et al., 2017; Depersin and Barthelemy, 2018). The SIM school (on simulation and complexity) and ECON (economic geography) papers follow quite different trends. SIM had a higher insularity score in the beginning (0.62), but decreases over the period to reach a 0.23 score in the period 2015-2017, which is the lowest score of all the subfields in this period. This means that the SIM cluster of papers appears to be the most open to other subfields. ECON starts as one of the lowest (0.15) and fluctuates then between middle values (from 0.24 to 0.43). The final value of 0.24 suggests that it is also a rather open school of thought. The low insularity values for SIM and ECON are consistent with the rather dispersed structure of these schools of thought as presented in Figure 2.4. The relatively low scores can be interpreted as a greater openness of these subfields, or at least as being less structured around a very specific debate or coherent research agenda.

The relations between the different subfields are visualised in Figure 2.5. In this figure, all the papers from the same subfield have been aggregated into a single node, as well as their out-citations. Between 1995 and 1999, patterns of citations between the different subfields are quite equally distributed. REG cites equally itself and ECON, WCN cites all the others, SIM is connected to WCN and ECON, and CSD cites the SIM school of thought but not itself, clearly showing that it originates in the latter. It is from the period 2000-2004 that internal citation starts taking the biggest share of citations. In terms of external citations, the distribution is also quite equal. It is really from 2005 that our data starts to show preferred relations between subfields. For the period 2005-2009, most of the external citations of REG and ECON go to WCN, and this tendency is confirmed the following period, as well as for 2015-2017. WCN is also the most cited subfield by SIM and CSD. Among the literature on interurban relationships, the research on world cities/world city networks is clearly the one that received the most attention. In return, this subfield seems to pay some attention to REG and ECON for the three last periods, and to the SIM school of thought in the most recent period. The SIM school pays attention to the research on city size (CSD), but this is less so the other way around. They have a common interest in complex systems theories, as was already emerging from our analysis of the vocabulary of those two subfields. Finally, REG and ECON also manifest a common interest, especially for the period 2015-2017 where most of the external citations of ECON go to REG.

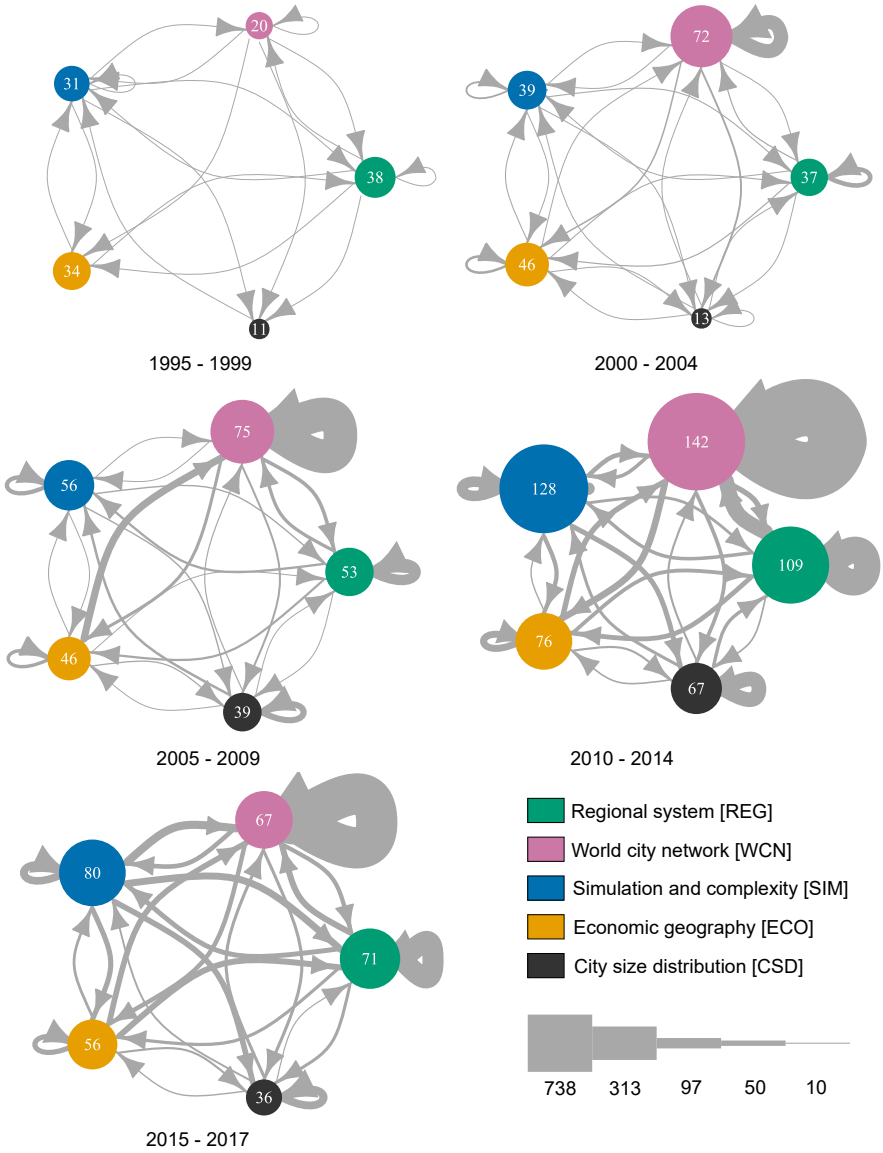

FIG. 2.5 Evolution of the citation patterns between the subfields

TOC

2.4 Conclusion and discussion

This paper presented the first bibliometric analysis of the research addressing relationships between cities, also referred to as 'urban systems' research. The analysis of the vocabulary in the papers led to the bottom-up identification of five clusters of terms corresponding to different schools of thought that have emerged over the last 23 years. Even though these schools manifest very different approaches, they can be split into two overarching groups. On the one hand, the research on world cities and on polycentric urban regions deals with quite specific objects of research. The former explores the top of the urban hierarchy by looking at global networks (mainly corporate networks and some transportation networks), while the latter studies very specific, polycentric settlement systems, characterised by the proximity of several small and medium-sized cities. These two fields depart from what they consider an existing reality and try to simplify it in order to understand it. To do so, they use models and heuristics, but the starting point is an empirical one. On the other hand, the theoretical and quantitative geography using notably simulation methods and the research cluster on city size distributions are both more interested in identifying general laws and mechanisms. They depart from hypotheses about causalities and try to explain or reproduce regularities presumably present in reality. According to its vocabulary and its central position in the co-word analysis, the subfield dealing with economic geography seems in between. The emphasis on rather abstract or generalised notions such as "cost", "wage", and "worker" would place it in the latter overarching group, but its strong focus on North America would include it in the former.

The bibliometric method developed in this paper has proven its capacity to study a set of publications by exploring both their content and their patterns in citations. First, the multi-step delineation procedure with adaptive queries and learning process allowed to collect a set of papers and book chapters with a minimum of noise and can be applied to other scholarly fields. The main advantage of this way of approaching literature is to limit the bias in the selection of the corpus by allowing for the possibility to include texts that are not known nor targeted a priori. The mapping of terms revealed effectively the different schools of thought addressing systems of cities. Again, using methods such as co-word analysis allows a bottom-up reconstruction of the different approaches in urban systems research, limiting the influence of predefined ideas of the field. Studying the vocabulary of papers proved to be a novel and adequate way to assess the primary units of analysis, the geographical scope, the methodologies and the thematic foci of different schools of thought, as well as several epistemological dimensions of each of them. Finally,

the hypernetwork approach, which combines semantic and citations networks from papers appears an efficient way of exploring the communication between different schools of thought.

Despite the small recent signs of openness, increasing fragmentation appears to be the main tendency over the last two decades. Three out of five schools of thought (REG, WCN and CSD) witnessed a quick and very substantial increase of their insularity score. All three schools of thought have had more that 50% of internal citations since 2005. It has to be noted that the very last period is characterised by a small average decrease of this insularity index, suggesting more integration between most schools of thought. But this is still insufficient evidence for a move from multidisciplinarity to interdisciplinarity in urban system research. Although such a move is generally propagated in science, it remains a question for further debate and research whether this is desirable in urban systems research, and whether it would bring us new insights. If the answer is confirmative, the next question is how to achieve this. As our analysis revealed, the quite strong ontological and epistemological differences between some of the schools of thought identified here do not necessarily allow for an easy generation, modification and recombination of ideas and approaches. Yet, we hope that our analysis urges others to explore this potential.

Bibliography

Alderson AS, Beckfield J (2004) Power and Position in the World City System. Am J Sociol 109:811–851. doi: 10.1086/378930

Arcaute E, Hatna E, Ferguson P, et al (2015) Constructing cities, deconstructing scaling laws. J R Soc Interface 12:20140745

Auerbach F (1913) Das Gesetz der Bevölkerungskonzentration. Petermanns Geogr Mitteilungen 59:74–76

Batty M (2013) The New Science of Cities. MIT Press

Batty M (2001) Polynucleated Urban Landscapes. Urban Stud 38:635–655. doi: 10.1080/00420980120035268

Beaverstock JV, Smith RG, Taylor PJ (2000) World-City Network: A New Metageography? Ann Assoc Am Geogr 90:123–134. doi: 10.1111/0004-5608.00188

Berry BJL (1964) Cities as Systems Within Systems of Cities. Pap Reg Sci 13:147–163. doi: 10.1111/j.1435-5597.1964.tb01283.x

Bettencourt LMA, Lobo J, Helbing D, et al (2007) Growth, innovation, scaling, and the pace of life in cities. Proc Natl Acad Sci 104:7301–7306. doi: 10.1073/pnas.0610172104

Bourne LS, Simmons JW (1978) Systems of cities: readings on structure, growth, and policy. 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. Hist Methods J Quant Interdiscip Hist 50:49–65. doi: 10.1080/01615440.2016.1237915

Bretagnolle A, Pumain D (2010) Simulating Urban Networks through Multiscalar Space-Time Dynamics: Europe and the United States, 17th-20th Centuries. Urban Stud 47:2819–2839. doi: 10.1177/0042098010377366

Burger MJ, Knaap B van der, Wall RS (2014) Polycentricity and the Multiplexity of Urban Networks. Eur Plan Stud 22:816–840. doi: 10.1080/09654313.2013.771619

Callon M, Courtial J-P, Turner WA, Bauin S (1983) From translations to problematic networks: An introduction to co-word analysis. Soc Sci Inf 22:191–235. doi: 10.1177/053901883022002003

Capello R (2000) The City Network Paradigm: Measuring Urban Network Externalities. Urban Stud 37:1925–1945. doi: 10.1080/713707232

Chavalarias D, Cointet J-P (2013) Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields. PLOS ONE 8:e54847. doi: 10.1371/journal.pone.0054847

Christaller W (1933) Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen. University Microfilms

Cottineau C, Hatna E, Arcaute E, Batty M (2017) Diverse cities or the systematic paradox of Urban Scaling Laws. Comput Environ Urban Syst 63:80–94. doi: 10.1016/j.compenvurbsys.2016.04.006

Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Syst 1695:1–9

Davoudi S (2003) EUROPEAN BRIEFING: Polycentricity in European spatial planning: from an analytical tool to a normative agenda. Eur Plan Stud 11:979–999. doi: 10.1080/0965431032000146169

De Goei B, Burger MJ, Van Oort FG, Kitson M (2010) Functional Polycentrism and Urban Network Development in the Greater South East, United Kingdom: Evidence from Commuting Patterns, 1981–2001. Reg Stud 44:1149–1170. doi: 10.1080/00343400903365102

Depersin J, Barthelemy M (2018) From global scaling to the dynamics of individual cities. Proc Natl Acad Sci 201718690. doi: 10.1073/pnas.1718690115

Derudder B, Taylor P, Ni P, et al (2010) Pathways of Change: Shifting Connectivities in the World City Network, 2000—08. Urban Stud 47:1861–1877. doi: 10.1177/0042098010372682

Ducruet C, Beauguitte L (2013) Spatial Science and Network Science: Review and Outcomes of a Complex Relationship. Netw Spat Econ 14:297–316. doi: 10.1007/s11067-013-9222-6

Ducruet C, Lugo I (2013) Cities and Transport Networks in Shipping and Logistics Research. Asian J Shipp Logist 29:145–166. doi: 10.1016/j.ajsl.2013.08.002

Engels TCE, Ossenblok TLB, Spruyt EHJ (2012) Changing publication patterns in the Social Sciences and Humanities, 2000–2009. Scientometrics 93:373–390. doi: 10.1007/s11192-012-0680-2

Friedmann J (1995) Where we stand: a decade of world city research. In: World cities in a world-system, Cambridge University Press. Cambridge, pp 21–46
Fujita M, Krugman P, Mori T (1999) On the evolution of hierarchical urban systems1. Eur Econ Rev 43:209–251. doi: 10.1016/S0014-2921(98)00066-X
Gibrat R (1931) Les inégalités économiques. Recueil Sirey
Haggett P, Chorley RJ (1969) Network Analysis in Geography. Edward Arnold
Hicks D (1999) The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics 44:193–215. doi: 10.1007/BF02457380
Kamalski J, Kirby A (2012) Bibliometrics and urban knowledge transfer. Cities 29, Supplement 2:S3–S8. doi: 10.1016/j.cities.2012.06.012
Kansky KJ (1963) Structure of transportation networks: relationships between network geometry and regional characteristics. University of Chicago.
Kloosterman RC, Musterd S (2001) The Polycentric Urban Region: Towards a Research Agenda. Urban Stud 38:623–633. doi: 10.1080/00420980120035259
Latour B, Woolgar S (2013) Laboratory Life: The Construction of Scientific Facts. Princeton University Press
Leydesdorff L (1998) Theories of citation? Scientometrics 43:5–25. doi: 10.1007/BF02458391
Leydesdorff L, Zhou P (2007) Nanotechnology as a field of science: Its delineation in terms of journals and patents. Scientometrics 70:693–713. doi: 10.1007/s11192-007-0308-0
Liu Z (2005) Visualizing the intellectual structure in urban studies: A journal co-citation analysis (1992-2002). Scientometrics 62:385–402
Meijers E (2007) From central place to network model: theory and evidence of a paradigm change. Tijdschr Voor Econ En Soc Geogr 98:245–259
Meijers E, Romein A (2003) Realizing Potential: Building Regional Organizing Capacity in Polycentric Urban Regions. Eur Urban Reg Stud 10:173–186. doi: 10.1177/0969776403010002005
Milanez DH, Noyons E, Faria LIL de (2016) A delineating procedure to retrieve relevant publication data in research areas: the case of nanocellulose. Scientometrics 107:627–643. doi: 10.1007/s11192-016-1922-5
Neal ZP (2008) The duality of world cities and firms: comparing networks, hierarchies, and inequalities in the global economy. Glob Netw 8:94–115. doi: 10.1111/j.1471-0374.2008.00187.x
Neal ZP (forthcoming) Is the urban world small? The evidence for small world structure in urban networks. Netw Spat Econ
Parr J (2004) The Polycentric Urban Region: A Closer Inspection. Reg Stud 38:231–240. doi: 10.1080/003434042000211114
Parr JB (2014) The Regional Economy, Spatial Structure and Regional Urban Systems. Reg Stud 48:1926–1938. doi: 10.1080/00343404.2013.799759
Pflieger G, Rozenblat C (2010) Introduction. Urban Networks and Network Theory: The City as the Connector of Multiple Networks. Urban Stud 47:2723–2735. doi: 10.1177/0042098010377368
Pred A (1977) City Systems in Advanced Economies: Past Growth, Present Processes, and Future Development Options. Wiley
Pumain D (2011) Systems of Cities and Levels of Organisation. In: Bourgine P, Lesne A (eds) Morphogenesis. Springer Berlin Heidelberg, pp 225–249
Pumain D (1997) Pour une théorie évolutive des villes. Espace Géographique 26:119–134. doi: 10.3406/spgeo.1997.1063
Pumain D (2006) Hierarchy in natural and social sciences. Dordrecht, The Netherlands : Springer
Pumain D, Guerois M (2004) Scaling laws in urban systems. St Fe Inst Work Pap 4
Raimbault J (2017) Exploration of an Interdisciplinary Scientific Landscape. ArXiv171200805 Cs
Sanders L (2014) Trois décennies de modélisation des systèmes de villes : sources d'inspiration, concepts, formalisations, Three decades of modeling systems of cities : sources of inspiration, concepts, formalization. Rev D'Économie Régionale Urbaine décembre:833–856
Sanders L, Pumain D, Mathian H, et al (1997) SIMPOP: A Multiagent System for the Study of Urbanism. Environ Plan B Plan Des 24:287–305. doi: 10.1068/b240287
Sassen S (1991) The Global City: New York, London, Tokyo. Princeton University Press
Taylor PJ (2001) Specification of the world city network. Geogr Anal 33:181–194
Taylor PJ, Derudder B (2015) World City Network: A Global Urban Analysis. Routledge
Van Eck NJ, Waltman L (2011) Text mining and visualization using VOSviewer. ArXiv Prepr ArXiv11092058

van Eck NJ, Waltman L (2014) Visualizing Bibliometric Networks. In: Ding Y, Rousseau R, Wolfram D (eds) Measuring Scholarly Impact. Springer International Publishing, Cham, pp 285–320
van Meeteren M (2016) From polycentricity to renovated urban systems theory: explaining Belgian settlement geographies. Ghent University
van Meeteren M, Poorthuis A, Derudder B, Witlox F (2015) Pacifying Babel's Tower: A scientometric analysis of polycentricity in urban research. Urban Stud 0042098015573455. doi: 10.1177/0042098015573455
Vapnarsky CA (1969) On Rank-Size Distributions of Cities: An Ecological Approach. Econ Dev Cult Change 17:584–595
Verhetsel A, Beckers J, De Meyere M (this issue) Assessing Daily Urban Systems (DUS) in Belgium: a network approach based on commuting flows, with special attention to gender and income differences, Netw Spat Econ
Waltman L, van Eck NJ (2012) A new methodology for constructing a publication-level classification system of science. J Am Soc Inf Sci Technol 63:2378–2392. doi: 10.1002/asi.22748
Waltman L, van Eck NJ, Noyons ECM (2010) A unified approach to mapping and clustering of bibliometric networks. ArXiv10061032 Phys
Wang H, He Q, Liu X, et al (2012) Global urbanization research from 1991 to 2009: A systematic research review. Landsc Urban Plan 104:299–309. doi: 10.1016/j.landurbplan.2011.11.006
Zhang S, Derudder B, Witlox F (2016) Dynamics in the European Air Transport Network, 2003–9: An Explanatory Framework Drawing on Stochastic Actor-Based Modeling. Netw Spat Econ 16:643–663. doi: 10.1007/s11067-015-9292-8
Zipf GK (1949) Human behavior and the principle of least effort: an introduction to human ecology. Addison-Wesley Press
Zitt M (2015) Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. Scientometrics 102:2223–2245. doi: 10.1007/s11192-014-1482-5
Zitt M, Bassecoulard E (2006) Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. Inf Process Manag 42:1513–1531. doi: 10.1016/j.ipm.2006.03.016

TOC

3 Using toponym co-occurrences to measure intercity relationships

Review, application and evaluation

This is the Author's Accepted Manuscript version (final draft post-refereeing as accepted for publication by the journal). The definitive, peer-reviewed and edited version of this article is published as: Meijers E. & Peris A. (2019) Using toponym co-occurrences to measure relationships between places: review, application and evaluation. International Journal of Urban Sciences. DOI: https://doi.org/10.1080/12265934.2018.1497526

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.

3.1 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 stressing the importance of network embeddedness for urban and regional development (Camagni and Capello, 2004; Neal, 2013a; Taylor and Derudder, 2016; McCann and Acs, 2011; Meijers et al., 2016, 2018; Camagni, 2017), discussing the existence of 'urban network externalities' (Capello, 2000; Burger and Meijers, 2016) 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 and Lees, 1986) and geographers (Bourne and 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 et al., 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 organisation of different types of functional linkages is not necessarily identical (Berroir et al., 2017; Burger, Meijers

and Van Oort, 2014a; Limtanakool et al., 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 and Diodato, 2014b). 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 and Derudder, 2016), such a proxy does not necessarily work well for smaller spatial scales (Lambregts, 2009; Burger et al., 2014a). 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 (Van Meeteren and Poorthuis, 2018; Schiff, 2015), 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 crowdsourced. 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 et al., 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 et al., 2011) or revisiting the world city network based on geolocated tweets (Lenormand et al., 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 together[9] ('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 et al., 2013) and public health (Thornton et al., 2017), its systematic application to the study of relationships between cities has only just started to develop (e.g. Hu et al., 2017), following some initial small-scale explorations of the potential of this method (e.g. Devriendt et al., 2008; Liu et al., 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).

9 Hu et al (2017) refer to this as 'semantic relatedness'

3.2 Measuring relationships between cities

3.2.1 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 analysing 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 et al., 2007; Pumain et al., 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 inter-city 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 (Nordlund, 2004; Neal, 2012, 2013b; Liu and Derudder, 2013), and it is not well capable of measuring relationships between (smaller) cities on the regional scale (Lambregts, 2009; Burger et al., 2014a). Other data allows to study intercity firms relations with actual relational data on ownership relations between headquarters and subsidiaries of multinational enterprises (Rozenblat et al., 2016).

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 relationships between cities. It can be obtained by looking at the infrastructure such as a railway, roads or postal road network (Bretagnolle and Franc, 2017; Derudder et al., 2014), by looking at the moves of vehicles such as ships (Ducruet et al., 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 and Witlox, 2005), train passengers (Berroir et al., 2017) or commuters, shoppers or business travellers (Burger, et al., 2014c; Nelson & Rae, 2016). Recently, flows of people have also been identified through geolocated posts of people on social media (Lenormand et al, 2015; Zhang et al., 2016), which allows to overcome the national dimension of data collection, but is not necessarily without representative bias. Another interesting source of relational data can be the mails and telephone calls (Zipf, 1946; Krings et al, 2009).

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 et al., 2014a; Choi et al., 2006; Ducruet et al., 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 analysing systems of cities.

3.2.2 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 et al., 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 and Cointet, 2013:2). So far, these 'elements' have included for instance organisations and firms (Vaughan and You, 2010); hyperlinks (Boulton et al., 2010; Salvini and Fabrikant, 2016); country names (Queyroi et al., 2015; Grasland et al., 2016) or even hashtags (Lorenz et al., 2018) in addition to the key words characterising scientific fields – see Peris et al. (2018) 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 and You, 2010:483), making co-occurrence analysis also an important tool for Webometrics.

In addition to keywords, people, papers, hyperlinks, countries, organisations 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 x 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 (2017) 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 to identify clusters of cities that are often mentioned together and consequently apply the toolkit of network analysis, it is hard if not impossible to conceptualise 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 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 organisation 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.

3.3 Research approach

3.3.1 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). '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.

3.3.2 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 personalised 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 and 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[10]. 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. 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 and 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 3.1) and having inspected a sample of pages with 20 to 25 unique co-occurrences, concluding that these should generally be included. Building on Rasool et al. (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.

10 commoncrawl.org

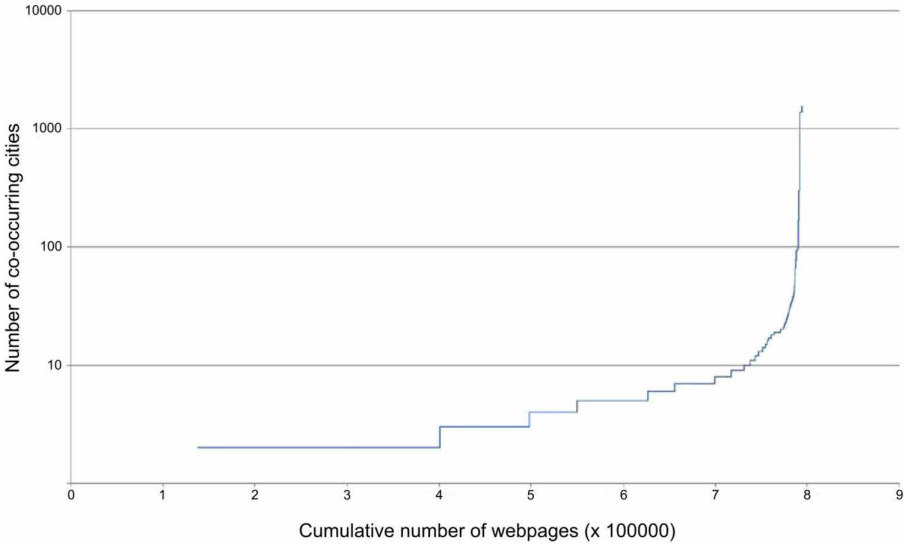

FIG. 3.1 Number of webpages plotted against the number of unique occurrences contained in these pages. Source: Brunner et al., 2017.

3.3.3 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.

- 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. 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 et al., 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 organisations, 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 two 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 3.1). 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.

TABLE 3.1 Potentially biased place names.

Source of bias	Frequency	Percentage[i]
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%

[1] *Relative to 1616 different place names ('unbiased place names' relative to 1639 places).*

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 surface 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.

3.3.4 Classification of co-occurrences

The filtered dataset allowed to count 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 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 et al, 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.

3.4 Results of the co-occurrence method

3.4.1 Overall pattern of co-occurrences

In this section, we will both visualise 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 (2014:100) 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 et al, 2016), detect outliers that may be caused by place name disambiguation, and to move beyond a simple visualisation 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, with 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 if the toponym co-occurrence method to places with different sizes. Table 3.2 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).[11]

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 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.

11 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.

TABLE 3.2 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)**
Frysian/Dutch place name different		.107 (.006)**		.368 (.021)**		.472 (.057)**
N city pairs	515,658	515,658	47,533	47,533	4,950	4,950
N 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.*

3.4.2 The spatial organisation of the Netherlands

While Figure 3.2 presents the pattern of absolute flows between the 100 largest places in the Netherlands, Figure 3.3 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).

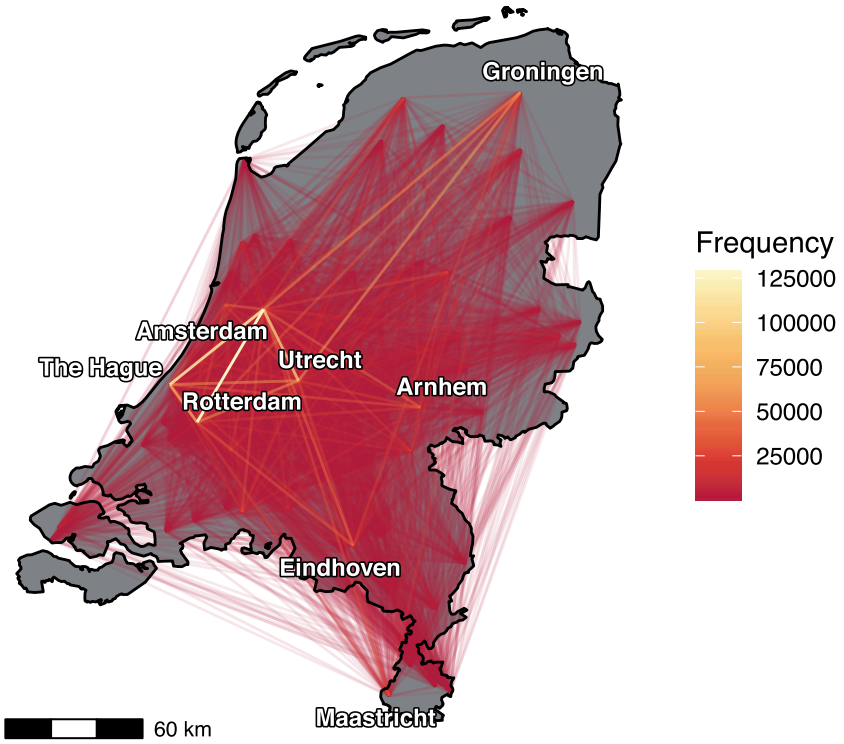

FIG. 3.2 Observed spatial organisation of the Netherlands based on the pattern of toponym co-occurrences.

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 3.2). 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 related to the main Randstad cities. However, the comparison of Figures 3.2 and 3.3 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).

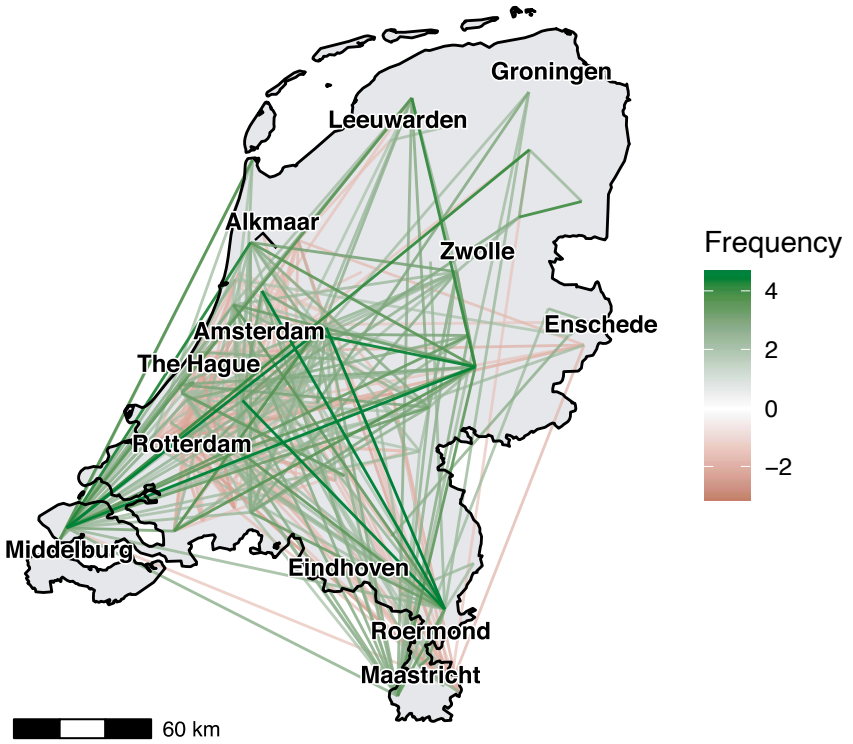

FIG. 3.3 Observed versus expected relations between Dutch places, based on toponym co-occurrences.[12]

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.

12 For clarity of the visualisation, only standardized residuals >=2 and <=-1 are displayed.

This can be further explored by calculating the sum of all unstandardized predicted and residual values and comparing these. Table 3.3 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 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 and 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).

TABLE 3.3 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

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 3.4) and relative flows (Figure 3.5).

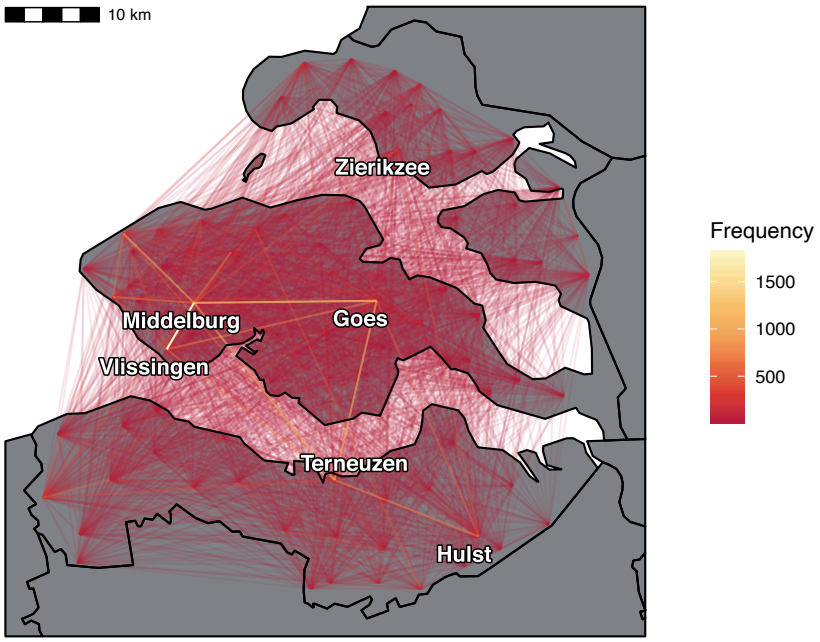

FIG. 3.4 Observed spatial organisation of Zeeland based on the pattern of toponym co-occurrences.

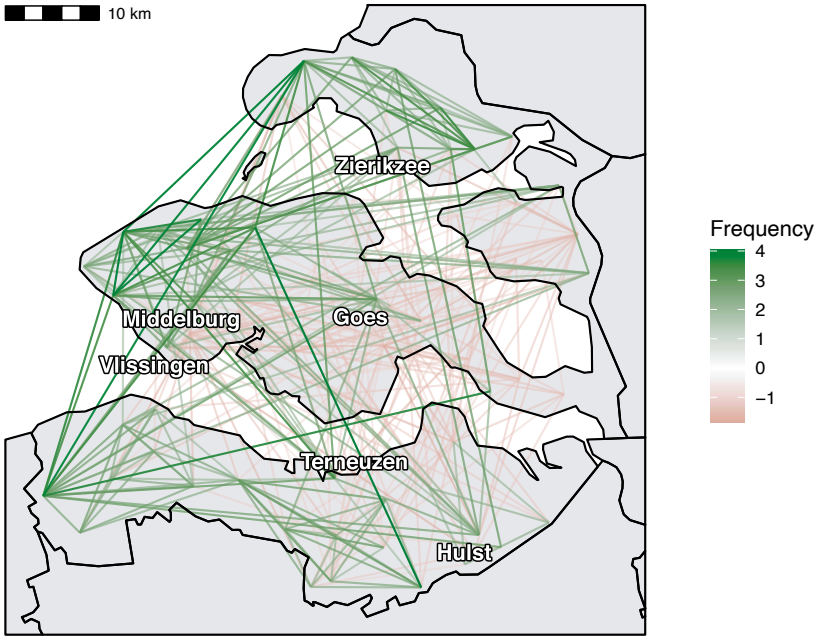

FIG. 3.5 Observed versus expected relations between places in Zeeland, based on toponym co-occurrences.[4]

TOC

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 he patterns could be tentatively expected and logically explained. The region appears rather well integrated, with a dominance of relationships that are more strong over relationships that are more weak. Figure 3.5 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 more strong, 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.

3.4.3 Classifying co-occurrences

The spatial organisation 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 et al., 2014a). 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 3.4 presents the number of city pair relationships categorised 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.

TABLE 3.4 Classified flows between places versus the gravity model.

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

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

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 categorised as 'shopping', up to 6.8% for 'education'. On average, 1,75% of all co-occurrences are categorised, 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.

3.5 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 organisation 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 organisation 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 Devriedt 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 categorise 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 categorising a website is not necessarily the same as categorising 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 CommonCrawl and the efforts to digitalise 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 interurban networks. After all, it is not accidental that the method was invented in this domain (Tobler and 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.

Acknowledgements

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 data collection 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).

Bibliography

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). Les systèmes urbains français : une approche relationnelle. *Cybergeo : European Journal of Geography*. Published online 6 February 2017. DOI : 10.4000/cybergeo.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. http://www.nltk.org/api/nltk.stem.html, 2017.

Boulton, A., Devriendt, L., Brunn, S., Derudder, B., & Witlox, F. (2010). City networks in cyberspace and time: Using google hyperlinks to measure global economic and environmental crises. ICTs for mobile and ubiquitous urban infrastructures: Surveillance, locative media and global networks (pp. 67-87).

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. and F. Witlox (2010). Networks of European Cities in Worlds of Global Economic and Environmental Change. *Fennia, 1881*, 37-49.

Brunner, T., Mališ, M., Reichert, G. & Agtmaal, P. van (2017). *UrbanSearch. Final report graduation project*. TU Delft. Delft: Delft University of Technology.

Burger, M.J., Meijers, E.J. & Oort F.G. van. (2014a). Multiple Perspectives on Functional Coherence: Heterogeneity and Multiplexity in the Randstad. *Tijdschrift voor economische en sociale geografie, 105*, 444-464.

Burger, M., Thissen, M., Oort, F. van & Diodato, D. (2014b). The Magnitude and Distance Decay of Trade in Goods and Services: New Evidence for European Countries. *Spatial Economic Analysis, 9*, 231-259.

Burger, M. & Meijers, E. (2016). Agglomerations and the rise of urban network externalities. *Papers in Regional Science*, *95*, 5-15.

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 hierarchicaldistributive approach. *Quality Innovation Prosperity, 21*, 31-48.

Camagni, R. & Capello, R. (2004). The city network paradigm: theory and empirical evidence. In Capello R, Nijkamp P (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 networks. *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. 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. *Cybergeo : European Journal of Geography*. DOI: 10.4000/cybergeo.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, Mass: Harvard University Press.
Hu, Y., Ye, X. & Shaw, S.-L. (2017). Extracting and analysing semantic relatedness between cities using news articles. *International Journal of Geographic 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.
Keßler, 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., Gonçalves, B., Tugores, A. & Ramasco, J.J. (2015). Human diffusion and city influence. *Journal of The Royal Society Interface*, *12*, 1-9.
Lenormand, M., Bassolas, A., & Ramasco, J. J. (2016). Systematic comparison of trip distribution laws and models. *Journal of Transport Geography, 51*, 158-169.
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 web pages, *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: Cherifi, C., Cherifi, H., Karsai, M., Musolesi, M. (eds) *Complex Networks & Their Applications*. VI. Complex Networks 2017 2017. Studies in Computational Intelligence, 689. Springer, Cham.
McCann, P. & Acs, Z.J. (2011). Globalization: countries, cities and multinationals. *Regional Studies,45*, 17-32.
Meeteren, M. van & 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. & 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.
Meijers, E. & Burger, M. (2017). Stretching the concept of 'borrowed size'. *Urban Studies, 54*, 269-291.
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.P. (2013a). *The Connected City: How Networks are Shaping the Modern Metropolis*. New York: Routledge.
Neal, Z. (2013b). Brute Force and Sorting Processes: Two Perspectives on World City Network Formation, *Urban Studies*, *50*, 1277-1291.
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. https://tel.archives-ouvertes.fr/tel-00008053/document.
Peris, A., Meijers, E. & Ham, M. van (2018). The evolution of the systems of cities literature: Schools of thought and their interaction. Working paper TU Delft.
Pred, A. (1977). *City Systems in Advanced Economies: Past Growth, Present Processes, and Future Development Options*. 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. (2016). 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 multirelational 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., Ma, 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*. 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.

4 One century of information diffusion in the Netherlands derived from a massive digital archive of historical newspapers

The DIGGER dataset

This is the Author's Accepted Manuscript version (final draft post-refereeing as accepted for publication by the journal). The definitive, peer-reviewed and edited version of this article is published as: Peris A., Faber W.J., Meijers E. & van Ham M. (2020) One century of information diffusion in the Netherlands derived from a massive digital archive of historical newspapers: the DIGGER dataset. Cybergeo, European Journal of Geography. DOI: https://doi.org/10.4000/cybergeo.33747

ABSTRACT Previous studies have highlighted the importance of having long term data for the study of cities, but such sources are relatively scarce. This is especially the case for data about relations between cities, which is a crucial aspect of urban dynamics. Over the last two decades, many efforts have been made to digitalize texts, including books and newspapers, which are primary sources on most of our societies. Researchers have shown that these massive digital archives can be used to identify macroscopic trends related to historical and cultural changes. The wealth of geographic information in such digital archives has not been used much, while they are very valuable for the study of cities. In this paper, we present DIGGER, a newly developed dataset that we built on Delpher, the digital archive of historical newspapers of the National Library of the Netherlands, by extracting geographical information from a selection of 102 million of news items. This dataset allowed us to study the spatial diffusion of information on and between the Dutch cities from a corpus of 81 newspapers published in 29 different cities between 1869 and 1994. This paper presents the method developed to build the dataset as well as the validation steps for the accuracy of the place name recognition. This dataset can be used to study the evolution of the Dutch urban system as well as aspects related to the spatial diffusion of information and geographical bias in media coverage.

KEYWORDS System of cities, flows, diffusion, history, database, text mining, geocoding

4.1 Background & Summary

We have designed DIGGER in order to study the evolution of the Dutch urban system by investigating information flows extracted from historical newspapers that go back to 1869. Newspapers are full of geographical information as most of news items include one or more place names. However, studying this geographical information systematically is not an easy task. In this project, we have geocoded place names contained in a selection of 102 million news items to build origin-destination matrices with places mentioned in the news items (o) and places where the newspapers were issued (d) for 125 years (t). It takes the form of a cube with 3 dimensions: origin, destination and time.

Information circulation has been identified as a key factor in urban dynamics. Classical urban literature has highlighted the importance of available information on locational decisions of individuals, groups and firms and of its role as prerequisite for other kinds of people and goods movements. An early paper of Zipf (1946) used local

newspapers to study the interactions between distant 'communities' and used this data in a gravity model. Allan Pred (1977) also used local newspapers from different American cities to measure the time it took for information to travel from one place to another. But because of the time and workforce needed for the data collection, these studies were limited to a very small number of cities or short periods of time.

However, with the recent development of computing techniques, it is now possible to upscale and systematize data collection from newspapers to analyse the information circulation at the level of an entire territory. Over the last 10 years, interdisciplinary teams of computer scientists and humanity scholars have worked on extracting patterns from massive textual corpora illustrating historical and cultural changes. A seminal study by Michel et al. (2011) showed the potential of this approach by compiling 5 million digitalized books to provide quantitative insights on the evolution of grammar, as well as the detection of events such as pandemics, the influence of certain thinkers, or the evolution of gender bias in vocabulary. While the importance of such an approach was widely acknowledged, the study received a number of critiques related to the book selection (Morse-Gagné, 2011), and the fact that it did not include newspapers, which were thought to better reflect their time due to the frequency of publication (Schwartz, 2011). Indeed, the written press was the primary source to access information from distant places in most of the industrial societies for a long period of time. More recently, a study on British newspapers has used more refined techniques such as Named-Entity recognition to study the content of a massive corpus of historical newspapers (Lansdall-Welfare et al., 2017). While this study could look more precisely at historical and cultural trends, the analysis of the geographical focus, which was not the core of the study, remained at the stage of visualisation. However, problems related to extracting spatial information from text where not addressed, including the variety of scales (an article can mention a street, a city, a country, etc.) and ambiguities in place names.

Over the past few years, researchers have been increasingly interested in extracting geographical information from unstructured or semi-structured text data (Grasland, 2019; Meijers and Peris, 2018; Tranos and Kefalas, 2018). Extraction of spatial information from text was also carried out to map and analyse the global scientific production with a bibliometric database (Maisonobe et al., 2016). Mining these huge amounts of textual data is an important challenge for social sciences because these textual sources contain much information on social and economic processes, which are very often tied to places.

The motivation for the collection of DIGGER is to build a dataset on city-to-city interactions in order to test hypotheses related to the evolution of an urban system through history through the prism of information flows. Our objective is to look for macroscopic spatial trends in the way information is diffused and how this is changing

over time. In the study related to this data paper, we investigate the changing role of distance, city size, cultural and administrative borders on the circulation of information on a sample of 31 local newspapers between 1869 and 1930 (Peris et al., 2021). We find evidences of a space-time contraction, with faraway places being increasingly covered. The changes in patterns of information flows are also characterized by a hierarchical selection process. Almost all newspapers report more and more about the 4 main cities of the Randstad (Amsterdam, Rotterdam, The Hague and Utrecht), at the expense of the intermediate provincial cities located close-by.

The following section presents aspects related to the methods used for the data collection. It focuses first on the selection of a corpus of newspapers from the digital archive and of a set of cities for which data is collected. Then, it presents issues in place names recognition and choices to deal with these issues. Afterwards, we present the tests that we did to have statistics on the accuracy of our method. We then present an example of use of the dataset by mapping the information field of different cities with information flows in 1871. Finally, we give some concluding remarks on the potential of this database and its possible uses.

4.2 Methods

4.2.1 Corpus selection

The first important step in any quantitative study using a text archive is to select a relevant corpus. The content of a digital archive might be influenced by many factors such as digitalization policies, projects targeting a specific part of the media landscape (a newspaper, a region or a time period) or copyrights issues. Carefully selecting the corpus can significantly reduce bias, and is necessary to create a dataset as representative as possible depending on the research question. We have applied four criteria in the selection of newspapers:

- the newspaper had to be issued after 1869;
- its publication place had to be in the Netherlands;
- the newspaper had to exist during at least two consecutive decades;
- and we dropped the many small newspapers that were published only during the Second World War.

The following paragraphs detail and justify the choices that have been made to select the final corpus of 81 newspapers.

We started by analysing the temporal coverage of Delpher. At the time the data collection started, there were 1970 different titles in the archive. Table 4.1 shows that most of the newspapers have a lifespan of 5 years or less. The very short lifespan of most of titles is consistent with the findings of Van Kranenburg et al. (1998) that show that during the period 1848-1997, most Dutch daily newspapers did not even survive a decade. Indeed, like any other firm, if a newspaper does not find a market (here a readership), it is most likely to disappear. The fact that some newspapers were able to survive during long periods of time is a proof that they were supported by a sufficiently large readership. Using the lifespan of newspapers is a crude but relatively reliable proxy for their importance. By selecting newspapers according to this time dimension, we ensure that they had a sufficient diffusion to stay alive at least two decades consecutively. This huge variability in duration is also reflected by the amount of news items published by the different newspapers.

TABLE 4.1 Summary statistics of the Delpher corpus

	Min.	Median	Mean	Max.	St. dev.
Life span (Delpher)	5	5	11.14	180	17.69
Published items (Delpher)	1	123	67,843	13,307,273	491,679.6
Lifespan (selection)	11	50.50	53.39	129	28.53
Published items (selection)	1316	744,390	1,530,737	13,307,273	2,165,452

There are also important fluctuations in terms of number of publication of news items across the three centuries that are covered by the database (Figure 4.1). While the period 1700-1800 is characterized by a relatively small number, one can see an increase after the middle of the 19th century, followed by a very rapid rise and a peak that culminates in the 1940s. This tendency reflect the history of the Dutch press. The increase in the second half of the 19th can be explained by the abolishment of a tax on newspapers – the '*dagbladzegel*' – that made them cheaper and affordable for a wider public. As we are interested by the amount of non-local information received by urban dwellers, we decided to take this time mark as our starting point, because from this period, newspapers became the backbone of information diffusion in the Netherlands. The following period is a period of development of the press, that ends in a peak during the Second World War, a period were many anti- and pro-

German newspapers were created, most of the anti-German being underground. This resulted in the presence of a lot of short lived newspapers only published during the Second World War (n=2139) that can be very interesting for historians interested in the war but less relevant for long term studies. The application of these 4 criteria resulted in a sub-corpus of 81 newspapers that still cover an important part of the Delpher archive.

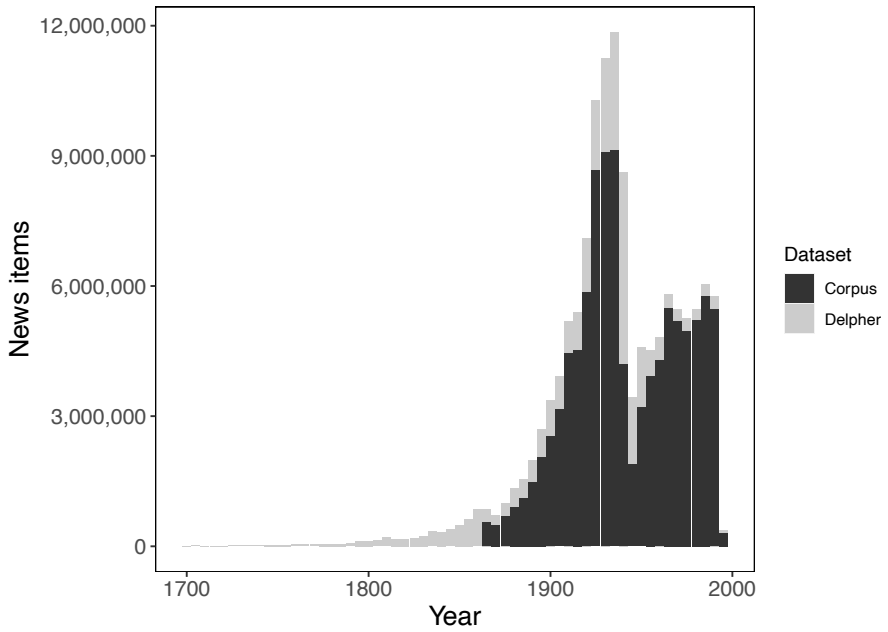

FIG. 4.1 News items per year in Delpher and in the sub-corpus

4.2.2 Selection of cities

In any study on cities, the selection of entities to work on is never a trivial operation. The city is indeed a very complex object and its definition varies among countries and epochs. In our case, defining our primary units of analysis is made difficult by the fact that the data collection is meant for a corpus that covers more than one century.

Cities can be defined according to many criteria, they can be continuous build-up areas, functional entities, designated by a certain level of urban functions or by administrative status. A definition that is often used is municipal entities above a certain threshold of population. Because we are interested in identifying cities in texts, we must go beyond these definitions and identify the terms that relate to cities in the common language. We adopt what Goodchild and Li (2011) call a "placial" perspective. These authors make a distinction between the *spatial perspective*, where the geographic information is organized by coordinate systems, and the *placial perspective*, that focusses on places as "named domains in human discourse". For them, one is not more important than the other as "the name people give to places and points of interest constitute a very significant form of geographical information". We decided to look at the terms people use to say where they live because place names have a stronger inertia than the boundaries of local governments. The "*woonplaatsnamen*", that can be literally translated as "names of places of residence", appeared as the most interesting concept. The *woonplaatsen* are used in the everyday language, they are the toponyms people include when writing down an address. They are generally associated with population centres such as cities, town or villages and their surroundings.

To allow a data collection in a reasonable amount of time, it is very important to work on a limited number of entities. For this reason, we are focusing only on the top of the system of settlements. We must therefore distinguish place names that cover urban places from the rest, and do so for the entire period that is studied. Similarly to a previous cross-temporal analysis of the Dutch urban system (Van der Knaap, 1980), we decided to depart from the current situation and keep the list of units of analysis consistent throughout the period covered by the data collection. The main advantage of this choice is that the same basis is used for the entire period, meaning that the number of units of analysis does not change with time. Nonetheless, we acknowledge that there are also some drawbacks. A place that could not be considered as urban in the beginning of the period but only at the end will be included in the data for every period. In return, a place with a population dipping under the threshold during the period will not be included at all.

We created a database on population per *woonplaatsen* with census data available at postcode level for the year 2011[13] and the geocoding API from PDOK[14] to identify to which *woonplaats* the postcode was attached to. We kept only the *woonplaatsen* with

13 http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=81310ned&D1=0&D2=a&HDR=T&STB=G1&VW=T

14 https://github.com/PDOK/locatieserver/wiki/API-Locatieserver

more than 10,000 inhabitants. This threshold is often used by statistical agencies and scholars as the lower limit to define urban centres, and significantly reduces the number of places to query for. The result of this selection is a set of 317 Cities. Figure 4.2 presents their locations obtained from the Geonames database[15].

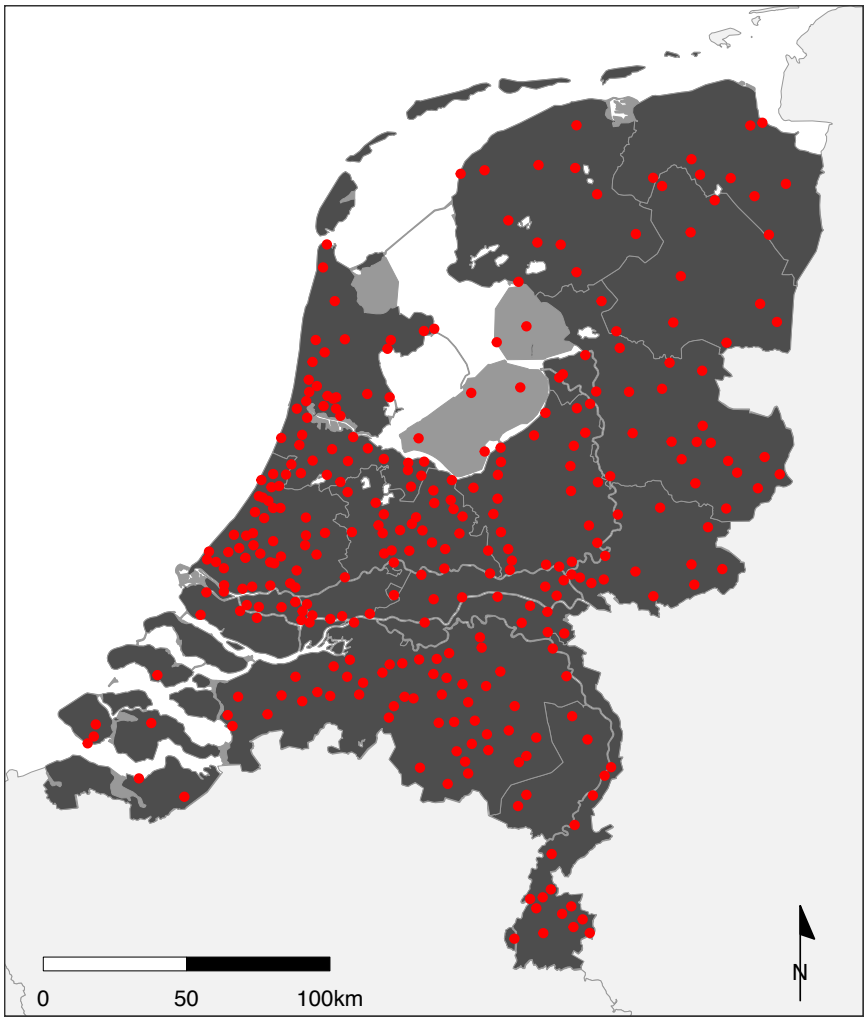

FIG. 4.2 Location of the 317 cities for which data is collected

15 https://www.geonames.org/

TOC

4.2.3 Classification of issues in place name recognition

City names, and more generally speaking place names, are subject to many ambiguities. These ambiguities can lead to important under- and overestimations when doing simple counts based on word frequencies. In a previous study (Meijers and Peris, 2018), different problems were identified in the case of the Dutch *woonplaatsen*. The most important sources of errors leading to false positives are listed below.

- Multiple meaning: some place names are similar to common names (i.e. “Huizen” = houses or “Dieren” = animals), verbs (i.e. “Leiden” = to lead, “Kampen”, to fight), and adjectives (i.e. “Houten” = wooden). This is the highest source of false positives.

- Homonymy: several places can have similar names. This is the case for “Katwijk”, which is at the same time a medium-sized coastal city in South Holland and a very small village in North Brabant. This can also be the case when a region and its most important city have the same name such as for Groningen and Utrecht.

- Family names: in quite some cultures, it is common to have a family name that relates to a place. In the list of cities that we are using for the data collection, (Van/Van der) “Beek”, (Van) “Dongen” and (Van) “Doorn” are among the top 100 most frequent family names in the Netherlands[16].

- Organisations: place names are sometimes used by organisations, firms or institutions in their name. Our university, Delft University of Technology, is a case in point. Given the intimate connection of the most of these organisations with their place, one could argue that this is less a problem as news items using these organisation names will often be referring to something related to or happening in that place.

In terms of false negatives, there is one important source of errors:

- Multiple names: this can be the case when the old name coexists with the newer one (i.e. Den Haag/’s-Gravenhage and Den Bosch/’s-Hertogenbosch), when working on a multilingual corpus, or when places are also referred to with an abbreviation. The case of Alphen aan den Rijn, sometimes also spelled Alphen a/d Rijn, or simply referred to as ‘Alphen’ by some, can be mentioned.

16 http://www.cbgfamilienamen.nl/nfb/documenten/top100.pdf

4.2.4 A trade-off between computation time and precision level

Most of the issues mentioned above can be avoided by using a combination of Named-Entity-Recognition (NER) and disambiguation algorithms. NER is a subtask of Natural Language Processing (NLP) that aims to locate and classify entities from a given text into pre-defined categories. Named entities can be locations, persons, organisations, dates, measures (money, weight, distance, percent…), etc. In our case, considering the size of the dataset, such a method could not be used for every city as it would have taken months to perform NER on the entire corpus. We decided to go for a mixed technique to retrieve the data on cities in a reasonable amount of time. The issues that could occur with the list of 317 cities were listed and dedicated solutions were selected to handle these issues (Table 4.2). NER was used only for ambiguous cases. Different types of NER algorithms exist. Their efficiency largely depends on the types of entities that are targeted, the domain and the language. Studies applied to historical newspapers have shown that the level of performance of these algorithms can differ significantly (Ehrmann et al., 2016; Mosallam et al., 2014). In this project we have used the multiNER software[17], a NER set-up developed by the research department of the National Library of the Netherlands for the enrichment of several Dutch text corpora. This software is combining the outputs of 3 different NER packages in order to increase the accuracy of the recognition by using a certainty score. The leading package is the Stanford NER[18], which was previously trained manually using annotated sheets of Dutch historical newspapers (Neudecker et al., 2014). The two following packages are spaCy[19] and polyglot[20], both using a pre-trained Dutch NER-model.

17 https://github.com/KBNLresearch/multiNER

18 https://nlp.stanford.edu/software/CRF-NER.shtml

19 https://spacy.io/

20 http://polyglot.readthedocs.io/en/latest/

TABLE 4.2 Issues in city name recognition and their solution.

Problem	Frequency	Share (%)	Method	Implementation
Multiple meanings	23	7.3	NER	Yes
Homonymy	15	4.7	NER + *disambiguation*	No
Family names	3	0.9	NER	Yes
Organisation	*Unknown*	*Unknown*	NER	No
Multiple names	7	2.2	Multiple string queries	Yes
Unambiguous place names	274	86.4	String queries	Yes

Table 4.2 shows that the vast majority of city names is not ambiguous (86.4%) and does not require the use of NLP techniques. For these 274 cities, we performed SRU[21] queries using city names as simple search terms to retrieve the relevant articles from the corpus. This operation could be done in a reasonable amount of time.

In order to have an idea of the problem of false positives in the case of city names with multiple meanings, we measured the number of hits for two city names – one unambiguous and one ambiguous – in 21 different newspapers from the corpus. The two cities that were selected are Best, a small town close to Eindhoven which has a name that is a very common word in Dutch (the superlative of "better", like in English), and Dordrecht, a bigger city in South-Holland which has a very low chance of having false positives. Figure 4.3 shows that the case of Best manifests a considerable difference between the two techniques and require the use of NER. For this reason, in the case of the 23 cities with multiple meanings, we first collected the relevant articles via string queries using the SRU protocol and used the multiNER software to see whether the city was considered as a named entities. We kept the articles when 2 out of the 3 NER packages agreed that it was a named entity (a place name) indeed.

21 http://www.loc.gov/standards/sru/

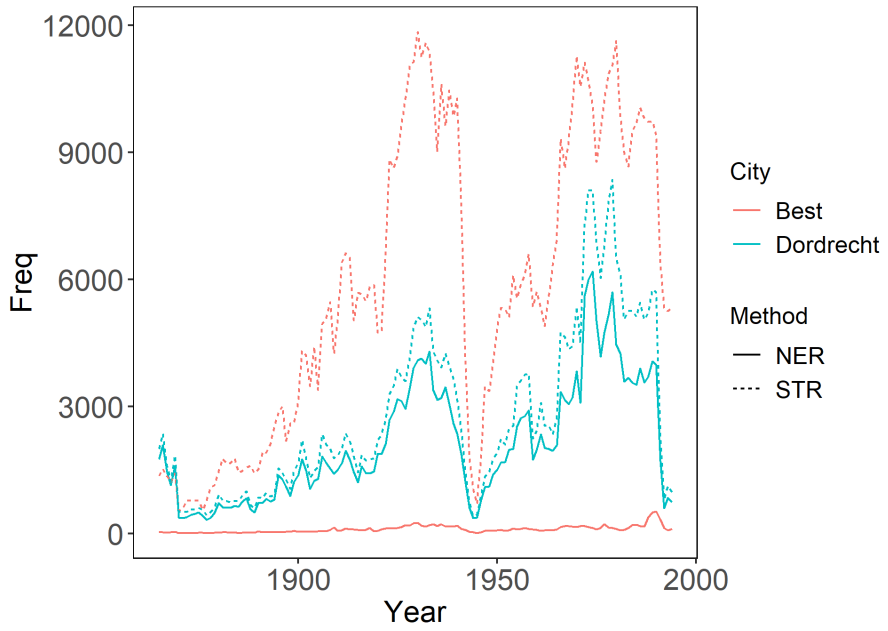

FIG. 4.3 Comparison of two retrieving techniques for Best and Dordrecht.
Dashed lines represent the simple string queries (STR) performed via the SRU protocol, the solid lines shows the frequency after using the NER algorithm.

NER was also used for the 3 cities that occur often in family names. As we noticed some misclassification on the kind of named entities by the multiNER software, we kept only the articles with named entities that exactly matches the city name. This way, we could drop the names of people that are composed of a first name (or initials), a family name, and sometimes a prefix in between (“van”, “de”, “van der”, etc.).

In the case of organisations, we could not apply NER because we had an insufficient knowledge of the organisations using the city names from the list. Such organisations would have been very difficult to identify considering the cross-temporal dimension. Moreover, as most of the time the use of a toponyms in the name of an organisation reflect relation with the place, this problem is not as important as multiple meanings and family names.

For cities with multiple names, multiple string queries via the SRU protocol were done. For example, we searched both for “Den Haag” and “‘s-Gravenhage” and aggregated the results afterward.

TOC

Finally, homonymy was the most difficult issue to handle. For a maximum level of precision, it would have been necessary to develop a specific disambiguation algorithm that uses the sentence around the named entity, the metadata of the newspaper (i.e. the place where the news item is published), as well as the importance of the possible places. However, we did not apply any disambiguation algorithm as the 15 cities from the list have homonyms of much smaller size (Figure 4.4). This limits the numbers of errors in their case.

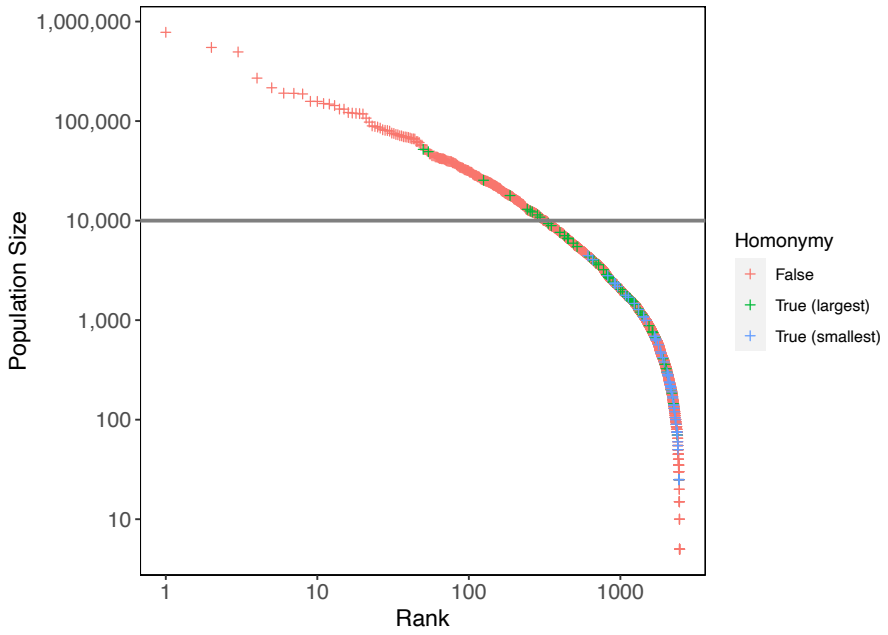

FIG. 4.4 Homonymy in settlement names.
The horizontal grey line represents the threshold above which the data is collected.

4.2.5 Data collection and structuration

The different steps of the data collection are summarized in Figure 4.5. They resulted in two files: one with the results of the data collection for the unambiguous city names (freq_count_STR.csv) and one for the ambiguous city names (freq_count_NER.csv). The file for unambiguous place names is structured the following way:

TABLE 4.3 Structure of the freq_count_str.csv file

ppn	city	type	year	freq
37631091X	Amsterdam	artikel	1871	347
37631091X	Amsterdam	advertentie	1871	149

The column *ppn* corresponds to a unique identifier given to each newspaper title. The column *city* is a character string corresponding to the city that is mentioned. The next variable, *year*, indicates the date. After that, *type* describes whether the city is mentioned in an article, an advertisement, some family announcements, or in the caption of an illustration. Finally, *freq* indicates the number of times this combination occurred. In this example, we can see that, according to the first line of the table, Amsterdam was mentioned in 347 articles of *De Maasbode*, a Rotterdam newspaper, in 1871.

The file for ambiguous place names is structured almost the same way. The only difference is that additional to the frequency returned by the simple string query, there is an extra column with the number of hits after performing NER on the individual articles returned after the first query:

TABLE 4.4 Structure of the freq_count_ner.csv file

ppn	city	type	year	Freq_str	Freq_ner
400337266	*Leiden*	artikel	1914	89	21
400337266	*Leiden*	advertentie	1914	47	35

Additionally to these two files, the dataset contains also elements allowing to spatialise the data such as a file containing metadata on the newspapers, including the coordinates of the place where they were published (np_metadata.csv) and a file with the coordinates of the cities mentioned (cities_information.csv). Other files are also included such as freq_count_corps.csv, that contains the total number of items published in each year for every newspapers, which allows for example to standardise the data. More detailed descriptions of the files can be found in the metadata of the dataset.

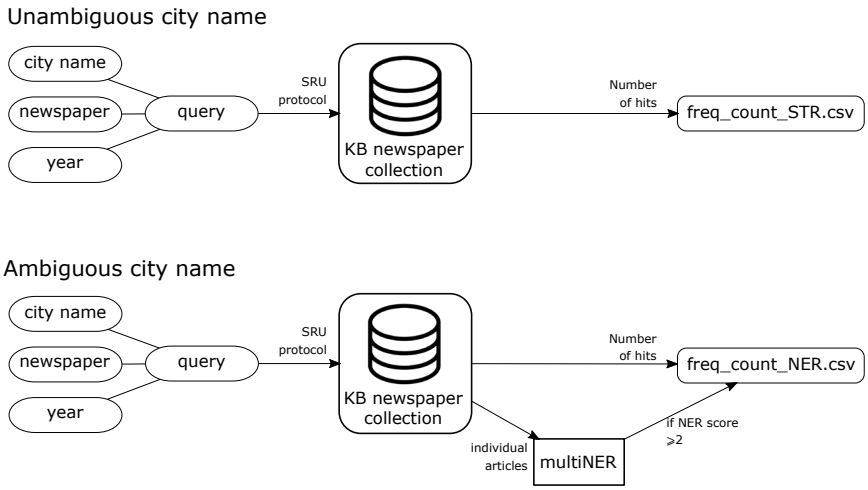

FIG. 4.5 Data collection algorithms

4.3 Technical Validation

In order to validate the accuracy of the dataset, we decided to manually assess the results of our algorithm on three different samples of news items containing in each cases articles, advertisements and family announcements. The division in three sets was done on a temporal basis. We divided the period for which we have newspapers in three equal time slots (1869-1910, 1911-1952, 1953-1994) and randomly selected 50 news items between each of these time marks. This separation in different sets was done because we were aware that the quality of prints significantly improved during this period, affecting the efficiency of the automatic recognition of characters (OCR) used during the digitalisation of the newspapers. This selection resulted in three tables similar with the structure shown in Table 4.5. Each line corresponds to a news item. The first column contains an identification number, the second column the year of publication, the third column is the plain text of the digitalized item. The next three columns corresponds to the different steps of the place name identification. STR is the result of a simple string query for unambiguous place names, NER column is the result of a string query for the places that are in the list of ambiguous place names, and NER result is the outcome of the NER algorithm on the ambiguous place name. For example, in the case of the third row of Table 4.5, the string ‘Goes’ has been identified in the text of the news item, but the multiNER did not classify it as a place name, so it does not appear in the ‘NER result column’.

TABLE 4.5 Structure of the sets used for sensitivity analysis

ID	YEAR	Plain text	STR	NER	NER result
1	1884	Keukenmeid of Huishoudster,Er biedt zich aan tegen half September een net Meïsja in eeu kleiu gezin als Of 31 1 ook is zij niet ongenegen eene ziekelijke Dame op te passen. Brieven franco, left. B, bij den Boekh. 1). RRAAIJ ENBRTNK, te Woerden. (5996)	Woerden		
2	1885	ZEE-MILITIE.De Burgemeeiter en Wethouders van Venloo nootfigen bij deze de lotelineen uit, die bij de Zee-Militie verlangen te dienen, zich daartoe bij hen aantemelden, ter plaatselijke Secretarie vóór den 1 April aanstaande. Venloo , den 12 Maart 1S86.			
3	1892	Burgerlgke Stand. GEHUWD: A. v. Dorp, jm. 31 en E. v. Vollenho ven, jd. 23 j., Pelikaanstraat 1. H. Pootman van Oije, wedr. 57 en M. L. Hazemijjer, wed. v. C. v. Hoek, 46 j., Hoogstraat 261. G.Kapsenberg, jm. 33 en J.A.v.der Goes,jd. 33}? L. Warande 106. J. H. Reidt, jm. 29 en A. F. v. Rjjn, jd. 26 j., Diergaardesingel 78. Huwelijks-Brieven en Verlovings-Circulaires worden gedrukt en spoedig afgeleverd, desverlangend geadresseerd ter drukkerij van het Nie uw sblad Goedkoop. Fijner papier naar keuze van i den besteller.		Goes	

We then counted the number of true positives, true negatives, false positives and false negatives to derive precision and recall indices for our three periods of time. These two indices are used to evaluate the outcome of an automated classification. The Precision P corresponds to share of relevant instances among the retrieved instances, and can be defined as:

$$P = \frac{tp}{tp + fp}$$

Where tp corresponds to the true positives and fp to the false positives. We also computed the recall R, which corresponds to the share of relevant instances that were correctly retrieved. This index takes the following form:

$$R = \frac{tp}{tp + fn}$$

Where fn corresponds to the number of false negatives. Table 4.6 shows the results of these two calculations for the three different periods.

TABLE 4.6 Results of the precision and recall tests

Period	*P*	*R*
1869-1910	0,91	0,92
1911-1952	0,98	0,96
1953-1994	0,99	0,88

The results of this validation process show an overall very good accuracy of our algorithm in the identification of place names in raw data. Most of the errors that we found in the randomly selected sample of articles were false negatives related to the quality of the OCR. This type of errors are almost unavoidable in quantitative analysis of digitalized texts. However, many efforts are being made by to constantly improve OCR quality.

4.4 Application: The information field of 15 Dutch cities in 1871

We present the maps resulting from the mapping of information field for 15 different cities in 1871 (Figure 4.6). Pred (1971) defines information fields as the total array of non-local contacts of individual places. Usually, these non-local contacts are likely to be high with nearby places, with which the frequency of interaction is important. In contrast, normally less information will likely be received from distant places. This pattern partly relates to the selection of information as being relevant for a group, given their pattern of interactions. Because of the considerable variability in the number of news items published in each newspaper we decided to plot the relative frequency of place-name mentions in comparison to the total number of news items published. To highlight the regional patterns in news coverage, we computed the Stewart potential with R package *SpatialPosition* (with an exponential function, span = 10000, beta = 3). These maps confirm the importance of distance for information flows as most of the attention is concentrated on the close-by cities and towns in 1871, with some attention to the big cities of the provinces of North and South-Holland.

A more extensive study on the diffusion of information between the Dutch cities and its evolution over time can be found in (Peris et al., 2021).

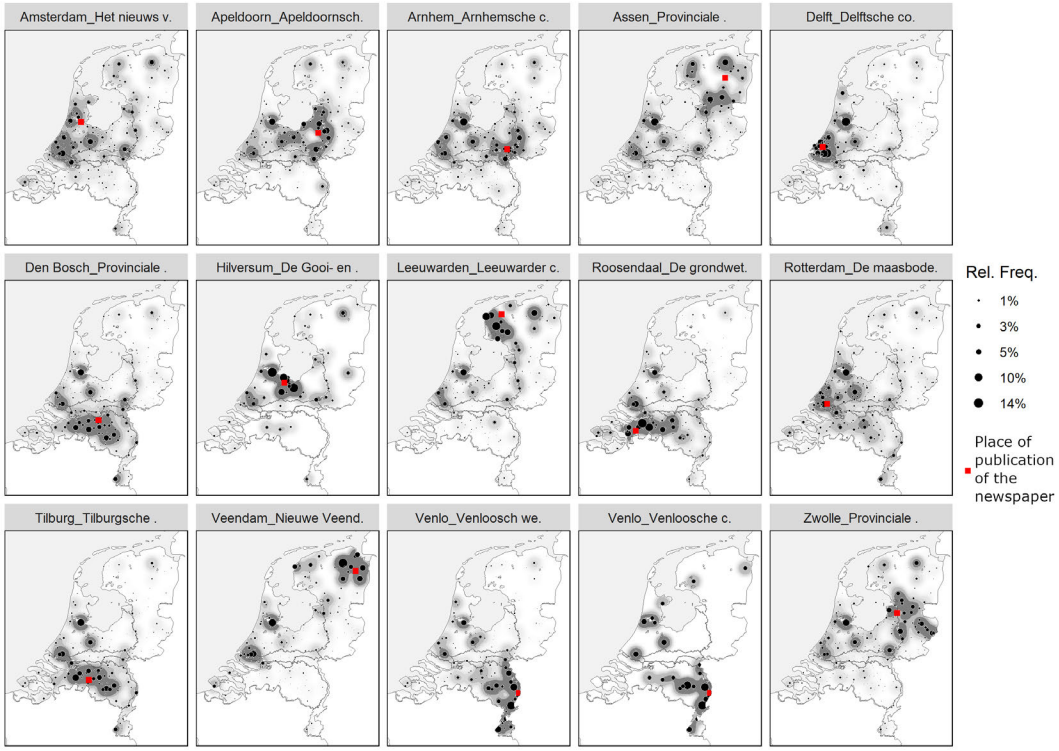

FIG. 4.6 Information field extracted from 15 local newspapers

4.5 Conclusion

Massive archives of digitalized text are full of geographic information. This is especially the case for archives of newspapers as these recorded the pulse of past societies. Quantitative analyses do not replace in depth readings, but they are a new way of looking at these sources and can reveal hidden patterns that appear only at the macroscopic scale. The reconstruction of the geography of information flows at different periods in time is such an hidden pattern that can only be observed through distant reading techniques. It allows to gain knowledge on the spatial organisation of territories through time.

However, extracting such patterns remains an important challenge from a methodological point of view. In this paper, we show the different steps that were needed to build a database on information flows between cities for a period of 125 years which we call DIGGER. It necessitated three main steps. The first one was to filter the different periodicals in order to keep only important ones. The second one was to select a sample of places that are consistent in terms of scale, toponomy and definition. Finally, the last step consisted in the building of an algorithm extracting accurate information from a massive dataset in a reasonable amount of time by using simple string queries and Named entity recognition (NER).

DIGGER can serve a wide variety of purposes. It has great potential for urban scholars to answer questions related to the dynamics of Dutch cities and the spatial diffusion of information, as well as by historians or media scientists interested in the geographical bias of news coverage. More generally, the methodology proposed in this data paper is of interest for people working on extracting geographic information from unstructured text data.

4.6 Dataset description

Language
English

Spatial coverage
317 cities and towns in the Netherlands (min y = 50.85, min x = 3.57, max y = 53.33, max x = 7.03). Reference system: ESPG 4326.

Temporal coverage
01/01/1869 – 31/12/1994

Creation date
September 2018

Dataset creators
- Peris A. (Department of Urbanism, Delft University of Technology, Delft, The Netherlands)
- Faber W. J. (Koninklijke Bibliotheek, The Hague, The Netherlands)

Format name and version

File	Description
cities_information.csv	geographical information on cities for which data has been collected
freq_count_corpus.txt	number of news items published by newspapers (per year/type)
freq_count_NER.csv	news flows between cities with ambiguous names and publication places of the newspapers (per year/type)
freq_count_STR.csv	news flows between cities with unambiguous names and publication places of the newspapers (per year/type)
np_metadata.csv	main information on the different periodicals included in the study
ppn_correspondance.txt	table to aggregate unique newspapers with multiple ppn code (unique identifier in Delpher database)
thesaurus_cities.txt	table to aggregate unique cities that have different names (ex. 's-Gravenhage = Den Haag)

Repository location

https://data.4tu.nl/repository/uuid:a14a1607-dafe-4a8a-aebc-d1c5cd66a588

Source

The digital archive of newspapers is accessible on the Delpher website (https://www.delpher.nl/). The data was created by querying the catalogue of the Koninklijke Biliotheek via a SRU protocol (http://jsru.kb.nl/sru/sru?query=).

Bibliography

Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F., 2016. Diachronic Evaluation of NER Systems on Old Newspapers 11.

Goodchild, M., Li, L., 2011. Formalizing space and place. CIST2011 - Fonder les sciences du territoire, Nov 2011, Paris, France. Proceedings du 1er colloque international du CIST, 177–183.

Grasland, C., 2019. International news flow theory revisited through a space–time interaction model: Application to a sample of 320,000 international news stories published through RSS flows by 31 daily newspapers in 2015. International Communication Gazette 1748048518825091. https://doi.org/10.1177/1748048518825091

Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., Team, F.N., Cristianini, N., 2017. Content analysis of 150 years of British periodicals. PNAS 114, E457–E465. https://doi.org/10.1073/pnas.1606380114

Maisonobe, M., Eckert, D., Grossetti, M., Jégou, L., Milard, B., 2016. The world network of scientific collaborations between cities: domestic or international dynamics? Journal of Informetrics 10, 1025–1036. https://doi.org/10.1016/j.joi.2016.06.002

Meijers, E., Peris, A., 2018. Using toponym co-occurrences to measure relationships between places: review, application and evaluation. International Journal of Urban Sciences 1–23. https://doi.org/10.1080/12265934.2018.1497526

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, T.G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L., 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331, 176–182. https://doi.org/10.1126/science.1199644

Morse-Gagné, E.E., 2011. Culturomics: Statistical Traps Muddy the Data. Science 332, 35–35. https://doi.org/10.1126/science.332.6025.35-b

Mosallam, Y., Abi-Haidar, A., Ganascia, J.-G., 2014. Unsupervised Named Entity Recognition and Disambiguation: An Application to Old French Journals, in: Perner, P. (Ed.), Advances in Data Mining. Applications and Theoretical Aspects, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 12–23. https://doi.org/10.1007/978-3-319-08976-8_2

Neudecker, C., Wilms, L., Faber, W.J., van Veen, T., 2014. Large-scale refinement of digital historic newspapers with named entity recognition 16.

Peris, A., Meijers, E., van Ham, M., 2021. Information diffusion between Dutch cities: Revisiting Zipf and Pred using a computational social science approach. Computers, Environment and Urban Systems 85, 101565. https://doi.org/10.1016/j.compenvurbsys.2020.101565

Pred, A.R., 1971. Large-City Interdependence and the Preelectronic Diffusion of Innovations in the U.S. Geographical Analysis 3, 165–181. https://doi.org/10.1111/j.1538-4632.1971.tb00360.x

Schwartz, T., 2011. Culturomics: Periodicals Gauge Culture's Pulse. Science 332, 35–36. https://doi.org/10.1126/science.332.6025.35-c

Tranos, E., Kefalas, P., 2018. Digging into digital archives: the evolution of the digital economy in the UK. Conference Paper.

van der Knaap, G.A., 1980. Population growth and urban systems development: a case study. M. Nijhoff.

Van Kranenburg, H.L., Palm, F.C., Pfann, G.A., 1998. The life cycle of daily newspapers in the Netherlands: 1848–1997. De Economist 146, 475–494.

Zipf, G.K., 1946. Some Determinants of the Circulation of Information. The American Journal of Psychology 59, 401–421. https://doi.org/10.2307/1417611

TOC

5 Information diffusion between Dutch cities

Revisiting Zipf and Pred using a computational social science approach

This is the Author's Accepted Manuscript version (final draft post-refereeing as accepted for publication by the journal). The definitive, peer-reviewed and edited version of this article is published as: Peris A., Meijers E. & van Ham M. (2021) Information diffusion between Dutch cities: revisiting Zipf and Pred using a computational social science approach. Computers, Environment and Urban Systems. DOI: https://doi.org/10.1016/j.compenvurbsys.2020.101565

ABSTRACT News travels fast and far, and the general idea is that the spatial extent of news coverage has increased over time. Information flows are always involved in systems of interdependent cities. This is the reason why George Zipf and Allan Pred, both pioneers of the urban systems literature, were eager to obtain data on these relations to understand urban system dynamics. However, because of limited resources in data acquisition, they restricted their studies to small samples of cities or short periods of time. By using novel computational social science techniques on a digital archive of historical newspapers, we could map and explore changes in the spatial extent of news coverage in the Netherlands at an unprecedented detailed scale for a period of 60 years. In this paper, we analyse 24 million news items mentioning 312 different cities and towns in a sample of 31 local newspapers. Thanks to this data, we were able to reconstruct the information field of urban readerships from different cities and how it changed over time. By analysing their evolution, we find evidence of space-time contraction with an increasing coverage of faraway places in the period

ranging from 1869 to 1929. However, this coverage is not evenly distributed but is characterized by a hierarchical selection process. Coverage of the largest cities in the Randstad increased at the expense of information flows from intermediate provincial cities. More generally, this paper shows how computational social science approaches may offer new ways of looking at urban dynamics with large text corpora such as digital archives of historical newspapers.

KEYWORDS System of cities, information flows, gravity model, Netherlands, historical newspapers

5.1 Introduction

When studying the organisation of systems of cities, research can build on many different types of empirical data: the circulation of people, goods, capital or information. Among these elements, information has a particular place in the research tradition on urban systems because it plays a crucial role in organising the complex patterns of networks and flows connecting cities. The tradition to focus on information started with early research by the pioneers of the system of cities literature analysing the content of local newspapers (Pred, 1973, 1977; Zipf, 1946a) and is still active today because of the availability of new data sources on human communications through mobile devices and social media (Grauwin et al., 2017; Krings et al., 2009; Stephens and Poorthuis, 2015).

Because information is a prerequisite to any other kind of exchanges, mapping its patterns helps to understand the organisation of the system of cities itself. This idea has been corroborated recently by researchers working on the development of the postal road network in France between the 17th and 19th century (Bretagnolle and Franc, 2017). They have shown that the development of an integrated communication network was concomitant with the development of the city-system at the national scale and the take-off in urbanisation rate. Having longitudinal data is very important in order to understand cities because the roots of their growth or decline are often situated far in the past. Inertia and initial advantages are indeed often major explaining factors of their situation in the urban hierarchy. However, collecting data on their relations, and especially information circulation remains a challenge because of the change in communication technology.

Some of the earliest research on urban systems has collected data from newspapers in order to analyse the circulation of information (Zipf, 1946a), because it was one of the only ways to develop a relational approach of cities at this time. But because of the cost of the data collection – involving analysing hundreds of actual paper newspapers – these studies were limited to a small number of cities, and very short periods of time. Today, thanks to the effort of digitization of historical newspapers undertaken by national libraries and the developments in data science, it is possible to upscale such analysis and apply distant reading techniques to extract similar data from millions of digitalized newspaper sheets. Digital humanities scholars have shown that these massive digital archives can be used to identify macroscopic trends related to cultural changes (Bod, 2013; Lansdall-Welfare et al., 2017), but very few studies have looked systematically at the geographical dimension of these archives.

In this paper, we look at the evolution of the circulation of information within the Dutch urban system with data extracted from Delpher, the digital archive of historical newspapers from the National Library of the Netherlands. This archive contains around 12 million newspapers pages containing even a much larger number of news items. We have selected a corpus of 31 local newspapers and collected the news referring to 312 settlements for the period 1869-1929. This represents a total number of 24 million of news items, which we use to analyse the evolution of the Dutch urban system over a long time period through its patterns of information flows. The overarching goal of the paper is to test the potential of this novel data source for reconstructing the evolving urban geography of an entire country. In order to achieve this goal, we will look for regularities in how flows of information develop over time and test different hypotheses related to the evolution of systems of cities.

Following our theoretical review (section 2), a variety of hypotheses will be formulated addressing the role of space and cities in information circulation. After a presentation of the novel data we use (3), and our initial research approach employing gravity modelling to understand the circulation of information over time (4), we test our hypotheses in the following section (5). Further explorations of our hypotheses using alternative approaches that handle spatial heterogeneity in information circulation are presented in section 6. Finally, we conclude in section 7.

5.2 Theory and background

5.2.1 Information flows in systems of cities

One of the main characteristics of cities is their ability to organize territories by articulating many types of networks and flows. Cities do not function in isolation but in systems at different scales from the regional to the global (Pumain, 2011). While recent research indicates that these relations are increasingly important (Meijers et al., 2016), it has been shown by historians that relations were central in the fate and fortune of cities long before the current globalisation (Hohenberg and Lees, 2009). For all these reasons, networks and flows are at the centre of the understanding of cities (Batty, 2013).

Information flows have long been recognized by geographers and urban scholars as an essential explanation of urban and spatial processes (Hägerstrand, 1967; Meier, 1962; Törnqvist, 1970, 1968). Researchers acknowledged very early that information was the most central resource in an urban system, because "without information about the economic risks and opportunities of the system, there would be no directed movement within the system" (Zipf, 1946a). The same idea was also emphasised in the work of Meier (1962), arguing that information and knowledge, that are conveyed by communications, were at the root of the mechanisms making economic growth possible in cities. Few years later, the empirical investigations of Allan Pred on the urban system of the United-States departed from the same hypothesis: "None of the economic actions and location decisions that underlie individual and collective urban growth can materialize unless preceded by information acquisition. None of the interurban commodity, capital, and human flows that are the outward expression of growth can transpire unless there is either the transmission of knowledge about demand, prices, and opportunities or some other form of information exchange." (Pred, 1973, p. 2).

The connections between information and other types of circulations is the reason why it was a major focus in quantitative geography research from the 1950s onward. The dissertation of Hägerstrand, initially published in Swedish in 1953 and translated into English later on by Pred (Hägerstrand, 1967), played a big role in putting the study of information patterns at the core of the research agenda of human geography. Hägerstrand's contribution was to propose models describing the adoption of innovations in the population of a Swedish region by means of

information dissemination through personal contacts. The notion of information field was presented for the first time in his work as an operationalisation of the probability of contact between individuals. In this seminal work focusing on a rural population, this probability was decreasing homogeneously with distance. However, Hägerstrand noted that in the case of a system of cities, the urban hierarchy would channel the course of diffusion. This element was empirically observed by Pred in his analysis of the system of cities of the United-States in the 19th century (Pred, 1973), focusing on the transmission of many innovations and identifying spatial biases in the availability of information and their pivotal role in explaining the diffusion patterns. In this study, he observed the prevalence of hierarchical diffusion processes but observed also some "neighbourhood effect" with small cities in the vicinity of big ones receiving innovations more early than would be expected given their position in the hierarchy – an early example of what would become known as 'borrowed size' (see Meijers et al., 2016).

While the previously mentioned works were mostly conceptual or limited to small samples of cities (19 in the case of Pred, 30 for Zipf), a more recent systematic exploration at the scale of an entire urban system confirmed the intricate interweaving of communication and urbanisation (Bretagnolle and Franc, 2017). Indeed, in their study they proved that in France, the emergence of an integrated national urban system and the take-off of urbanisation were concomitant with the development of a large scale communication network of postal roads around the middle of the 18th century.

The aforementioned body of literature had influence beyond the Swedish and North-American quantitative geography. In the 1990s works from urban sociologists on the global cities (Sassen, 1991) and the networked society (Castells, 1996) were influenced by the conception of cities as centres processing information inherited from the urban literature of the 1950s-1970s. The empirical counterpart of this emerging research tradition, and especially the work of the Globalization and World Cities research group (Derudder et al., 2003; Taylor, 2001), is explicitly focusing on the flows of high order information establishing and sustaining the command and control functions of world cities.

Information flows thus summarize many other interactions, and from that perspective, it is not surprizing that recent research has attempted to redraw the boundaries of regions by looking at the flows of communications between people, for instance through mobile electronic devices (Ratti et al., 2010). Because information is tightly connected to other flows and relations, looking at its patterns allows to study the system itself.

5.2.2 Analysing information flows

Information flows can be studied in many ways. It is possible to study the message itself (a news item, a phone call, a letter, etc.), or the infrastructure that is necessary to carry the message (postal roads, the internet backbone, etc.).

In the case of studies focusing on the messages, newspapers have been among the first sources of data to analyse relations between cities. They were initially used to assess which distant cities were the most salient in the pages of a given city newspapers (Zipf, 1946a), and how much time it took for the information to travel between two places (Pred, 1977, 1973). This can be explained by the fact that newspapers used to be the main sources of information before the democratization of the radio and the television. In more recent periods, newspapers are in competition with many other sources of information (the radio, the television and now social media) but researchers have highlighted that they remain interesting sources. In a paper discussing the scarcity of data to study relations between cities as the global scale, Beaverstock et al. (2000) investigate the potential of using business news taken from a city's newspapers and present this data source "as the answer to the empirical problem of studying medium-term trends in world city relations."

Until very recently, the content was still gathered manually, limiting the scope of such analysis. But an emerging trend of research, to which our paper contributes, explores ways of automating the analysis of cities through their mentions in media or textual materials (Chen et al., 2017; Meijers and Peris, 2018; Salvini and Fabrikant, 2015). While in our case we are focusing on local newspapers receiving information from distant cities (creating a directed relation), recent studies have also explored alternative methods such as looking at co-occurrences of places in documents. This methods allow to retrieve undirected relations by looking at the frequency of co-mentions of toponyms in a given textual corpus such as online news (Hu et al., 2017), Wikipedia articles (Salvini and Fabrikant, 2015) and webpages (Meijers and Peris, 2018).

Human communications can now also be analysed through exchanges between individuals on social media (Decuyper et al., 2018) or via devices such as mobile phones. The use of phone calls to map relations is an old tradition in quantitative geography and urban planning (Board et al., 1970; Zipf, 1946a), but the recent availability of very detailed datasets generated by mobile devices has given a new impulse to such studies of information exchange (Grauwin et al., 2017; Krings et al., 2009; Lambiotte et al., 2008).

Some researchers also use proxies in order to characterize collaboration networks along which information is potentially exchanged. For instance, the world city network literature (see Peris et al., 2018 for a definition of such research schools in urban systems research) focuses on the corporate networks of big transnational firms. These scholars assume that the exchanges of high-order information and knowledge between offices of the same firms located in different places are the main driver of today's economy (Taylor and Derudder, 2015). Another interesting proxy to study exchanges of high-order information between cities is scientific collaborations that can be extracted bibliometric databases (Maisonobe et al., 2016).

Finally, other studies showed that looking at the infrastructure supporting the information exchanges can also be an interesting, albeit indirect proxy of information diffusion. This is true for past networks such as the postal roads between cities (Bretagnolle and Franc, 2017) or contemporary ones such as the internet backbone (Choi et al., 2006).

5.2.3 Regularities in the way information travels

Despite the very different ways in how information is approached in previous studies, some of them have arrived at quite similar conclusions. Here, we list the main findings on information circulation between cities, which can generally be seen as regularities in the way information travels.

The most frequently mentioned feature in information diffusion processes is that the size of the place emitting information, and the distance between the origin of the information and its destination play a major role. This regularity has been initially discovered by Zipf (1946) for news coverage, newspaper diffusion, phone calls and telegram messages. Since then, it has been observed with more contemporary datasets and notably mobile phone communications (Krings et al., 2009; Lambiotte et al., 2008), contradicting the idea that our contemporary society has been characterized by the "death of distance". The explanatory power of 'size' is due to the likelihood of more events occurring in bigger communities than in small communities. The importance of the second dimension – the distance – can be explained by the fact that information coming from different places will not be of the same importance as the value of a news item decreases proportionally with the distance to its origin. Zipf's model assumes that the amount of valuable information received in j from a given populated place i will be in proportion to M_i / D_{ij} where M_i is the population of place i and D_{ij} is the distance between i and j, a formulation very close to the well-known gravity model, of which several

formulations exist for studying information diffusion (Grauwin et al., 2017; Krings et al., 2009; Lambiotte et al., 2008).

Studies have also highlighted the importance of administrative and cultural borders in hampering the flows of communications. This has been observed recently by researchers working on mobile phone calls in several countries (Grauwin et al., 2017). In their study they show that internal administrative borders largely reduce interactions and that such networks exhibit a strong nested hierarchical structure. The authors came up with a "hierarchical model" using the size of the interacting entities and a parameter changing with the probability for two persons from these locations to communicate, this probability being based on the "hierarchical distance" between the locations. Such impact of territorial borders has also been identified with cultural borders in countries, such as the linguistic one (Lambiotte et al., 2008), echoing other researches on the "territorial effect" (Grasland, 2010).

While working on very different datasets, researchers have highlighted very strong regularities in the way information travels. However, these studies were mostly cross-sectional and do not look at how these factors have evolved through time. In this research, we have assembled a massive dataset on interurban information flows based on recently digitalised historical local newspapers. We want to test the findings mentioned above with our dataset that covers a different geographical and temporal context, and add a cross-temporal dimension by looking at how these factors influencing the circulation of information have changed over time. Our hypotheses are:

- **Hypothesis 1 (H1):** Larger cities tend to be covered relatively more in news than smaller cities.
- **Hypothesis 2 (H2):** Geographical distance plays a role in hampering the probability of receiving news from a distant place.
- **Hypothesis 3 (H3):** Over time, information flows cover larger distances and the 'information field' of people increases.
- **Hypothesis 4 (H4):** Cultural and administrative borders hamper the circulation of information.
- **Hypothesis 5 (H5):** The hampering effect of cultural and administrative borders on the circulation of information has decreased over time.

5.3 Empirical data

5.3.1 Newspaper data

For the period we are studying, newspapers can be considered as the backbone of information diffusion; they were the means through which knowledge on more distant places beyond the self-experienced space was spread. For access, one did not have to be subscribed as newspapers were also hung on displays throughout the city. Newspapers were central in shaping geographical knowledge and imaginaries of the wider public (Frémont, 1976) and also played an important role in spreading relevant economic information through advertisements and price reports (Pred, 1977, 1973). As historians have argued: "Mapping the increased circulation of newspapers and mail provides a context for studying news items, adverts, tool catalogues, posters, railway timetables, and letters from loved ones – all communications from afar – and teasing out clues therein about changes in villagers' imagined geography of distant places (capital cities, the seaside, the nearest town, and sites of job prospects a long way from home)" (Schwartz et al., 2011).

The data on cities mentioned in historical local newspapers comes from the DIGGER dataset, a dataset that we created by using Delpher, the digital archive of historical newspapers of the Koninklijke Bibliotheek (the National Library of the Netherlands). This dataset was built by querying massively the catalogue of the library through a search and retrieve protocol and by running named entity recognition (NER) algorithms when necessary in order to correctly identify news items containing place names. The creation of this dataset is described extensively in Peris et al. (2020). Initially, this digital archive contained 12 million of newspapers pages. However, not all sources (newspapers) were of the same importance, for instance as they existed only for a short period. In order to build a relevant corpus we used two criteria related to the spatial and temporal coverage:

- The newspapers had to target a spatially bounded readership (this was necessary to construct origin-destination matrices);
- The newspapers had to be published in at least two consecutive decades (as we are interested in evolution of patterns of information flows)

The first criterion means that we basically excluded national newspapers. Finally, we manually removed two newspapers because many years were missing. This research focuses on the period between 1869 and the end of the year 1930. The initial time mark of this research was selected because it was the year of the abolition of a tax on newspapers – the '*dagbladzegel*' – that reduced significantly their price. The final year (1929) was chosen as it was the last census year available before the Second World War. The Second World War meant a big shock to the Dutch media landscape. Indeed, several newspapers that had been controlled by, and supported the German occupant were dissolved after the war, their place being taken by resistant journals that had started underground. Also, after the war, the share of households having a radio or a television grew very rapidly, competing with the printed press as being the main source of information.

In total, our corpus contains 31 newspapers, located in 24 cities[22]. The newspapers and their main characteristics can be found in Appendix 1. Figure 5.1 shows the location of the places where they were published as well as the 312 cities and towns for which data was collected. These 312 cities and towns are the '*woonplaatsen*' that have a population of more than 10,000 inhabitants in 2013 (this obviously excludes cities in the reclaimed polders of Flevoland province that were built after our study period. More details about the selection of the settlements can be found in Peris et al. (2020).

In this dataset, the information flows are defined as follows: First, we have a set of cities c_i, c_j, c_k, …, c_n, located in a territory. An information flow f_{ij} is occurring when the local readership of the newspaper published in the city c_j can access a news item about another city c_i. Several mechanisms can lead to such an information flow. It can be because an event occurs in c_i and the editorial board of the newspaper publish in c_j thinks that this event is of interest for the local readership, or because firms located in c_i are paying to advertise in the newspaper of c_j. The total amount of information received in city c_j from c_i for a given period t is:

$$F_{ijt} = \sum_{i \neq j}^{n} f_{ijt}$$

22 Amsterdam, Apeldoorn, Arnhem, Assen, Breda, Delft, Den Bosch, The Hague, Doetinchem, Eindhoven, Enschede, Groningen, Heerenveen, Heerlen, Helmond, Hilversum, Leeuwarden, Middelburg, Nijmegen, Roosendaal, Rotterdam, Tilburg, Venlo, Zwolle

TOC

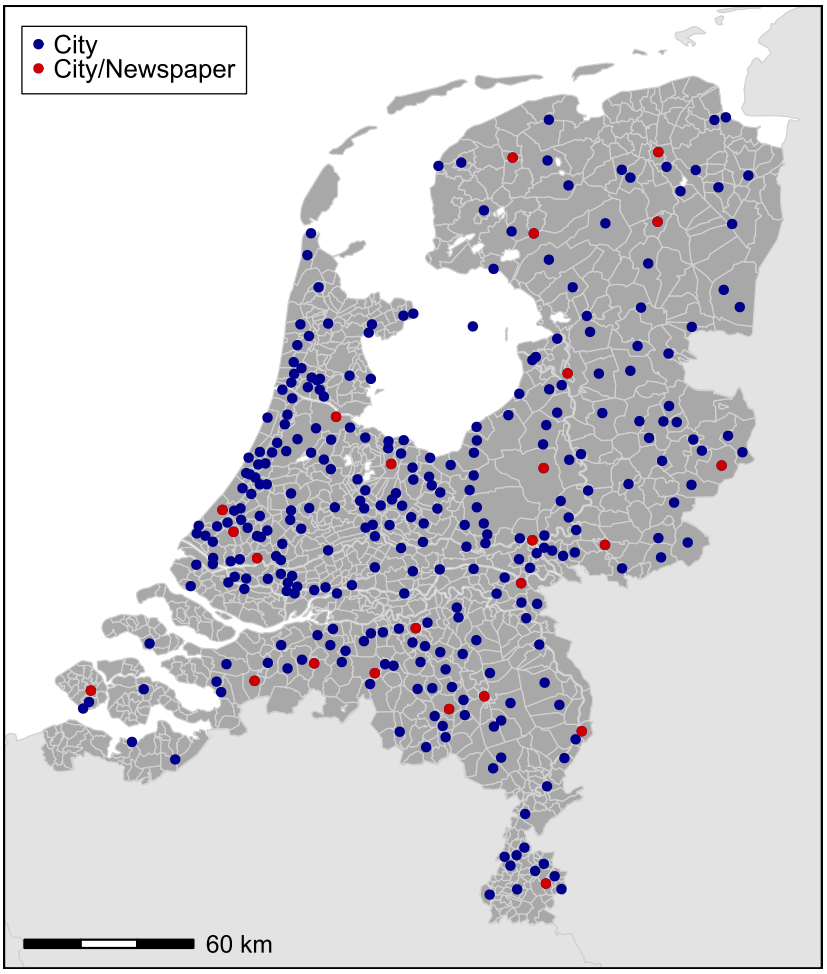

FIG. 5.1 Place of publication of the newspapers and cities for which the data has been collected.

To give an example, Figure 5.2 shows the information field of 6 different cities for the ten year period 1881-1891. The figure shows that an important majority of the information flows are spatially restricted . In the case of Amsterdam and the Hague, a great part of the flows come the main cities of North and South-Holland (Amsterdam, Rotterdam, The Hague and Utrecht), as well as smaller cities such as Haarlem, Leiden and Delft. Some attention is also paid to Groningen in the North and Arnhem in the East. In the case of the newspapers from Leeuwarden, Tilburg, Venlo and Zwolle, a general patterns emerge: an important focus on the medium and small cities from their respective provinces as well a good coverage of Amsterdam. Rotterdam is also well covered, and The Hague to a lesser extent.

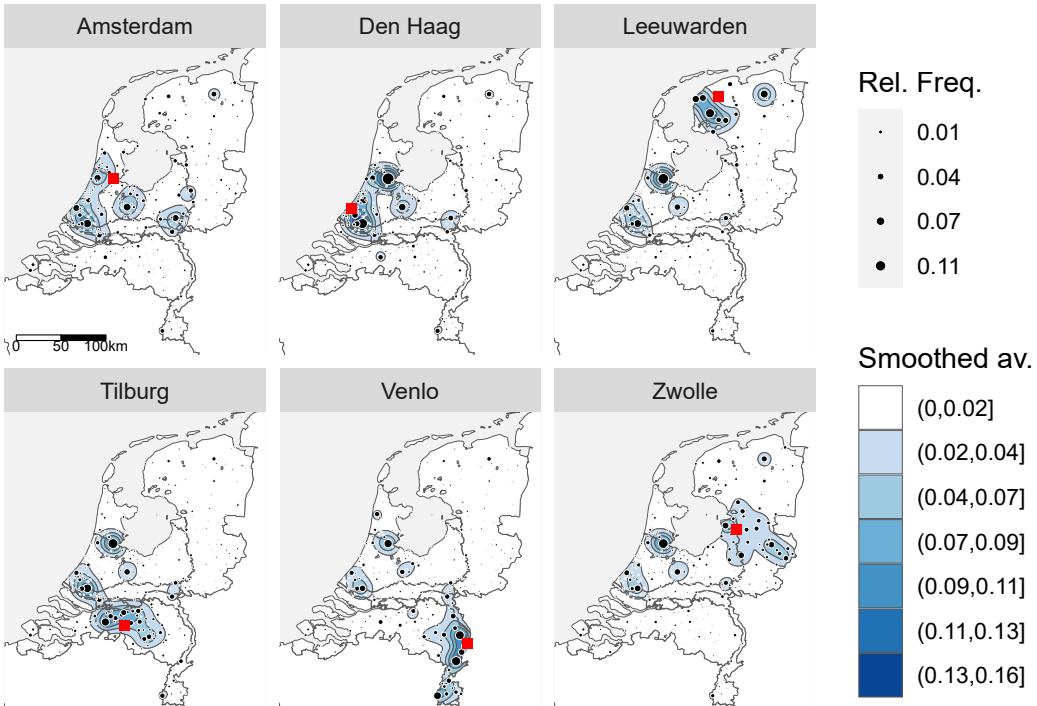

FIG. 5.2 Information field of 6 cities for the period 1881–1890.
The red square represents the city in which the newspaper is published. To highlight clusters of cities frequently mentioned and regional focus, we computed the smoothed average with the Stewart potential function of the R package SpatialPosition (span: 10 km, function: exponential, beta: 2) (Giraud et al., 2019)

TOC

5.3.2 Historical urban populations and boundaries

Defining cities and urban population is an important step in any study on urban systems. The set of cities for which data was collected was defined according to the current population of the country. All the settlements that are above 10,000 inhabitants (a threshold commonly used by statistical agencies and researchers to define an 'urban' place) were taken. The best longitudinal data we could access is the Dutch national census which started in 1795. It is accessible for some years as linked open data through the CEDAR API[23] (Ashkpour et al., 2015). For some missing entries or incoherent patterns (i. e. very sharp increase or decrease of the population between two years of census) we went back to the digitalized versions of the original census books[24] to manually correct the database if needed. The geometries of the municipalities were accessed through the NLGIS API[25]. As we are dealing with longitudinal data, it is important to harmonize the dataset. There are several ways for the harmonization of urban population over time depending on the initial definition of cities (morphological, political or functional), and the spatial resolution of the dataset (Cottineau, 2014). In our case, as we are dealing with irregular areal data with changing boundaries, we decided to go for a method that corrects for the municipalities that have been absorbed over the years by bigger ones. This method implies three steps:

- The spatial delineation of cities for a baseline year (usually the last available)
- The identification of past municipalities that have been absorbed to form the city at the baseline year.
- The spatial aggregation of these municipalities under a certain threshold of population.

It is possible that two settlements are located in the same municipality. In such a case, we summed their frequency of mention in the newspapers and took the population of the entire municipality. Finally, in order to have an approximation of yearly population data, we did a linear interpolation between the different census years. This data on historical urban populations is available on a github repository[26].

23 http://lod.cedar-project.nl/cedar-mini/sparql

24 http://www.volkstellingen.nl/nl/index.html

25 http://nlgis.nl/api/

26 https://github.com/AntoinePrs/information_field/tree/master/data

5.4 Research approach

5.4.1 A gravity framework to model changes in information diffusion

In order to systematize the analysis on such rich and multidimensional data, it is necessary to adopt a modelling framework. For that we use the gravity model, which is widespread in the studies of information flows and spatial interaction in general. The basic assumption behind this model is that flows between spatial units are proportional to the product of their size and inversely proportional to the distance between them. The origin of this model can be traced back to the middle of the 19th century, when people studying migrations and markets suggested an analogy with Newton's law of gravitation to explain the intensity of flows in the proximity of big population centres. But it is really from the 1940s that spatial interaction models of the gravity type started to spread among geographers, economists and engineers. It was for instance in contexts as different as explaining the catchment areas of universities (Stewart, 1942), intercity trips of persons (Zipf, 1946b) or information from distant places whether they come from news articles or telephone calls (Zipf, 1946a). In the following decades, the model also started to be used by engineers and planners in land use and traffic forecasting. The most common formulation of the model is:

$$F_{ij} = k\frac{M_i^{\beta} M_j^{\gamma}}{D_{ij}^{\alpha}}$$

where F_{ij} is the intensity of the interaction between i and j, M_i and M_j correspond respectively to the mass term (i.e. population, number of jobs, etc.) associated with the spatial units i and j, D_{ij} is a measure of distance between them, k is a constant of proportionality and α, β and γ are parameters to be estimated. β and γ can be understood respectively as the potential to generate and attract flows, and α is the parameter corresponding to the 'friction' of distance. The higher this exponent the more distance will play a role in lowering the intensity of interactions. While this model it nothing new, it has recently experienced a regain in interest due to the increasing availability of interaction and flow data (Grasland, 2019; Krings et al., 2009).

The gravity model also received some critiques due to its lack of universality. In a paper, Simini et al. (2012) have criticized the use of adjustable parameters "that vary from region to region" and proposed a universal model – the radiation model – that is parameter-free. According to them, the radiation model could predict a wide range of interactions such as commuting, long-term migration, phone calls and freight flows with a better accuracy than the gravity model. In addition to the fact that other model comparisons did not find that the radiation model outperforms the gravity model in prediction power (Commenges, 2016; Masucci et al., 2013), the context-free dimension of the model can be seen as a limit. Indeed, according to Commenges (2016), the calibration phase in gravity modelling that Simini et al. (2012) present as a weakness, allows in fact to create knowledge on the type of spatial interaction that is studied in a given context and to incorporate this knowledge in the model. This argument is in line with the conception of this model by Burger et al. (2019) which present it as a tool to "gauge" the level of interaction between spatial units. For these reasons, in order to address our hypotheses, we propose to study systematically the changes in parameters value of the gravity model calibrated with our dataset.

5.5 Results

5.5.1 Effects of size, distance and borders over the entire period

We first designed general models taking into account the entire period (1869-1929). The first model is a simple gravity type equation only taking into account the mass of the emitting city:

$$\log\left(F_{ijt}\right) = k + \alpha \log\left(M_{it}\right) + \beta \log\left(D_{ij}\right) + \quad _j$$

Where M_i corresponds to the population of the emitting city, $_j$ represents a newspaper specific fixed-effect, k is a constant and D_{ij} stand for the distance between the emitting and the receiving cities. For this model and the following ones, we use Euclidian distance. Of course, more complex ways of measuring distance could be worthwhile to consider, especially travel time between cities. However, in this paper, we are interested in the distortion from a simple isotropic situation by analysing the patterns of news flows. Further investigations on the role of infrastructure development and technological innovations on the changing patterns

of these interactions would be valuable – because they would touch upon the very vivid debate on the structuring effect of transportation (Offner, 1993; Raimbault, 2020) – but they are out of the scope of this paper. For the implementation of the model we are using a multiple regression method based on an Ordinary Least Square (OLS) regression. As we are using newspapers fixed-effects, our model corresponds to what Wilson (1971) defines as an "attraction constrained" model in his family of spatial interaction models. The hypothesis behind this choice is that news can originate from cities around the publication place in a gravity model way, but the total number of news is known.

The second model introduces a variable capturing the territorial effect associated with provincial borders.

$$P_{ij} = \begin{cases} 1, & if \ \ p_i = p_j \\ 0, & otherwise \end{cases}$$

Where p_i is the province of city i and p_j is the province of city j. Provinces in the Netherlands are not only administrative entities. For most of them, their origin can be traced back to the medieval times. They are inherited structures with a certain degree of functional and cultural coherence. The resulting model is the following:

$$\log\left(F_{ijt}\right) = k + \alpha \ \log\left(M_{it}\right) + \beta \log\left(D_{ij}\right) + \gamma \ P_{ij} + \ _j$$

Finally, in the third generic model, we add a variable capturing another type of territorial effect associated with an assumed hampering role of the Rhine river. $R_{ij} = 1$ for North-North or South-South interactions, and $R_{ij} = 0$ otherwise. This variable is not only supposed to represent the physical discontinuity but also some historically grown cultural differences. For instance, it formed the northern boundary of the Roman empire, and is a gross dividing line between a majority of Catholics to the south of the river, and a majority of Protestants to the north. Being from 'below the river(s)' and 'above the rivers' is a widespread geographical reference used in daily language. The southern provinces of Limburg and North-Brabant were even incorporated later in the country. A the third model aims to catch the effect of this cultural dividing line:

$$\log\left(F_{ijt}\right) = k + \alpha \ \log\left(M_{it}\right) + \beta \log\left(D_{ij}\right) + \gamma \ P_{ij} + \delta \ R_{ij} + \ _j$$

Results of these three models can be found in Table 5.1. They show that only size and distance explain already well how news circulates. The two parameters associated with them are highly significant and 54 % of the variance in the

dependent variable is predicted by these two variables. As expected, city size plays a positive role on the probability of emitting news ($\alpha = 1.09$) and the distance between the news source and where the news is published plays a negative role ($\beta = -0.95$): it is less likely that news from far away appears in a local newspaper. The next two models show that being located in the same province, as well as being located on the same side of the Rhine-Meuse rivers plays a positive role on the circulation of information ($\gamma = 0.5$ and $\delta = 0.33$ for model 3). However, adding these two variables does not increase explained variance substantially due to the fact that they partly capture the distance effect. Based on these results, it seems possible to affirm that the information field of the urban readership of the cities for which we have data is spatially bounded. During the period 1869-1930, urban dwellers are more likely to encounter news from nearby and big centres, which confirms our hypothesis H1 and H2. Based on this data, we can also identify some territorial effects associated with the provincial scale and some inertia of the North-South divide that have been structuring in Dutch history. The hypothesis H3 seems also valid for the period 1869-1930.

TABLE 5.1 Results of the global models

	Dependent variable		
	Log(Fij)		
	(1)	(2)	(3)
k	-2.009*** (0.025)	-3.036*** (0.028)	-3.425*** (0.029)
log(Mi)	1.088*** (0.002)	1.089*** (0.002)	1.087*** (0.002)
log(Dij)	-0.947*** (0.003)	-0.733*** (0.004)	-0.688*** (0.004)
Pij		0.609*** (0.008)	0.498*** (0.005)
Rij			0.331*** (0.005)
Newspaper Fixed-effect	TRUE	TRUE	TRUE
Observations	313,995	313,995	313,995
R^2	0.545	0.552	0.557
Adjusted R^2	0.544	0.552	0.557
Residual Std. Error	1.217	1.207	1.200
F Statistic	11,727.670***	11,717.750***	11,617.010***
Note : *p<0.1 ; **p<0.05; ***p<0.01			

5.5.2 The evolution of information circulation over time

To study the evolution of the circulation of information over time, a fourth model has been implemented. The model integrates time-varying exponents for the different factors hampering the diffusion of news.

$$\log\left(F_{ijt}\right) = k + \alpha \log\left(M_{it}\right) + \beta_t \log\left(D_{ij}\right) + \gamma_t P_{ij} + \delta_t R_{ij} + \varepsilon_j$$

As the result of this model consists of many different parameter values corresponding to the different years, they could not fit in a table. To enhance readability and interpretability, they were plotted in a diagram with the x-axis presenting the year and the y-axis the corresponding value of the parameter (Figure 5.3). Using time-varying exponents improved again slightly the predictive power of the fourth model ($R^2 = 0.57$). In this first specification of a dynamic gravity model, we observe that the parameter associated with distance β_t is becoming less negative between 1869 and 1930. This would suggest a reduction of the friction of distance on the circulation of news over the time period. However, we do not observe a continuous decrease but a slightly fluctuating trend with some drops in the years before the beginning of the 20th century and at the end of the 1910s. In the case of the parameter γ_t and δ_t, associated respectively with being or not in the same province and being on the same side of the Rhine-Meuse river, we observe much more fluctuations over time. Despite these important fluctuations, we can observe a slightly increasing tendency for both γ_t and δ_t. Such results are paradoxical because while β_t is becoming less negative, indicating an increasing probability of receiving news from far away cities, γ_t and δ_t, that are associated with short distance interactions, are increasing. One possible interpretation of these changes in parameters values would be that both short and long distances interactions are increasing, while middle range distance interactions are decreasing. The intermediate places would be sort of jumped over. This interpretation would be in line with previous works on "space-time contraction" that show that in the case of transportations networks, the increase in speed is associated with fewer stops and a weakening of intermediate places and smaller cities (Bretagnolle and Pumain, 2010; Janelle, 1968).

Finally, we implemented a fifth model with a time-varying parameter associated with the mass of the cities M_{it}. This parameter corresponds to the potential of generating information in respect to size.

$$\log\left(F_{ijt}\right) = k + \alpha_t \log\left(M_{it}\right) + \beta_t \log\left(D_{ij}\right) + \gamma_t P_{ij} + \delta_t R_{ij} + \varepsilon_j$$

The goodness of fit of this model is similar to the previous one with $R^2 = 0.57$. Surprisingly, the evolution of the parameter associated with distance is opposite. This means that when controlling for the changing potential to generates news relative to the size, the general tendency indicates a growing probability of short distance interactions. This is not necessarily contradictory with the previous interpretation as part of the long distance interactions that are associated with big cities and its impact on the β_t could be captured by the parameter α_t.

However, we can observe that the trend in the evolution of parameters in a longitudinal gravity model is largely influenced by the specification of the equation and the variables taken into account. This has already been raised by the economic literature that focuses on international trade and has been coined the "distance puzzle" (Brun, 2005). Further works on the applicability of the gravity model in cross-temporal analysis are needed. This also shows the limits of focusing on a single indicator such as the distance friction to summarize complex spatial phenomena. No definitive conclusions can be derived from these dynamic gravity models. However, they provide some hints on a possible space-time contraction and that effects found are not general for all cities, instead suggesting that there is spatial heterogeneity that is hard to capture with general models.

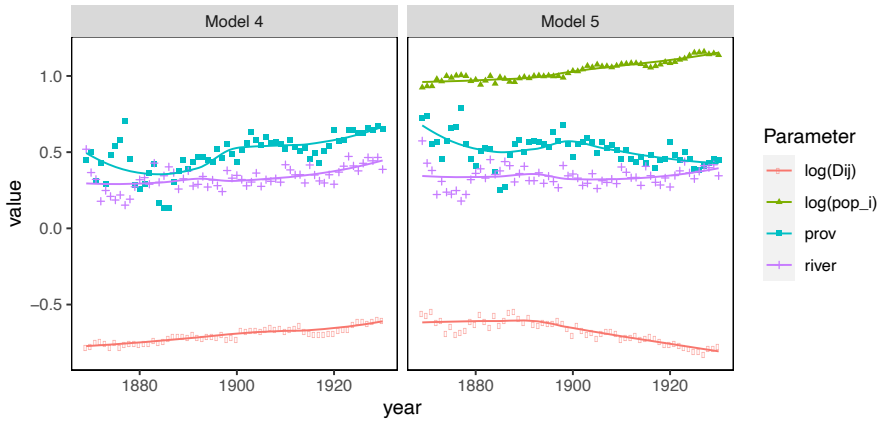

FIG. 5.3 Results of the models 4 and 5

TOC

5.6 Spatial heterogeneity and city trajectories

We have seen in the preceding section that the gravity model does not allow looking at dynamics of individual cities, even though some results suggested that effects may differ for different types of cities. In the following subsections we will look at disaggregated data at the city scale in order to better understand some dynamics suggested by the models. We will first look at the average distance travelled by information flows in the form of news items.

5.6.1 Average distance of information flows

An alternative way of looking at the changes in the spatial extent of the information field of cities is to compute the average distance of information flows for every year and every city for which we have one or more newspapers. For any city j, the index will be computed the following way:

$$\overline{D}_{jt} = \frac{1}{n}\sum_{i=1}^{n} F_{ijt}.d_{ij}$$

For example, if a newspaper based in Delft publishes two news items about Rotterdam (located at 12.9 km), one about The Hague (8.3 km) and one about Amsterdam (54.4 km) at time t, $\overline{D}_{jt} = 22.125$. The results of this index are shown in Figure 5.4. In order to compare the changes for different cities, we have subtracted the value of the first year to all other years. This way, the lines depart from 0. The value of the initial year $\mathrm{D}(\mathrm{t1})$ is presented also in the figure. While this first index is characterized by strong disparities, they are not interesting to interpret as they are mostly related to the relative position of a city toward all other cities. Therefore Amsterdam, Rotterdam, Hilversum and Arnhem have lower values than Eindhoven, Venlo and Groningen that are located close to the borders. Of greater interest is the trend of these lines. For this, we computed a linear regression where the average distance of information flows is considered as a linear function of time. The parameter associated with the slope is displayed on Figure 5.4. One of the first conclusions that can be drawn is that except for two cases (Rotterdam and Middelburg), the spatial information field of all cities have expanded over the period we are studying. This result tends to confirm our hypothesis about the

fact that information flows occur over larger distances (H3). However, there are important variations in the trend of these time series. The cities that have the most increasing trend are Heerlen (0.55), Groningen (0.29), Enschede (0.21), Eindhoven (0.18), Venlo (0.13), Heerenveen (0.13) and Doetinchem (0.1). These cities are all located on the outskirts of the country in the Northern provinces of Friesland and Groningen, in the South (Limburg and North-Brabant) or in the most western parts of the eastern province Gelderland and Overijssel. At the opposite of this trend, cities from the core area of what is today known as the Randstad (composed of the most urbanized areas of North- and South-Holland) have flatter trends. This is especially the case for the big centres such as Amsterdam (0.02), Den Haag (0.03) and Rotterdam (-0.01), but not for the smallest centres such as Delft (0.06) and Hilversum (0.09). So, in peripheral cities interest in the largest cities in the core increased, but this was hardly true vice versa. This trend corresponds to the change in the Dutch urban system described by van Engelsdorp Gastelaars and Wagenaar (1981) as the 'rise of the Randstad'. For these authors the process of political, cultural and economic integration of the Netherlands that took place at the end of the 19th century and in the beginning of the 20th century has resulted in an increasing polarisation toward the core (the Randstad) and a rising contrast between this region and the peripheries (the eastern, northern and southern regions). This process can be related to what van der Knaap (1980) has observed in his study of the Dutch urban system between 1840 and 1970. For him, the development of the inter-city linkages has resulted in an increasing concentration of the population in the largest cities (Amsterdam, Rotterdam, The Hague and Utrecht), where 25 % of the population of the Netherlands were living in 1910, when the polarization was the highest. Our data shows well this polarization process with peripheral cities reporting increasingly more often about distant Randstad cities, while the average distance of news flows to cities in the Randstad core region does not change much.

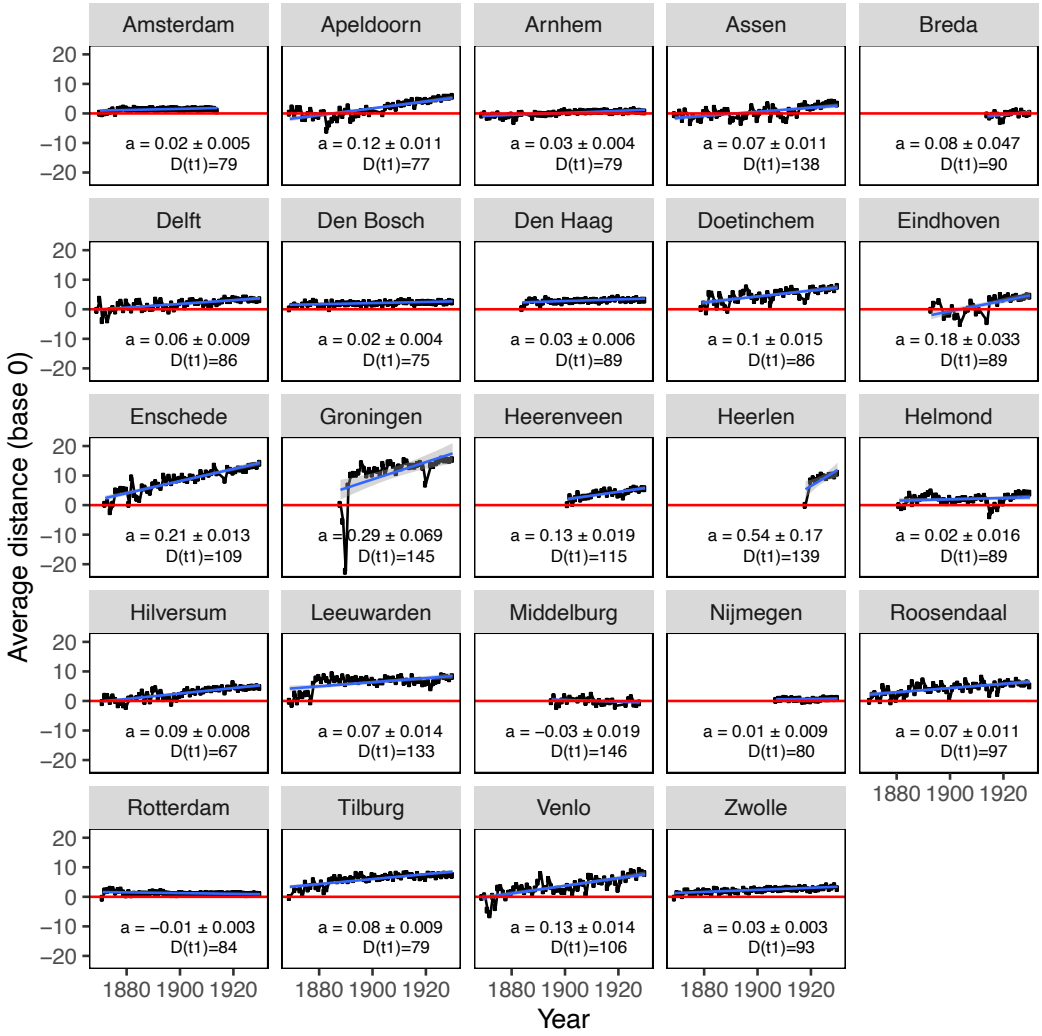

FIG. 5.4 Yearly evolution of the average distance of information flows for each cities

5.6.2 The spatial distribution of origins of information flows

While the average distance of information flows is an interesting index to look at rather general trends, its aggregated dimension does not allow to describe the spatial complexity of the changes in information field. For instance, as Figure 5.5 shows, in the case of the information field of Enschede, one can observe three different processes. First, the close-by cities of Almelo and Hengelo, as well as smaller nearby towns, remain very well covered in all three periods mapped. Second, more and more information from the big cities of the Randstad, and especially The Hague are published. The latter may have to do with the rising importance of the nation state, and The Hague is the seat of Dutch central government. Finally, the intermediate cities of Deventer and Arnhem, that used to be important sources of information in the period 1871-1890 are less important in the period 1911-1930. A quite different pattern can be observed in the case of Rotterdam. The drop of long distance information flows coming from Maastricht, Den Bosch, Nijmegen and Breda tends to confirm the pattern revealed by the average distance index. For more close by interactions, one can see that Den Haag skyrocketed as a source of information for Rotterdam readers.

To systematise this analysis for all the cities for which we have newspapers, we plotted the kernel density distribution of the distance travelled by information flows for 4 different periods of time (1870-1884, 1885-1899, 1900-1914, 1915-1929). Figure 5.6 shows the changes in coverage with respect to distance. One of the most striking patterns that emerges from this visualisation is the very clear increase of information flows coming from the Randstad area (the vertical red lines indicate the distance to the four largest Dutch cities Amsterdam, Rotterdam, The Hague and Utrecht). This is true for almost all the cities except Delft, which tends to cover less and less these big urban centres and Hilversum, which focuses less on nearby Amsterdam and Utrecht but more on Rotterdam and The Hague. Most of the other cities for which we have newspapers tend to focus more and more on the Randstad area. It demonstrates the rise of an economic and political core region in the Netherlands. This attention for nearby urban centres, including close-by medium-sized cities (represented by a green line in the Figure), decreases in relative terms. This is especially visible in the case of Apeldoorn, Arnhem, Den Bosch, Doetinchem, Eindhoven, Enschede, Leeuwarden, Tilburg and Venlo and Roosendaal where the density polygons of the most recent years are very clearly more and more inflated towards the right end of the x-axis of the graph. Notable exceptions to this tendency are Helmond and Groningen that receive more news flows from nearby cities and towns.

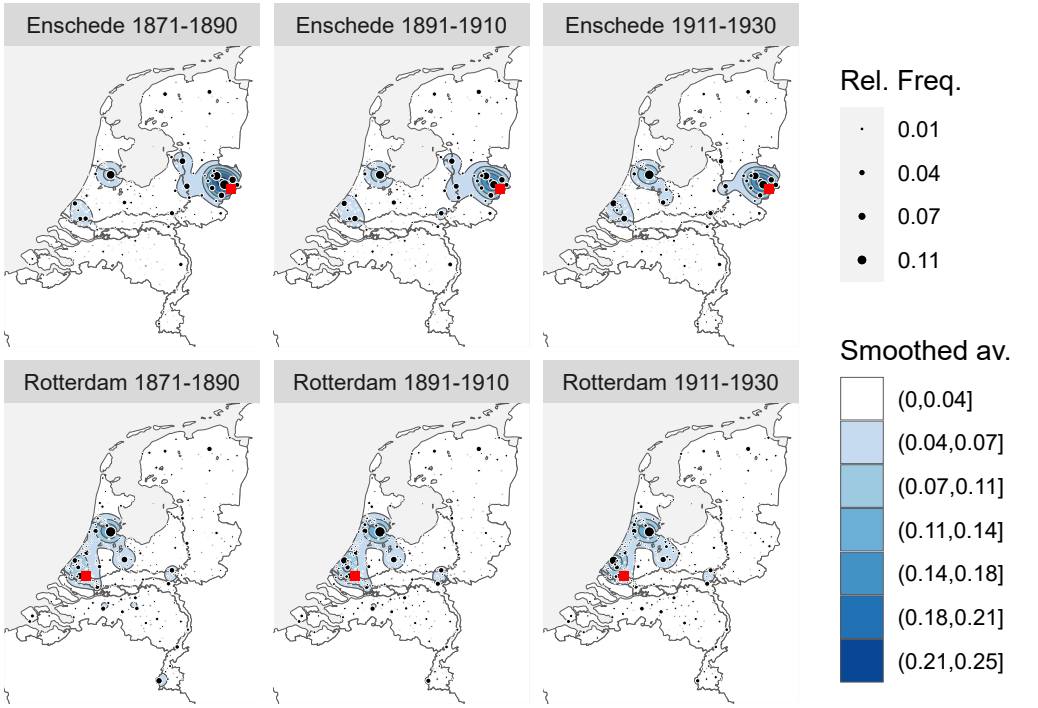

FIG. 5.5 Evolution of the information field of Enschede and Rotterdam.
Smoothed averages were computed with the Stewart potential function of the R package SpatialPosition (span: 11 km, function: exponential, beta: 2) (Giraud et al., 2019).

When analysing the pattern for the three Randstad cities we have data for, an opposite trend can be observed. For the three of them, the cities in the immediate proximity receive more and more attention. This might be caused by what van Engelsdorp Gastelaars and Wagenaar (1981) refer to as the "suburbanisation of the Randstad-centres" that happened at the end of the period we are studying. This phenomenon is described as the expansion of the daily system of the big urban centres due to the congestion in the core city generated by sharp increase of economic activities and population. This could also explain the stagnation or decrease of the average distance of news flows observed previously for the big Randstad cities, with the small settlements located in the periphery of big centres receiving more attention at the same time as being incorporated in wider functional entities.

TOC

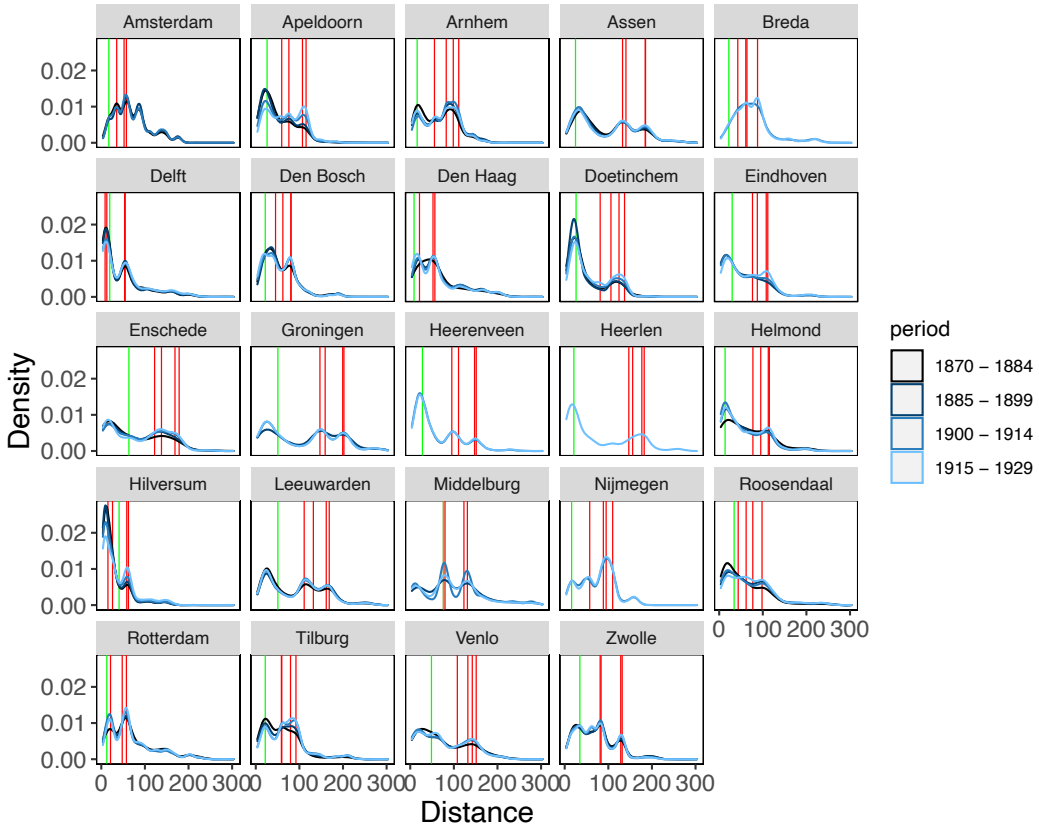

FIG. 5.6 Kernel density distribution of the origin of information flows.
The vertical red lines represent the location of the four big cities of the Randstad (Amsterdam, Rotterdam, The Hague and Utrecht). The green line represents the closest important city (in the top 15 most populated place at least once during the period), that is not located in the Randstad.

5.7 Conclusions and discussions

Our findings tend to confirm previous research that shows that space matters greatly in the process of diffusion of information. We confirmed that for the period between the 1869 and 1930, the size of cities emitting information and the distance between cities are important and significant factors explaining the circulation of information (as hypothesised in H1 and H2). Moreover, the hampering dimension of provincial borders and the North-South divide that have been structuring the Netherlands for a long time remained important over the entire period (as hypothesised by H4).

While we could achieve relatively good predictions with the dynamic gravity models used to analyse the evolution of the circulation of information, we could not draw clear conclusions on the influence of the different factors over time. At this stage of the analysis, we could not confirm the hypotheses H3 and H5 about the evolution of the hampering effect of distance and borders. However, more descriptive analysis on the average distance over which information travels as well as the spatial distribution of the origins of information flows allowed us to confirm the existence of a space-time contraction at this period (H3). We could also identify one of the main driving factors behind this increase of long distance interactions. Our analysis of the information flows between the Dutch cities revealed that the period is clearly characterized by a polarization around its main economic, political and demographic core that from 1938 onwards would be referred to as the "Randstad" (Meijers, 2019), and that became a major focus of the Dutch planning debate in the following decades (van Meeteren, 2020). Newspapers published outside of this area tend to report increasingly often on the Randstad cities, most of the time at the expense of the closer-by medium-sized cities. This process can be related to the hierarchical selection within the urban system in the context of space-time contraction (Bretagnolle and Pumain, 2010; Janelle, 1968). Such a process has also been observed by van der Knaap (1980) in a study on the Dutch urban system that also relates changes in the urban hierarchy to the "increasing scale of the spatial organisation of society".

The data allows for much more analysis and is available for other researchers (Peris et al., 2020). It could for instance be worthwhile to make a clear link with media studies, for instance through exploring the association between the political/religious orientation of newspapers and their spatial information field. Also, while we limited our analysis to simple counts of news items, it may be valuable to also incorporate the content of these news items so that information flows could be classified. This would require employing more advanced content analysis techniques

and machine learning. We see a particularly interesting challenge in relating the image (positive/negative) that is created of particular cities elsewhere and how this has influenced the spatial behaviour of firms and individuals, for instance with respect to migration. Further research could also explore how changes in transportation technology influence the patterns of information diffusion. Finally, the pattern of co-occurrences of place names in news items could be used to identify the relationships between cities, in a similar vein as was recently done for the CommonCrawl Web Archive (Meijers and Peris, 2019).

To our knowledge, this research is among the first systematic explorations of the geographic dimension of digitalised, historical text archives, and certainly the first to explore the spatial dimension of the digitalised newspaper archive in the Netherlands, a resource that has so far been the exclusive domain of the digital humanities. What we essentially show is the feasibility of using a computational social science approach to construct completely novel geographically relevant data sets, allowing us to reconstruct the (evolution of) the spatial organisation of a territory over time. Throughout the world, many programs are currently investing in the digitalisation of such archives. This means that our experiences could be relevant for those wanting to exploit the wealth of geographical information hidden in them.

Appendix

List of newspapers included in the study					
ID	Title	Publication place	First year	Last year	News items
37631091X	De maasbode	Rotterdam	1871	1939	1,110,870
376311770	Tilburgsche cr.	Tilburg	1869	1931	668,214
376312912	Delftsche cr.	Delft	1866	1945	520,290
398540756	Pr. Drentsche en Asser cr.	Assen	< 1865	1950	566,305
398541485	Haagsche courant	Den Haag	1884	1939	1,294,979
398543062	Pr. Overijsselsche en Zwolsche cr.	Zwolle	< 1865	1944	867,612
398825769	De grondwet	Roosendaal	1870	1955	335,136
398831475	Pr. Noordbrabantsche en 's Hertogenbossche cr.	Den Bosch	< 1865	1941	866,093
398831920	Tubantia	Enschede	1872	1942	400,711
399290591	Nieuwe Apeldoornsche cr.	Apeldoorn	1911	1945	250,668
400336960	Pr. Geldersche en Nijmeegsche cr.	Nijmegen	1907	1941	392,495
400337010	Venloosch weekblad	Venlo	< 1865	1898	64,013
400337029	Venloosche cr.	Venlo	1887	1908	52,521
400337088	Apeldoornsche cr.	Apeldoorn	< 1865	1924	210,547
400337266	Eindhovensch dagblad	Eindhoven	1914	1938	215,873
400337274	De Peel- en Kempenbode	Eindhoven	1893	1911	61,536
400337282	De Zuid-Willemsvaart	Helmond	1881	1944	294,252
400337452	Bredasche cr.	Breda	1914	1939	178,837
400337789	Arnhemsche cr.	Arnhem	< 1865	1950	783,105
400383756	Nieuwe Tilburgsche Cr.	Tilburg	1879	1944	1,012,536
400915138	Nieuwe Venlosche cr.	Venlo	1909	1941	241,965
401028933	Limburger koerier : pr. dagblad	Heerlen	1920	1975	546,984
832005797	Nieuwsblad van Friesland	Heerenveen	1901	1994	479,167
832401439	De Gooi- en Eemlander	Hilversum	1871	1950	554,704
83245351X	Limburgsch dagblad	Heerlen	1918	1994	470,076
83249562X	Het nieuws van den dag	Amsterdam	1870	1914	1,697,465
832564818	Rotterdamsch nieuwsblad	Rotterdam	1878	1944	2397235
832915580	De Graafschap-bode	Doetinchem	1879	1947	583,000
833013246	Nieuwsblad van het Noorden	Groningen	1888	1994	1,298,678
852115210	Leeuwarder courant	Leeuwarden	< 1865	1994	1,457,970
852121741	Middelburgsche courant	Middelburg	< 1865	1928	470,537

Acknowledgement

This work was funded through a VIDI grant (452-14-004) provided by the Netherlands Organisation for Scientic Research (NWO), and through the researcher-in-residence program of the Koninklijke Bibliotheek (KB), the national library of the Netherlands. The authors would like to thank the staff of the research department of the KB for their support and welcome.

Bibliography

Ashkpour, A., Meroño-Peñuela, A., Mandemakers, K., 2015. The Aggregate Dutch Historical Censuses. Historical Methods: A Journal of Quantitative and Interdisciplinary History 48, 230–245. https://doi.org/10.1080/01615440.2015.1026009

Batty, M., 2013. The New Science of Cities. MIT Press.

Beaverstock, J.V., Smith, R.G., Taylor, P.J., Walker, D.R.F., Lorimer, H., 2000. Globalization and world cities: some measurement methodologies. Applied Geography 20, 43–63. https://doi.org/10.1016/S0143-6228(99)00016-8

Board, C., Davies, R. j., Fair, T. j. d., 1970. The structure of the South African space economy: An integrated approach. Regional Studies 4, 367–392. https://doi.org/10.1080/09595237000185361

Bod, R., 2013. Who's afraid of Patterns?: The Particular versus the Universal and the Meaning of Humanities 3.0. BMGN-Low Countries Historical Review 128, 171–180.

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. https://doi.org/10.1080/01615440.2016.1237915

Bretagnolle, A., Pumain, D., 2010. Simulating Urban Networks through Multiscalar Space-Time Dynamics: Europe and the United States, 17th-20th Centuries. Urban Stud 47, 2819–2839. https://doi.org/10.1177/0042098010377366

Brun, J.-F., 2005. Has Distance Died? Evidence from a Panel Gravity Model. The World Bank Economic Review 19, 99–120. https://doi.org/10.1093/wber/lhi004

Burger, M., van Oort, F., Meijers, E., 2019. Examining Spatial Structure Using Gravity Models, in: D'Acci, L. (Ed.), The Mathematics of Urban Morphology, Modeling and Simulation in Science, Engineering and Technology. Springer International Publishing, Cham, pp. 471–479. https://doi.org/10.1007/978-3-030-12381-9_21

Castells, M., 1996. The Rise of The Network Society: The Information Age: Economy, Society and Culture. Wiley.

Chen, Y., Yan, F., Zhang, Y., 2017. Local name, global fame: The international visibility of Chinese cities in modern times. Urban Studies 54, 2652–2668. https://doi.org/10.1177/0042098016646674

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. https://doi.org/10.1111/j.1471-0374.2006.00134.x

Commenges, H., 2016. Modèle de radiation et modèle gravitaire - Du formalisme à l'usage. Rev. Int. Geomat. 26, 79–95. https://doi.org/10.3166/RIG.26.79-95

Cottineau, C., 2014. L'évolution des villes dans l'espace post-soviétique. Observation et modélisations. (phdthesis). Université Paris 1 Panthéon-Sorbonne.

Decuyper, A., Gandica, Y., Cloquet, C., Thomas, I., Delvenne, J.-C., 2018. Measuring the effect of node aggregation on community detection. arXiv:1809.08855 [physics].

Derudder, Taylor, Witlox, Catalano, 2003. Hierarchical Tendencies and Regional Patterns in the World City Network: A Global Urban Analysis of 234 Cities. Regional Studies 37, 875–886. https://doi.org/10.1080/0034340032000143887

Frémont, A., 1976. La région, espace vécu. Presses universitaires de France.

Giraud, T., Commenges, H., Boulier, J., 2019. SpatialPosition: Spatial Position Models.

Grasland, C., 2019. International news flow theory revisited through a space–time interaction model: Application to a sample of 320,000 international news stories published through RSS flows by 31 daily newspapers in 2015. International Communication Gazette 1748048518825091. https://doi.org/10.1177/1748048518825091

Grasland, C., 2010. Spatial Analysis of Social Facts, in: Handbook of Quantitative and Theoretical Geography. Advances in Quantitative and Theoretical Geography. Faculty of the Geosciences and Environment of the University of Lausanne, pp. 000–046.

Grauwin, S., Szell, M., Sobolevsky, S., Hövel, P., Simini, F., Vanhoof, M., Smoreda, Z., Barabási, A.-L., Ratti, C., 2017. Identifying and modeling the structural discontinuities of human interactions. Scientific Reports 7, 46677. https://doi.org/10.1038/srep46677

Hägerstrand, T., 1967. Innovation Diffusion as a Spatial Process. University of Chicago Press.

Hohenberg, P.M., Lees, L.H., 2009. The Making of Urban Europe, 1000-1994. 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. https://doi.org/10.1080/13658816.2017.1367797

Janelle, D.G., 1968. Central place development in a space-time framework. The Professional Geographer 20, 5–10. https://doi.org/10.1111/j.0033-0124.1968.00005.x

Krings, G., Calabrese, F., Ratti, C., Blondel, V.D., 2009. Urban gravity: a model for inter-city telecommunication flows. J. Stat. Mech. 2009, L07003. https://doi.org/10.1088/1742-5468/2009/07/L07003

Lambiotte, R., Blondel, V.D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., Van Dooren, P., 2008. Geographical dispersal of mobile communication networks. Physica A: Statistical Mechanics and its Applications 387, 5317–5325. https://doi.org/10.1016/j.physa.2008.05.014

Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., Team, F.N., Cristianini, N., 2017. Content analysis of 150 years of British periodicals. PNAS 114, E457–E465. https://doi.org/10.1073/pnas.1606380114

Maisonobe, M., Eckert, D., Grossetti, M., Jégou, L., Milard, B., 2016. The world network of scientific collaborations between cities: domestic or international dynamics? Journal of Informetrics 10, 1025–1036. https://doi.org/10.1016/j.joi.2016.06.002

Masucci, A.P., Serras, J., Johansson, A., Batty, M., 2013. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. Phys. Rev. E 88, 022812. https://doi.org/10.1103/PhysRevE.88.022812

Meier, R.L., 1962. A Communications Theory of Urban Growth. Joint Center for Urban Studies of the Massachusetts Institute of Technology and Harvard University.

Meijers, E., Peris, A., 2018. Using toponym co-occurrences to measure relationships between places: review, application and evaluation. International Journal of Urban Sciences 1–23. https://doi.org/10.1080/12265934.2018.1497526

Meijers, E.J., 2019. Herkomst van het concept Randstad: de strijd om de locatie van de nationale luchthaven. Geografie Januari 2019, 32–33.

Meijers, E.J., Burger, M.J., Hoogerbrugge, M.M., 2016. Borrowing size in networks of cities: City size, network connectivity and metropolitan functions in Europe. Papers in Regional Science 95, 181–198. https://doi.org/10.1111/pirs.12181

Offner, J.-M., 1993. Les « effets structurants » du transport : mythe politique, mystification scientifique. L'Espace géographique 22, 233–242. https://doi.org/10.3406/spgeo.1993.3209

Peris, A., Faber, W.J., Meijers, E., van Ham, M., 2020. One century of information diffusion in the Netherlands derived from a massive digital archive of historical newspapers: the DIGGER dataset. Cybergeo : European Journal of Geography.

Peris, A., Meijers, E., van Ham, M., 2018. The Evolution of the Systems of Cities Literature Since 1995: Schools of Thought and their Interaction. Netw Spat Econ 18, 533–554. 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. Wiley.

Pred, A., 1973. Urban growth and the circulation of information: the United States system of cities, 1790-1840. Harvard University Press.

Pumain, D., 2011. Systems of Cities and Levels of Organisation, in: Bourgine, P., Lesne, A. (Eds.), Morphogenesis. Springer Berlin Heidelberg, pp. 225–249. https://doi.org/10.1007/978-3-642-13174-5_13
Raimbault, J., 2020. Hierarchy and co-evolution processes in urban systems. arXiv:2001.11989 [physics].
Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., Strogatz, S.H., 2010. Redrawing the Map of Great Britain from a Network of Human Interactions. PLOS ONE 5, e14248. https://doi.org/10.1371/journal.pone.0014248
Salvini, M.M., Fabrikant, S.I., 2015. Spatialization of user-generated content to uncover the multirelational world city network: Environment and Planning B: Planning and Design. https://doi.org/10.1177/0265813515603868
Sassen, S., 1991. The Global City: New York, London, Tokyo. Princeton University Press.
Schwartz, R., Gregory, I., Thévenin, T., 2011. Spatial History: Railways, Uneven Development, and Population Change in France and Great Britain, 1850-1914. The Journal of Interdisciplinary History 42, 53–88.
Simini, F., González, M.C., Maritan, A., Barabási, A.-L., 2012. A universal model for mobility and migration patterns. Nature 484, 96–100. https://doi.org/10.1038/nature10856
Stephens, M., Poorthuis, A., 2015. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. Computers, Environment and Urban Systems, Special Issue on Volunteered Geographic Information 53, 87–95. https://doi.org/10.1016/j.compenvurbsys.2014.07.002
Stewart, J.Q., 1942. A Measure of the Influence of a Population at a Distance. Sociometry 5, 63. https://doi.org/10.2307/2784954
Taylor, P.J., 2001. Specification of the world city network. Geographical analysis 33, 181–194.
Taylor, P.J., Derudder, B., 2015. World City Network: A Global Urban Analysis. Routledge.
Törnqvist, G., 1970. Contact systems and regional development. Lund University Press.
Törnqvist, G., 1968. Flows of Information and the Location of Economic Activities. Geografiska Annaler. Series B, Human Geography 50, 99–107. https://doi.org/10.2307/490320
van der Knaap, G.A., 1980. Population growth and urban systems development: a case study. M. Nijhoff.
van Engelsdorp Gastelaars, R., Wagenaar, M., 1981. The rise of the 'Randstad', 1815–1930, in: Patterns of European Urbanisation Since 1500. Routledge.
van Meeteren, M., 2020. A prehistory of the polycentric urban region: excavating Dutch applied geography, 1930–60. Regional Studies 1–14. https://doi.org/10.1080/00343404.2020.1800629
Wilson, A.G., 1971. A family of spatial interaction models, and associated developments. Environment and Planning A 3, 1–32.
Zipf, G.K., 1946a. Some Determinants of the Circulation of Information. The American Journal of Psychology 59, 401–421. https://doi.org/10.2307/1417611
Zipf, G.K., 1946b. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. American Sociological Review 11, 677–686. https://doi.org/10.2307/2087063

TOC

6 Mapping functional regions in the Netherlands by analysing individual residential and job histories

This chapter explores the potential of individual level (micro) data for urban systems research. In contrast to the other empirical chapters, this paper has not yet been published, nor submitted, and it is still shaped as a research report documenting initial findings of this exploration. Since I believe that these are interesting and relevant, I decide to include this chapter nevertheless. I am thankful to Clémentine Cottineau, Evert Meijers and Maarten van Ham for the fruitful discussions that helped me progressing in this research project.

ABSTRACT

In this chapter, we present an alternative to the use of commuting flows in order to define functional regions from a multiscalar perspective by considering people that move without changing jobs ('stable workers'). Because these people continue working at the same work place but commute from two different locations, we can deduct that these two locations belong to the same labour and housing market area. Using a hierarchical clustering algorithm, we extract the nested hierarchy of functional regions in the Netherlands by using these residential migration flows. By far most studies that discern functional regions use aggregated data, but one of the main advantages of our method is that it is based on micro-level or individual data, which allows to explore how functional regions differ for different type of people, hence taking into account individual level heterogeneity. Here we focus on the difference in housing and labour market areas of younger workers compared to more senior workers.

6.1 Introduction

The question of the scale of the city is a lively debate in urban geography. Identifying the right scale at which the city functions as a coherent entity is very important to target the right area for policy intervention or service calibration. Over the past few years, a number of studies have highlighted the limitations of using singular and arbitrary choices of spatial scales. For example, work on the impact of city definitions on the scaling exponent of the relation between city size and urban quantities assessing performance (Arbabi et al., 2019; Cottineau et al., 2017), or the importance of scale in the measurement of exposure to the socio-spatial context (Petrović et al., 2018). In the Netherlands, the ways cities are conceptualised have had important implications on spatial planning policies. The traditional focus on polycentric urban regions like the Randstad has recently given way to a focus on individual cities because of the assumed importance of agglomeration economies. However, according to Poorthuis and Meeteren (2019), the studies supporting this debate "often compare apples to oranges by ignoring the definitional fluidity in the geometry and the scale of the city".

There is consensus in urban research that cities should be defined on a functional basis (Lobo et al., 2020). Cities emerge from a multitude of interactions among which labour and housing markets play an important role. They can be conceived as 'daily urban systems', which is the territory of daily interactions (Pumain, 2011). Their shape is largely dependent on the distance which can be covered by a human during a day, an element that is influenced by population distribution and transportation technologies. However, the capacity people have to move around the city is largely influenced by individual factors. Research on mobility in the Netherlands and the Randstad has shown that age, personal income, and level of education are positively associated with commuting between urban regions in comparison with commuting within urban regions (Burger et al., 2014). Other research has shown that teleworking leads to longer commutes, and therefore affects the geography of labour markets, especially for managerial positions and knowledge workers (de Vos et al., 2018). These studies echo the literature on dynamic segregation in cities and metropolitan regions showing important differences in the functional space of income groups (Le Roux et al., 2017; Netto et al., 2018).

Functional regions attached to cities are usually defined in a two-step process: the selection of a core city based on a minimum density threshold, and the identification of a surrounding commuting zone based on the minimum share of commuters traveling to the centre. However, in the Netherlands, exhaustive commuting data

does not exist which makes it difficult to delineate FUAs and have precise information on the commuters. Even more important however, is that departing from the idea of having a core city with a functional hinterland is increasingly obsolete given the dominance of polycentric urban structures. There are often multiple cities making up one functional area.

Therefore, in this paper we develop an approach inspired by housing market area delineation in order to delineate functional regions. Housing market areas are defined by O'Sullivan et al. (2004, p. 42) as "the geographical area within which most people both live and work and where most people moving home (without changing job) will have sought a house". This idea resembles an earlier conception of housing market area by Bourne that was never applied because there was no data to implement this definition of a "contiguous geographical area, more or less bounded, within which it is possible for a household to trade or substitute one dwelling unit for another without also altering its place of work or its pattern of social contacts (Bourne, 1981, p.71)".

Our aim here is to empirically delineate functional regions in the Netherlands, by using a multi-scalar perspective, and allowing for the exploration of different functional spaces of socio-demographic groups. This paper presents the initial steps necessary to fulfil this bigger aim. We present a method to build a multiscalar delineation of functional regions in the Netherlands by using microdata and we explore group heterogeneity by focusing on people moving between the defined regions but keeping their job. We refer to these people as 'stable workers'. Because people continue working at the same place but commute from two different locations, we assume that these two locations belong to the same labour and housing market area. For this study we use detailed micro-level data of Statistics Netherlands.

In the next part, we detail our data collection strategy to create our network of people moving without changing job and present summary statistics of these movements (2). Then, we present a method to derive the nested hierarchy of functional regions using these flows (3). In a fourth section, we present the results of the functional regionalisation for the entire group of stable workers and lay the foundations of the analysis of the heterogeneity of these functional relations by looking at the patterns of movement of two different population groups (4). Conclusions of these explorations are then presented, and followed by possible directions for future research (5).

6.2 Residential relocation of stable workers as an indication of functional distance

6.2.1 Data

Functional regionalisations are based on measures of 'functional distance', that summarise the intensity of relations between two nodes or places. They are usually distortion of Euclidian distance. Our operationalization of functional distance for delineating regions is based on the concept of spatial arbitrage. We consider that an individual moving without changing job will consider the new place of residence to be an appropriate substitute for accessing the workplace, and consequently that the former and new place of residence belong to the same housing and labour market. The aggregation of these movements between 'equivalent' places constitutes our measure of functional distance. In this study, functional regions are then defined as the area in which the majority of these movements are self-contained.

The data that we used is anonymised individual level data extracted from several administrative registers and provided by Statistics Netherlands (CBS). It covers the population living in the Netherlands. For demographic information such as age, place of residence and residential migrations, we used different files made from municipal registers[27]. As we are interested in the working-age population, we selected people between 18 and 65 years old for every years used in our study (2011-2018). Data on residential migration was extracted from the successive addresses of individuals. For privacy reasons, the addresses are encrypted but can be connected to 100 and 500m grid cells in order to spatialize the information.

Data on individual employment is derived from another dataset that is based on the information collected by the tax authorities of the Netherlands[28]. It contains monthly data on all employment contracts of people working in the Netherlands. The link between individuals and companies can be derived from an identifier representing

27 Gbapersoontab for demographic information and Gbaadresobjectbus for the successive places of residence.

28 The dataset is called Spolisbus

TOC

an employee-employer relationship. To account for cases of mergers or acquisitions that could result in a changed job ID while someone actually continues to work under the same conditions, we used another file allowing to correct for such organisational changes[29] that hence provides more consistent job IDs over time. From this data, we could extract contract duration, hours worked, monthly income and the economic sector people are working in (the categorization is based on collective labour agreements).

Thanks to this data it was possible to know when someone has moved from a place *A* to a place *B* without changing job. Individuals moving without changing jobs will be referred to as 'stable workers' in the rest of this chapter. Figure 6.1 describes the three steps necessary to create the network of residential migration of such stable workers.

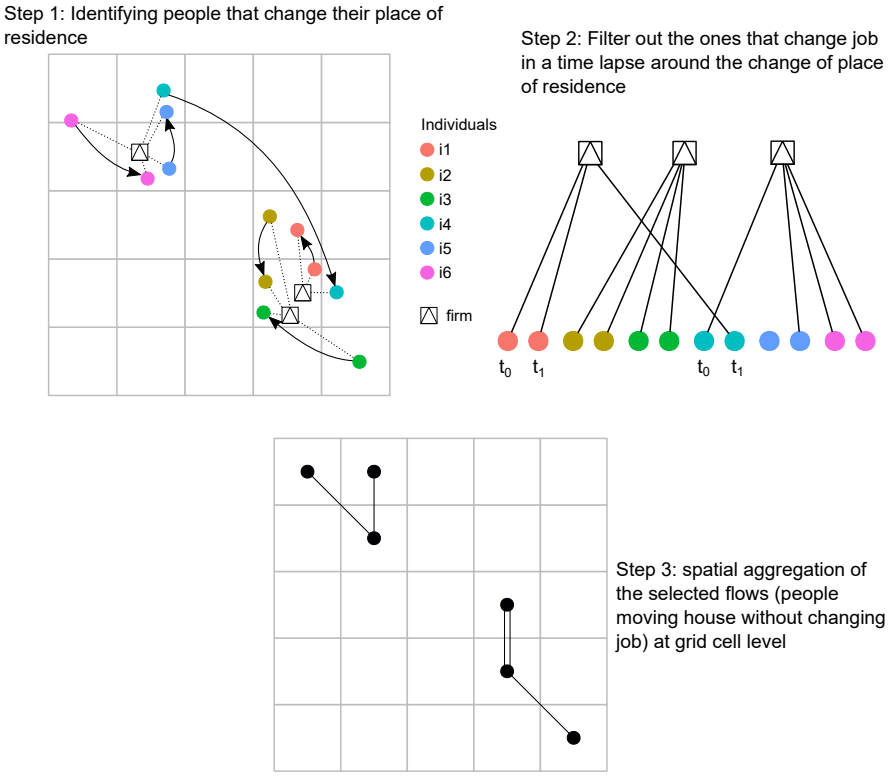

FIG. 6.1 Steps for the creation of the network

29 The Spolislongbaantab

6.2.2 Description of the sample of stable workers

In order to create our group of stable workers, we first filtered out people that are working less than 12 hours per week, which, according to Statistics Netherlands (CBS) is the threshold for participation in the labour force. Then, we looked at people occupying the same job position in a company without discontinuity during the period ranging from 6 months before and 6 months after the moving date. We collected this data for 8 years, between 2011 and 2018. Descriptive statistics of these moves can be seen in Table 6.1. In total, our data contains 10,3 million of residential moves for a period of 8 years. Of these, 3,9 million movements (37.6%) involve stable workers. 3,3 million (32.2%) could be characterised as 'unstable workers': people for whom a residential move cam coupled with a change of jobs. The remaining 30.2% percent involve residential moves of people not active in the labour force.

Clear differences between those groups appear when looking at average distance of a residential move. On average, a worker changing jobs will move 25 km away from its origin while a stable worker will move only 11 km away on average. It was to be expected that movements of stable workers stay more local. When looking at the median distance of a residential move, we can see that 50% of all movers move less than 4 km away, with again a sensible difference between the workers changing jobs and the one keeping their position (respectively 5.5 and 3 km).

TABLE 6.1 Distance of migration for stable and unstable workers (in km)

Group	Number	1st Qu.	Median distance	Mean distance	3rd Qu.
All migrants	10,396,690	1.3	4	19.25	17.59
Stable workers migrants	3,907,518	1.08	3.04	11.48	9.79
Workers changing job	3,351,228	1.65	5.55	25.33	30.68

Many studies have shown important differences in terms of what we could call 'spatial flexibility' of individuals, which seems to depend on their socio-economic and demographic characteristics (Burger et al., 2014; de Vos et al., 2018). To get an idea of whether such spatial flexibility would translate in the distance of residential migration, we computed summary statistics for different population groups: two age groups (stable workers between 25 and 34 years old, and between 40 and 54 years old), and within these two groups, workers earning more that the median growth monthly income, and workers earning less. In the Netherlands, income and educational level obtained highly correlate (Moonen et al., 2011). For the younger group, the median gross monthly income is set at ~2660 €, while for the older group, it is set at ~3200€.

We can observe differences in terms of distance of residential moves of stable workers depending of their age and income. The group that is likely to move furthest away comprises relatively younger workers, who have above-median earnings compared to their contemporaries, probably due to being higher educated. At the other end, the group that stays closest to their current residential location when moving involves workers between 40 and 54 earning less than the median income.

While it is important to keep an eye of those differences following from individual-level heterogeneity, we will nevertheless explore the potential of this data to delineate functional regions for the entire population of stable workers in the next section.

TABLE 6.2 Summary statistics of residential migration for different groups of stable workers

	Group	N	1st Qu.	Median distance	Mean distance	3rd Qu.
Stable workers per age group	25-34	1,748,542	1.12	3.09	10.26	9.62
	40-54	860,001	0.91	2.55	9.82	8.15
Stable workers per age group and income category	25-34 & < median income	873,369	1.06	2.91	9.66	8.51
	25-34 & > median income	873,324	1.20	3.30	12.26	11.16
	40-54 & < median income	429,991	0.85	2.40	8.04	6.99
	40-54 & > median income	429,990	0.94	2.75	11.60	9.73

6.3 Delimiting functional areas

6.3.1 Clustering approach

In this section, we apply a method to delineate functional regions in a bottom-up way by aggregating all the migrations of individual stable workers into a network of residential moves, and subsequently define clusters in this network. Partitioning space into sets of optimum regions is an old problem in geography (Haggett and Chorley, 1969). This process allows to reduce the complexity of the data by creating meaningful spatial units. The main principle behind any type of functional regionalisation is to merge elementary spatial units when they are characterized by high levels of interactions with each other. It allows to create a new regional ordering at higher spatial scales, and varying the level of interactions allows for a multiscalar approach to defining functional regions.

Several approaches exist to delineate functional regions. They can be grouped in two main groups: rule-based approaches and clustering. First, rule based approaches are characterized by the definition of initial criteria and thresholds – in other words, rules - in order to select spatial units and group them. The 'area of attraction' of French cities (INSEE, 2020) or the OECD/EU Functional Urban Regions (Moreno-Monroy et al., 2020) are based on such principles. They identify a core with a density criterion, and select surrounding areas which reach a certain share of residents commuting to the centre. Second, clustering approaches are based on the idea of grouping objects in a such a way that maximizes similarity of the objects within each group compared to the outside group(s). Consequently, this is a bottom-up approach. In the context of the definition of functional regions, one can think of network percolation (Arcaute et al., 2016; Raimbault, 2019), or the many variants of hierarchical clustering used for defining Travel-to-Work areas in Britain (Coombes et al., 1986), commuting regions in Sweden (Landré, 2012), housing markets (Brown and Hincks, 2008) and functional economic regions in Australia (Mitchell and Watts, 2010).

The choice of the method depends on the nature of the data and the objective of the regionalisation. In our case, our objective is to create multiple definitions with a bottom-up perspective, making the clustering approach more appropriate. Moreover, because we do not depart from a predefined population centre, this approach is better suited for a territory characterised by polycentricity.

In its original form, our data corresponds to a weighted directed network where the nodes are the centroids of grid cells and the weights of the edges correspond to the number of stable workers that have migrated between the two grid cells. As migrations are relatively rare events and because the number of possible combinations of origins and destinations corresponds to the square of the number of spatial units, the edges' weight had a flat-tailed distribution. After transformation, we ended up with a sparse matrix. To reduce the size of the matrix and consequently the computing time, we aggregated the initial spatial units (100 by 100 m) into 1 km grid cells. The Intramax procedure appeared as the most adapted to our data. This clustering algorithm is particularly well adapted to small scale units with low values because its iterative mechanism allows to create a snow-ball effect.

The intrazonal maximization or Intramax procedure, is an algorithm of hierarchical clustering (Masser and Brown, 1975; Masser and Scheurwater, 1980). It has been designed in order to create functional regions that maximise the self-contained flows within the aggregations of basic data units and that minimize incoming and outgoing flows. It is a bottom-up procedure that starts with the basic spatial units and that merges pairs at every steps until all the units are grouped into a single cluster. At every iteration of the regionalisation mechanism, it looks for the pairs of spatial units for which the objective function is maximized. The objective function is computed in the following way:

$$max \ \frac{F_{ij}}{O_i D_j} + \frac{F_{ji}}{O_j D_i}$$

Where

$$O_i = \sum_j Fij$$

And

$$D_j = \sum_i Fij$$

As Eq.1 shows, the aggregation procedure is based on the relative strength of the interaction between two pairs of units, which allow to overcome size effects. For N number of units it results in $(N-1)$ possible number of clusters. In many studies, the number of clusters is selected arbitrarily after visual inspections, however some researchers (Landré, 2012; Mitchell and Watts, 2010) have proposed to identify

breaking points in the evolution of the overall self-contained interactions that indicate important merges in the system. Similarly to the phase transitions in a hierarchical percolation process (Arcaute et al., 2016), they can be interpreted as a change in scale of analysis.

A well identified issue with the Intramax procedure is that it often gets trapped in local optima resulting in fragmented regions (Alvanides et al., 2000). In order to avoid this problem, we removed from our data the spatial units with very low number of inhabitants (< 150 inh./km² in 2017) and low numbers of outgoing stable workers that could bias the results (< 25 outgoing migrations of stable workers during the 8 years period).

The Intramax analysis was performed with *R*. We adapted a program from Charlton (2015) by adding the identification of breaking points to perform our multiscale analysis. The computation of self-containment statistics as well as the visualisations were also done with *R*.

6.4 Results

6.4.1 Functional regionalization for the whole population of stable workers

Figure 6.2 shows the evolution of the origin-based self-containment statistics at each iteration of the algorithm. In the initial situation, about 54 % of the outgoing flows are self-contained within the 1 km grid cells. At the end of the procedure, when all the spatial units are merged into a single cluster, 100 % of the flows are obviously self-contained.

In Figure 6.2b, the number of clusters at significant breaking points is pictured. The first important breaking point that we identified is at 119 clusters, at this stage of the procedure, the share of self-contained flows is around 75%. Big cities did not yet merge into unique clusters, which is due to the relative strength of interactions between two cells in a big city being relatively low because of the many opportunities for relocations with other cells that exist. As we cannot go through all the steps of

the analysis, we focus on the most important breaking points. The dendogram in Figure 6.3 shows the successive merges of clusters from the step with 40 clusters to the final iteration resulting in one cluster. This Figure allows to visualise the nested organisation of functional regions in the Netherlands according to the relocation of stable workers. The spatial footprint of these clusters is also displayed on the maps of Figure 6.4 for six different iterations of the Intramax procedure.

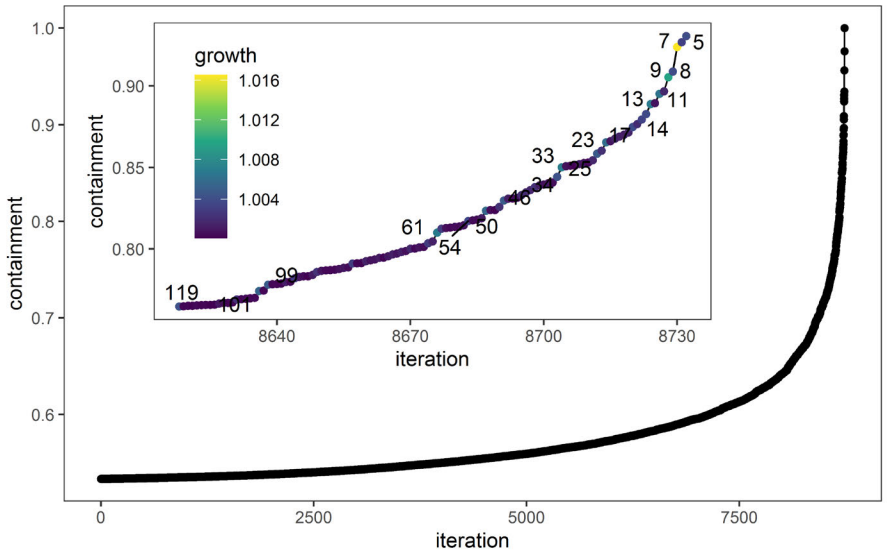

FIG. 6.2 Statistics on self-containment of the flows at different iterations of the algorithm

Looking at the results of this procedure starting from the end gives a good picture of the functional organisation of the country. The first split that can be observed (C=2)[30] resembles to the very well-known dividing line *'beneden de rivieren/boven de rivieren'* (under the rivers and above the rivers). This line corresponds more or less to the limit of the Roman expansion and later on to the border of the Spanish Netherlands, and is often presented as an important socio-cultural divide within the country, also reflecting a dividing line between Protestants in the north and Catholics in the south. However, the border obtained with our functional regionalisation does not perfectly overlap with the rivers. The most northern part of North-Brabant

30 We write C=2 for the step that resulted in two clusters, C=3 for three clusters, etc.

and Limburg belongs to the northern cluster. Another interesting element is that Zeeland, that was historically more attached to the Northern provinces, belongs to the southern cluster. Surprisingly, Goeree-Flakkee, the Northern island of the Rhine-Meuse-Scheldt delta, although being close to Zeeland from a geographic and cultural point of view, is functionally more connected to the Northern cluster, meaning that the provincial borders of South-Holland makes sense.

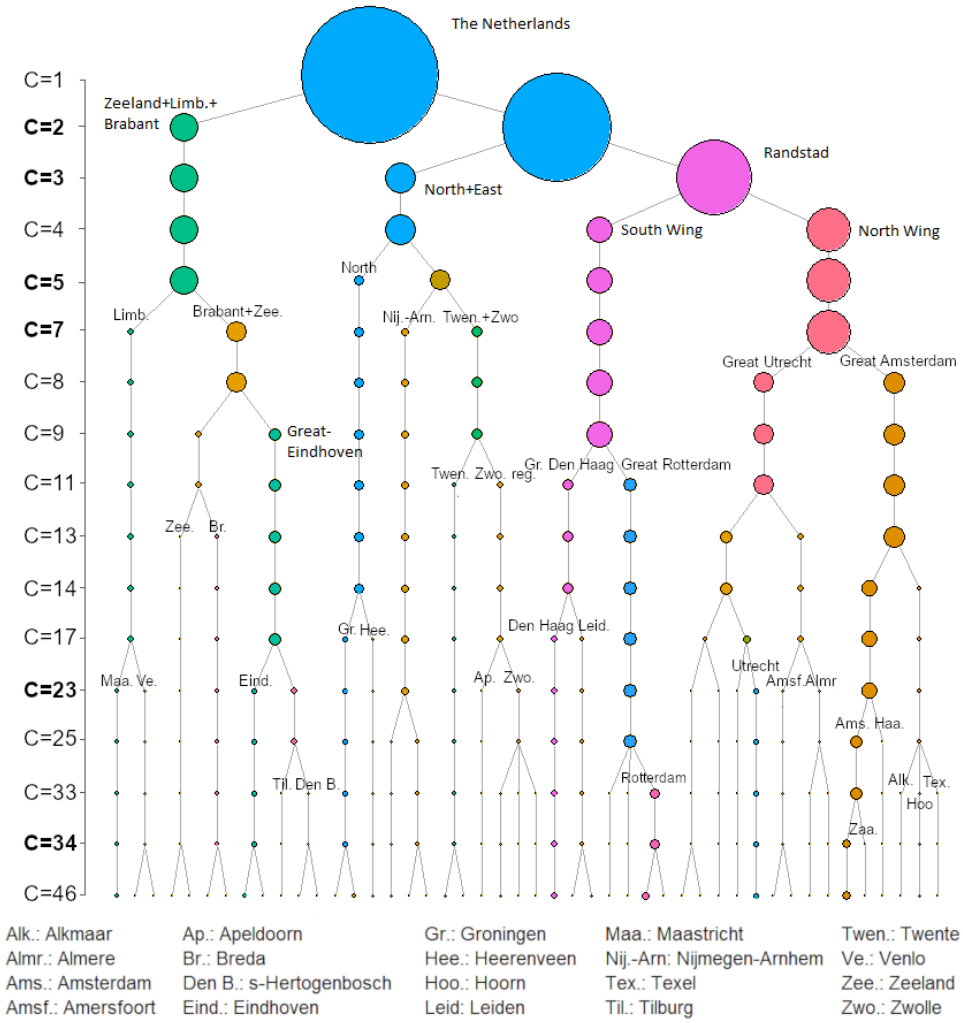

FIG. 6.3 The nested hierarchy of functional regions as extracted from the Intramax procedure.
The size of the dots represents the number of self-contained flows within a cluster. Colours correspond to the ones used for the maps in Figure *6.4. For readability reasons, not all iterations are pictured. We show here only important breaking points between the one resulting in 40 clu*sters (on the bottom of the figure) and the one resulting in 1 cluster (at the top of the figure).

TOC

For C=3, we observe a split between the North and East and an area covering the provinces of North-, South-Holland, Utrecht, the south of Flevoland and parts of Gelderland. This functional region corresponds to an extended definition of the Randstad that extends more far than we commonly assume.

For C=4, this extended Randstad splits into a North and a South Wing. The South Wing includes Rotterdam and The Hague, but also Dordrecht, Delft and Zoetemeer. Leiden, often presented as a link between the North Wing and the South Wing is integrated within the southern functional region. The North Wing, covers Amsterdam, Utrecht, Haarlem, Hilversum, part of Flevoland (including Almere and Lelystad), and goes up to Alkmaar, Hoorn and even Den Helder. Noticeable is also that this functional region extends over a much larger area than the functional area of Rotterdam-The Hague.

At C=5, the large North-East region splits in two parts. While it is common to speak about the North of the Netherlands by referring to the three provinces of Groningen, Drenthe and Friesland, our analysis shows that the southern edges of this ensemble (Meppel and Hoogeveen) are actually more functionally integrated with Overijssel and Gelderland.

For C=7, we were surprised by the fact that Zeeland and Noord-Brabant stay a functional entity while Limburg splits off, because of perceived greater cultural coherence between Limburg and Noord-Brabant. It is also remarkable that, in contrast with some planning ambitions of the Eastern provinces that present the axis from Zwolle to Nijmegen as structuring a structuring element of the East, parts of this axis end up in different clusters.

Regarding major urban corridors, we can also observe that a Greater Amsterdam and a Greater Utrecht region appear at C=8 while Rotterdam and The Hague only split at C=11, showing that the latter two form a much more coherent entity from a functional point of view.

The map for C=23 presents an extended Rotterdam metropolitan area that goes all along the Meuse estuary between Dordrecht and Hoek van Holland and that includes both banks of the river. Although some parts such as Westland are adjacent to The Hague, they belong to the functional region of the port city of Rotterdam at this stage of the procedure. We can also observe a very large cluster around Amsterdam. It covers both sides of the Amstel, with Zaandam and Purmerend being included, and goes up to Haarlem. In the East, we now observe a split between the two-pole region of Apeldoorn and Deventer and the one of Twente (Almelo, Hengelo and Enschede). The large sphere of influence of Zwolle gets also separated from the rest of the East. This area covers also the Noordoostpolder and goes up to the German border by including some of the Southern edges of the Drenthe and Friesland provinces.

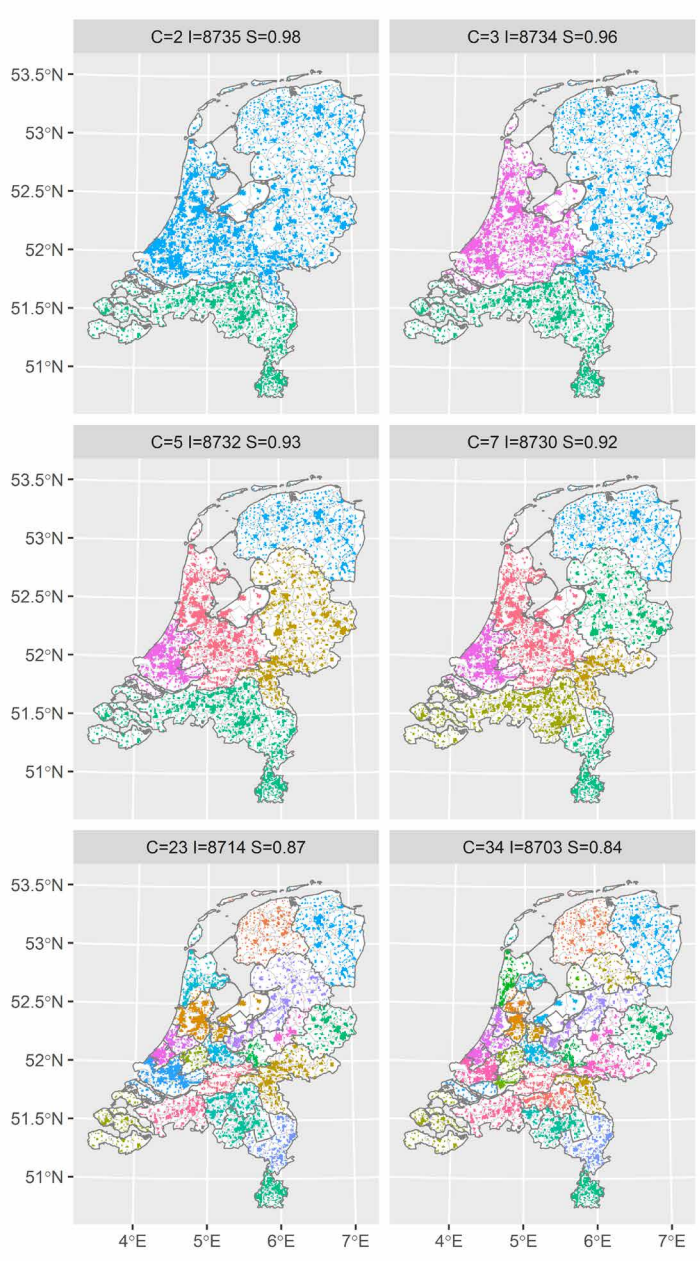

FIG. 6.4 Spatial representation of the clusters at 6 different iterations of the Intramax procedure. *C corresponds to the number of clusters, I is the number of iteration before arriving to this result and S is the overall share of self-contained flows. Because the maps show many clusters with redundant colours, we highlighted the clusters with light grey boundaries showing the fusion of the municipalities covered by the clusters.*

TOC

6.4.2 Comparison of 34 clusters with arbeidsmarktregio's

The map resulting from the Intramax procedure with C=34 manifests resemblances with the officially used delimitation of labour market areas in the Netherlands, the so-called arbeidsmarktregio's ('labour market areas' in Dutch) of which there are 35. Arbeidsmarktregio's had to be defined after a 2012 Law (Wet Structuur Uitvoeringsorganisatie Werk en Inkomen) required a single registration system for each labour market where all labour supply and demand were gathered and matched, and in each region all relevant stakeholders were supposed to define and execute a strategy for labour market development. This cooperation is supposed to provide service to employers by providing a single point of contact in each region for labour issues. Next to employer's organisations, the main stakeholders are the municipalities and the Employee Insurance Agency (UWV), an autonomous administrative authority commissioned by the Ministry of Social Affairs and Employment to implement employee insurances and provide labour market and data services. It is important to note that this definition of labour markets is not based on data, but on perceptions of those stakeholders about what constitutes their labour market region. There is a transfer procedure for municipalities when they decide to join a neighbouring labour market region, which occasionally happens. In 2018, the last year included in our data, there were 35 arbeidsmarktregio's. It is interesting to compare our bottom up functional delimitation based on 'stable workers' with the delimitation based on perceptions of local stakeholders in order to see how empirics and perceptions match.

Figure 6.5 compares the two regionalisations. We named our functional regions extracted from the Intramax procedure for 34 clusters after the name of the most populated municipality included in them. First, it has to be highlighted that the boundaries of our functional regions and the ones of the arbeidsmarktregio's sometimes overlap well. It is the case for 6 regions: Maastricht (6) Leiden (11), Zaanstad (20), Hilversum (29), Ede (24) and Leeuwarden (13) that overlap respectively with the arbeidsmarktregio's of *Zuid-Limburg*[31], *Holland Rijnland*, *Zaanstreek/Waterland*, *Gooi en Vechtstreek*, the *FoodValley* and *Friesland*.

31 Arbeidsmarktregio's are referred to in italic and mapped in Appendix 1.

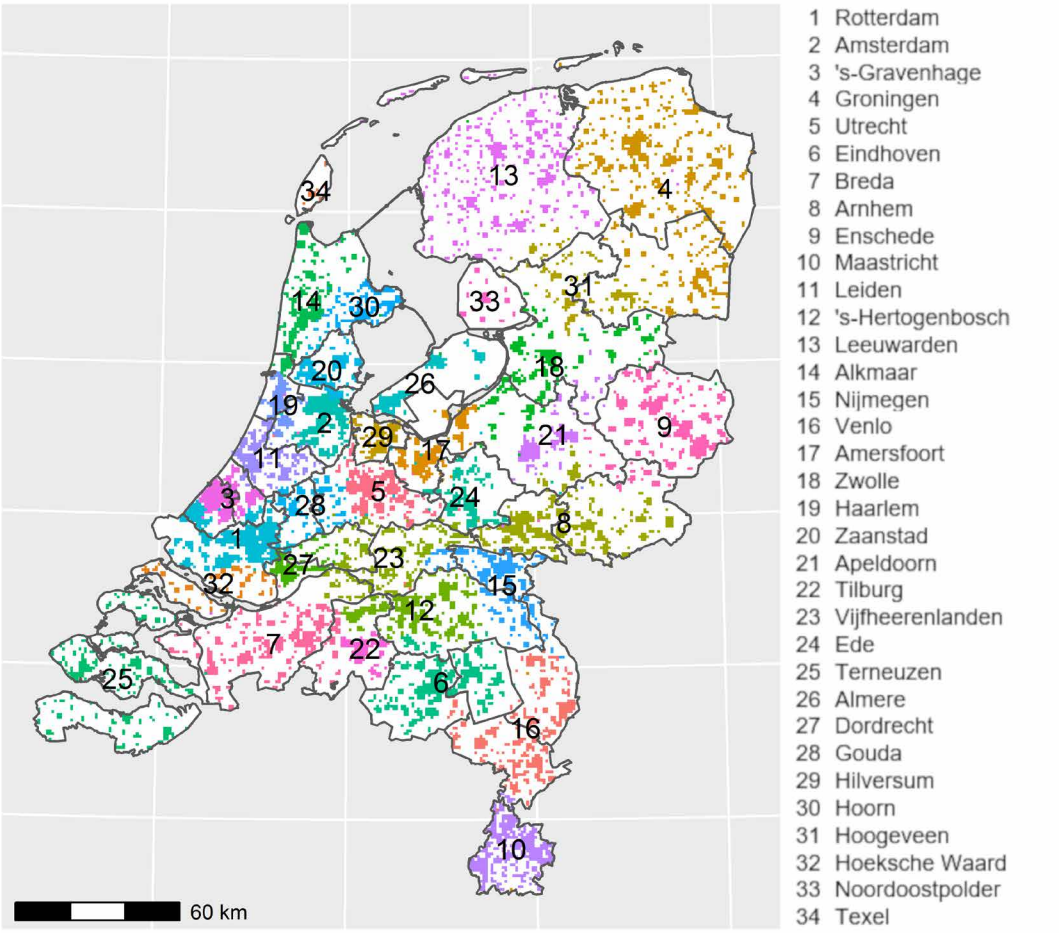

FIG. 6.5 Comparison of the 34 functional regions derived from the Intramax analysis with the arbeidsmarktregio's.
The functional regions are differentiated by colours and the arbeidsmarktregio's are surrounded by a grey line.

TOC

Some other regions present minor differences. The functional region of Terneuzen (25) does not include the two peninsulas of Tholen and Sint-Philipsland, that are attached to Breda region (7). The functional region of Gouda (28) covers a larger area than its arbeidsmarktregio counterpart, and consequently, the area of influence of Utrecht (5) appears slightly smaller. The regions of Amersfoort (17), Enschede (9) and Haarlem (19) are also larger. The municipality of Westland, that is attached to The Hague according to the arbeidsmarktregio definition is actually integrated in the urban continuum along the Meuse estuary that includes Rotterdam, Schiedam, Vlaardingen and Maassluis.

Some large arbeidsmarktregio's also split if we compare with our bottom-up empirical approach. It is the case for Zwolle region (18), that is distinct from Hoogeveen (31), Almere (26) that does not belong to the same cluster as Noordoostpolder (33). The arbeidsmarktregio called *Noord-Holland Noord* is also characterized by two functional regions, one structured around Alkmaar (14) and one around Hoorn (30).

Perception-wise distinct arbeidsmarktregio's also end up merged in the results of our procedure, for example a large cluster around Eindhoven that reunites *Zuidoost-Brabant* and *Noordoost-Brabant* and Venlo that merges *Midden-Limburg* and *Helmond-De Peel*.

The algorithm did not result in similar self-containment thresholds (Table 6.3). They vary from 0.69 for Zaanstad (20) to 0.93 for Maastricht (10). It is notable that the lower the share of self-contained flows, the more likely a region is to integrate on a higher scale with more iterations. This is especially the case for Amsterdam functional region (2) and its neighbours that manifest relatively low levels of self-containment. Amsterdam achieved indeed only 72% of self-contained residential relocation of stable workers, Haarlem 74% and Zaanstad 69%.

TABLE 6.3 Description of the 34 clusters

ID	Name	Self-contained flows	total outgoing flows	Share of self contained flows
1	Rotterdam	244,056	293,884	0.83
2	Amsterdam	212,154	293,523	0.72
3	‘s-Gravenhage	179,715	222,755	0.81
4	Groningen	150,094	165,007	0.91
5	Utrecht	144,096	191,079	0.75
6	Eindhoven	139,220	161,100	0.86
7	Breda	124,002	142,985	0.87
8	Arnhem	111,586	132,241	0.84
9	Enschede	108,580	119,731	0.91
10	Maastricht	101,588	109,712	0.93
11	Leiden	98,863	121,533	0.81
12	‘s-Hertogenbosch	97,884	119,342	0.82
13	Leeuwarden	86,883	97,226	0.89
14	Alkmaar	86,166	101,821	0.85
15	Nijmegen	76,458	93,688	0.82
16	Venlo	72,746	84,414	0.86
17	Amersfoort	70,736	89,130	0.79
18	Zwolle	70,454	84,844	0.83
19	Haarlem	67,125	91,289	0.74
20	Zaanstad	55,911	80,663	0.69
21	Apeldoorn	53,786	65,354	0.82
22	Tilburg	53,722	68,368	0.79
23	Vijfheerenlanden	53,052	69,371	0.76
24	Ede	53,051	67,318	0.79
25	Terneuzen	51,620	57,016	0.91
26	Almere	47,731	63,436	0.75
27	Dordrecht	46,790	60,574	0.77
28	Gouda	40,233	53,593	0.75
29	Hilversum	37,978	52,412	0.72
30	Hoorn	32,301	39,846	0.81
31	Hoogeveen	25,064	30,721	0.82
32	Hoeksche Waard	16,978	22,248	0.76
33	Noordoostpolder	8,440	10,448	0.81
34	Texel	1,591	1,979	0.80

The residential flows of stable workers are not balanced. Figure 6.6 shows that the flows of stable workers from Amsterdam to its neighbouring functional regions is systematically higher than the other way around. This is the case for Haarlem and Zaanstad but also Hilversum and Almere. Many hypotheses regarding the directionality could be listed. It could be that households of stable workers, due to their position in the life cycle or economic conditions, are moving to peripheries or smaller centres to access larger or cheaper houses. By doing so they would integrate the functional regions at a higher scale. Such hypotheses can be explored thanks to microdata.

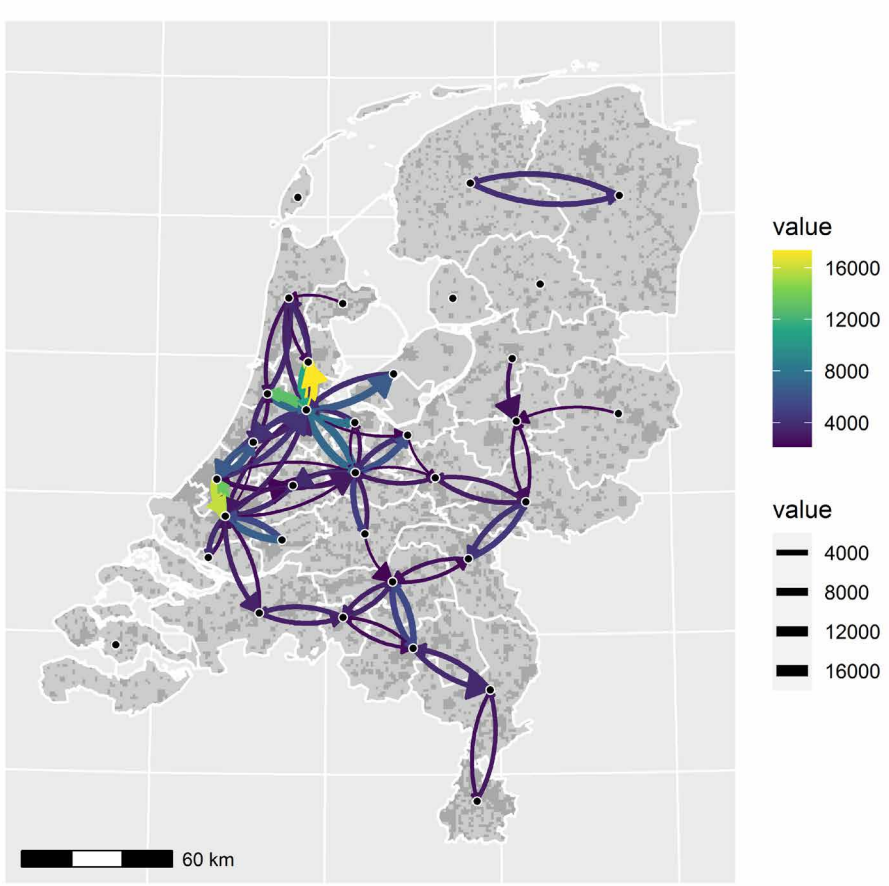

FIG. 6.6 Main inter-cluster migrations of stable workers based on results for C=34. Only the 80 most important flows in absolute values are represented.

6.4.3 Exploring the heterogeneity of daily urban system integration at different scales

In this last part of the exploration of the use of microdata to understand the functional integration of regions, we test whether two population groups have different patterns of residential relocation while keeping their job. We took the two age groups used in part 2.2. (stable workers between 25 and 34 years old, and between 40 and 54 years old) and computed the share of people moving within their functional regions and between two functional regions, the latter group participating more to integrating the system at a higher scale. We tested it for two different scales in order to see whether such results would still hold depending on the definition of functional region used. The two scales of analysis are depicted in Figure 6.7. We used the boundaries of the step resulting in 34 clusters described above, and the boundaries of the step resulting in 46 clusters that produce smaller functional regions.

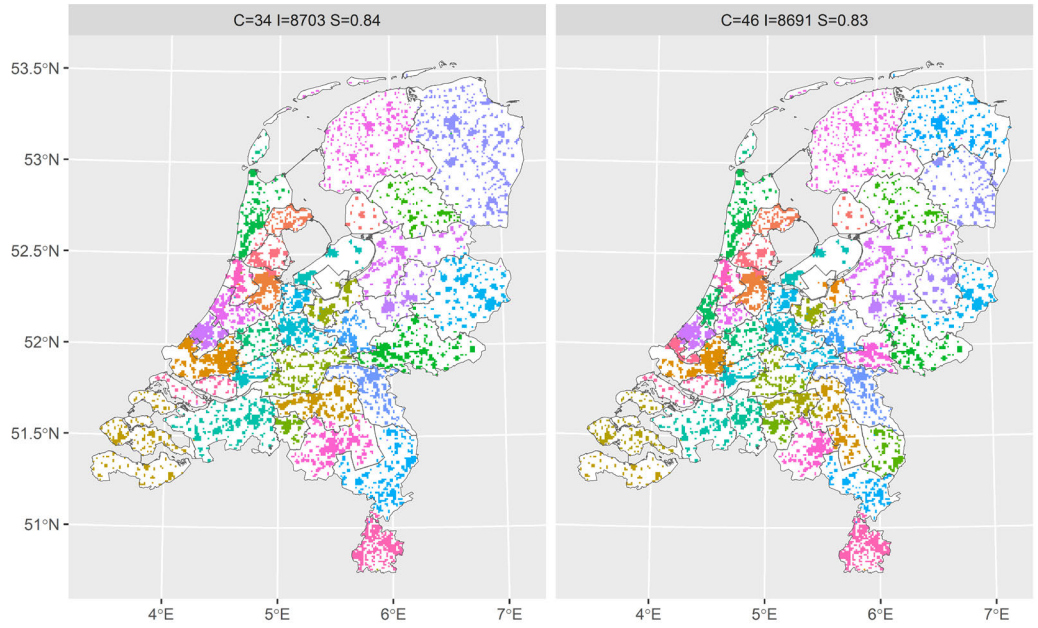

FIG. 6.7 Maps of the two scales of analysis.
Because the maps show many clusters with redundant colours, we highlighted the clusters with light grey boundaries showing the fusion of the municipalities covered by the clusters.

TOC

The result of this analysis shows that stable workers between 25 and 34 are more likely to move between two functional regions Table 6.4. Among the 1,519,525 movements we observed for this group, 19.11 % are inter-cluster movements using the regionalisation with 34 clusters and 21.2 % when using the regionalisation with 46 clusters. For the group of stable workers between 40 to 54 years old, that is smaller in absolute value (766,804 movements), 16.11 % of the movements are happening between two clusters for C=34, and 17.99 % for C=46. Hence, this proportion is consistently lower than for younger workers that manifest higher spatial flexibility. When taking only the clusters located in our extended definition of the Randstad (see map for C=3 in Figure 6.4), the proportion of movements between clusters is slightly higher for all the groups, which seems to indicate a higher spatial flexibility of workers in this area and more integration of the functional regions.

TABLE 6.4 Intra- and inter-cluster migration for two population groups

		C=34		C=46		
		share intra-cluster	share inter-cluster	share intra-cluster	share inter-cluster	N
Netherlands	all	81.55	18.45	79.48	20.52	3,458,651
	25-34	80.89	19.11	78.8	21.2	1,519,525
	40-54	83.89	16.11	82.01	17.99	766,804
Extended Randstad	all	80.55	19.45	79.18	20.82	1,847,962
	25-34	79.54	20.46	78.17	21.83	830,211
	40-54	82.63	17.37	81.37	18.63	407,747

6.5 Conclusion

From this exploration of the potential of microdata, we can first conclude that using residential relocations of stable workers provides an interesting alternative to commuting data for delimiting functional regions. Descriptive statistics of the distance of the movements indicates that they stay more local than the ones of workers changing jobs. We highlighted also differences in terms of migration behaviour of different subgroups of stable workers. Younger stable workers earning more than the median salary of their age group are the ones that cover the longest moving distance on average. Older stable workers earning less than the median income of their age group are the ones staying the closest to their initial location.

Using the Intramax procedure, we could create functional regions from the network of residential relocations of stable workers. Looking at the different steps of the procedure, we could identify meaningful functional regions, their nested hierarchy and their integration at a higher scale. The empirical analysis resulting in 34 regions provided an alternative to the existing official division of labour market areas that were created on the basis of perceptions. Because of their strong functional basis, the regions resulting from our hierarchical aggregation procedure could inform the perception-based procedure to define arbeidsmarktregio's to foster the achievement of a better labour market matching in the future.

We also investigated which age group of our population of stable workers is the most inclined to migrate between two functional regions, which indicates an integration of two regions at a higher scale. Using two scales of analysis, we showed that workers between 25 and 34 years old are more likely to move between two functional regions without changing jobs.

This last analysis of heterogeneity in relocation behaviour and labour market participation at higher spatial scales of the urban system has remained rather simple for time issues. Here, we want to suggest several future research directions in order to continue this type of analysis. First of all, the population groups that we created were rather crude. An important step to better characterize individuals would be to add information on their occupation and household characteristics. It has been demonstrated that the spatial flexibility of people depends largely on the type of jobs they are doing. Occupations do not have the same requirements for presence on the work site, and people that can telework or adapt their hours have more spatial flexibility on the housing market. In this research, we only looked at individuals because our elementary unit of analysis was the worker keeping its job or not, however, adding information on their household is necessary to understand their pattern of migration. Indeed many factors influencing migration operate at household level (size, occupational status of the partner, housing occupancy status, disposable income). Finally, while the Intramax method resulted in meaningful results in the delineation of functional regions at multiple spatial scales for the entire population of stable workers, further work will be necessary to adapt such clustering algorithm to the delineation of functional space of population subgroup.

We have demonstrated that it is possible to explore the spatial organisation of the Netherlands into functional regions using individual level register data. This data has the great advantage of being exhaustive, having fixed categories over time, and being available since the end of the 1990s and frequently updated. By providing information on individual level heterogeneity in spatial behaviour, it opens new possibilities for analysing the functional space of different population subgroups, aspects that are important to better understand the organisation of cities and urban systems.

Appendix

Source: CBS

The Dutch arbeidsmarketreio's in 2018

Bibliography

Alvanides, S., Openshaw, S., Duke-Williams, O., 2000. Designing zoning systems for flow data. GIS and Geocomputation. Innovation in GIS 7, 115–134.

Arbabi, H., Mayfield, M., Dabinett, G., 2019. Urban performance at different boundaries in England and Wales through the settlement scaling theory. Regional Studies 53, 887–899. https://doi.org/10.1080/00343404.2018.1490501

Arcaute, E., Molinero, C., Hatna, E., Murcio, R., Vargas-Ruiz, C., Masucci, A.P., Batty, M., 2016. Cities and regions in Britain through hierarchical percolation. Royal Society Open Science 3, 150691. https://doi.org/10.1098/rsos.150691

Bourne, L.S., 1981. The Geography of Housing. Wiley.

Brown, P.J.B., Hincks, S., 2008. A Framework for Housing Market Area Delineation: Principles and Application. Urban Studies 45, 2225–2247. https://doi.org/10.1177/0042098008095866

Burger, M.J., Meijers, E.J., van Oort, F.G., 2014. Multiple Perspectives on Functional Coherence: Heterogeneity and Multiplexity in the Randstad. Tijdschr Econ Soc Geogr 105, 444–464. https://doi.org/10.1111/tesg.12061

Coombes, M.G., Green, A.E., Openshaw, S., 1986. An Efficient Algorithm to Generate Official Statistical Reporting Areas: The Case of the 1984 Travel-to-Work Areas Revision in Britain. Journal of the Operational Research Society 37, 943–953. https://doi.org/10.1057/jors.1986.163

Cottineau, C., Hatna, E., Arcaute, E., Batty, M., 2017. Diverse cities or the systematic paradox of Urban Scaling Laws. Computers, Environment and Urban Systems, Spatial analysis with census data: emerging issues and innovative approaches 63, 80–94. https://doi.org/10.1016/j.compenvurbsys.2016.04.006

de Vos, D., Meijers, E., van Ham, M., 2018. Working from home and the willingness to accept a longer commute. Ann Reg Sci 61, 375–398. https://doi.org/10.1007/s00168-018-0873-6

Haggett, P., Chorley, R.J., 1969. Network Analysis in Geography. Edward Arnold.

INSEE, 2020. Méthode de constitution des aires d'attraction des villes 2020.

Landré, M., 2012. Geoprocessing Journey-to-Work Data: Delineating Commuting Regions in Dalarna, Sweden. ISPRS International Journal of Geo-Information 1, 294–314. https://doi.org/10.3390/ijgi1030294

Le Roux, G., Vallée, J., Commenges, H., 2017. Social segregation around the clock in the Paris region (France). Journal of Transport Geography 59, 134–145. https://doi.org/10.1016/j.jtrangeo.2017.02.003

Lobo, J., Alberti, M., Allen-Dumas, M., Arcaute, E., Barthelemy, M., Bojorquez Tapia, L.A., Brail, S., Bettencourt, L., Beukes, A., Chen, W.-Q., Florida, R., Gonzalez, M., Grimm, N., Hamilton, M., Kempes, C., Kontokosta, C.E., Mellander, C., Neal, Z.P., Ortman, S., Pfeiffer, D., Price, M., Revi, A., Rozenblat, C., Rybski, D., Siemiatycki, M., Shutters, S.T., Smith, M.E., Stokes, E.C., Strumsky, D., West, G., White, D., Wu, J., Yang, V.C., York, A., Youn, H., 2020. Urban Science: Integrated Theory from the First Cities to Sustainable Metropolises (SSRN Scholarly Paper No. ID 3526940). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3526940

Masser, I., Brown, P.J.B., 1975. Hierarchical Aggregation Procedures for Interaction Data. Environ Plan A 7, 509–523. https://doi.org/10.1068/a070509

Masser, I., Scheurwater, J., 1980. Functional Regionalisation of Spatial Interaction Data. An Evaluation of Some Suggested Strategies: Environment and Planning A. https://doi.org/10.1068/a121357

Mitchell, W., Watts, M., 2010. Identifying Functional Regions in Australia Using Hierarchical Aggregation Techniques. Geographical Research 48, 24–41. https://doi.org/10.1111/j.1745-5871.2009.00631.x

Moonen, L., Otten, F., Pleijers, A., 2011. Inkomens en positie op de arbeidsmarkt, Sociaaleconomische trends, 1e kwartaal 2011. Den Haag: Centraal Bureau voor de Statistiek.

Moreno-Monroy, A.I., Schiavina, M., Veneri, P., 2020. Metropolitan areas in the world. Delineation and population trends. Journal of Urban Economics 103242. https://doi.org/10.1016/j.jue.2020.103242

Netto, V.M., Meirelles, J.V., Pinheiro, M., Lorea, H., 2018. A temporal geography of encounters. Cybergeo : European Journal of Geography. https://doi.org/10.4000/cybergeo.28985

Petrović, A., van Ham, M., Manley, D., 2018. Multiscale Measures of Population: Within- and between-City Variation in Exposure to the Sociospatial Context. Annals of the American Association of Geographers 108, 1057–1074. https://doi.org/10.1080/24694452.2017.1411245

Poorthuis, A., Meeteren, M. van, 2019. Containment and Connectivity in Dutch Urban Systems: A Network-Analytical Operationalisation of the Three-Systems Model. Tijdschrift voor economische en sociale geografie n/a. https://doi.org/10.1111/tesg.12391

Pumain, D., 2011. Systems of Cities and Levels of Organisation, in: Bourgine, P., Lesne, A. (Eds.), Morphogenesis. Springer Berlin Heidelberg, pp. 225–249. https://doi.org/10.1007/978-3-642-13174-5_13

Raimbault, J., 2019. Multi-dimensional Urban Network Percolation. arXiv:1903.07141 [physics].

TOC

7 Conclusions and discussion

7.1 Introduction

This PhD research has developed alternative methods of measuring patterns of interrelationships between cities, and used them to advance our understanding of the spatial organisation and the evolution of the Dutch system of cities. The underlying knowledge gap that motivated this work was that, for many years, urban system research has focused on aspects such as the concentration of populations, economic activities, and urban functions but was limited in focus on the actual networks connecting cities and the impact they have on urban dynamics. During the first wave of urban systems research in the 1960s-1980s, only a limited number of studies were actually looking at relational data and the mention of flows and networks was mostly metaphorical. This slow development of network approaches to urban systems was mainly due to limitations in computational power and data availability. More recently, these issues were partly addressed by the development of 'big data'. The availability of big geospatial databases related to the development of locational-aware technologies has been the cause of a new surge of interest for urban systems research, sometimes gathered under the label of 'urban analytics' and 'urban sciences'. Cross-fertilisation with the emerging field of network sciences and contributions from physicists and computer scientists have led to an opening up of new frontiers in urban research; notably the possibility to work on very large population groups at very fine spatial and temporal scales. However, these approaches are not without limitations. Beyond common statistical problems related to these sources such as ecological fallacy, self-selection bias and non-representativeness, the fact that this data is generated by fast changing technologies with fluid categories makes it difficult to create datasets that are consistent over time to analyse cities from an urban systems perspective.

For this dissertation, I was inspired by a recent stream in urban systems research that builds on assembling and analysing of data sources such as archives and registers with modern computational methods to analyse patterns of interurban relationships. In this PhD research, I proposed to explore these approaches further in the context of the Netherlands. The aim was to develop new methods to measure relations between cities that are consistent trough time and over the entire Dutch urban system. In order to achieve this goal, the potential of three datasets was explored: two text archives and register data. In a first experiment to fulfil this aim, the co-mentions of city names in an Internet archive were looked at. This was an opportunity to upscale the 'co-occurrence method' and to approach text as data and to identify the challenges of dealing with such sources. The research question was*: to what extent can the toponym co-occurrence method be used to identify the patterns of relations between cities in a systematic way?* (*RQ1).* To answer the issue about time consistency, the strategy has been to use another database containing text data: the Delpher archive that includes millions of digitally searchable historical newspapers. The second research question was about dealing with the unstructured aspect of these sources: *how can we extract data on relations between cities for a long period of time from a massive archive of historical newspapers? (RQ2).* The third question was about the usefulness of such data sources. Understanding *to what extent are information circulation extracted from historical newspaper archives reflecting the long term evolution of the territorial organisation of a system of cities? (RQ3).* Finally, in the final empirical chapter of this dissertation, the scale of analysis of urban system research is questioned because intra- and interurban processes are increasingly difficult to differentiate as urbanisation intensifies, and especially in the Dutch context where numerous big cities are located very close to each other. The final research question asks: *to what extent can longitudinal individual data be used to build bottom-up definitions of cities and urban systems? (RQ4).* The next section presents the outcome of this research process.

7.2 Summary of the research results

Chapter 2 of this thesis, published in *Networks and Spatial Economics*, has presented an overview of the system of cities literature since the mid-1990s. It started with the observation that the landscape of research appears to be rather fragmented. This fragmentation is illustrated well by the frequency of the use of the expression 'paradigm change' by researchers willing to position themselves

in opposition with 'classical' approaches. This study of the evolution of the urban systems literature does not take the form of a typical literature review study. Rather it adopts a bibliometric approach to analyse a set of 1,491 papers on intercity relationships from 1995 onward. The methodology I used is based on the hyper-network approach, which combines the analysis of semantic and citation networks to give a 'bird's eye view' of the scientific field. First, a semantic network based on co-occurrences of words in the titles and abstracts of the set of papers is extracted. This allowed for an identification of the different subfields or schools of thought on the systems of cities. Second, the pattern of citations of the papers developed within these schools were analysed to understand the connections between the different subfields. This analysis allowed for a distinction between five different lexical fields in urban systems research. Strong differences in terms of methods, scale of analysis, thematic focus, main agents identified and influences from other disciplines have become manifest, which warrants a discussion about specific schools of thought.

The first school of thought that was identified is the research on world cities (WCN). This stream of research explores the top of the urban hierarchy by looking at global networks (mainly corporate networks and some transportation networks). The geographical scope of most research in this cluster is either at a global scale or on the local scale of the cities, making up this world city network. From a methodological point of view, this field is characterised by benchmarking studies as well as vocabulary from network analysis. The second group of publications that appeared clearly from this analysis is the cluster about regional urban systems (REG). This cluster is centred around the analysis of polycentric urban regions, with much attention paid to their performance, competitiveness, and sustainability. This field of research has connections with planning. A third cluster organised around the notions of complex systems and simulation also appeared (SIM). This cluster is characterised by a profusion of terms related to modelling and simulation. On the one hand, it contains publications grounded in theoretical geography with mentions of the Central Place model, spatial interaction modelling and network analysis. On the other hand, notions and methods from complexity theory are also very visible with references to simulation methods, predictions and complex systems. A fourth cluster gathers the vocabulary of studies dealing with aspects associated with city size (CSD). It contains all the publications dealing with the descriptions and explanations behind the regularity in city size distributions that have fascinated researchers for more than a century, a debate that is still vivid today and re-energized by the question of scaling laws. Finally, a last cluster characterized by the vocabulary of economic geography also appeared (ECON). From an empirical point of view, the publications refer mostly to a North American context. In terms of citation network structure, this group was less coherent than the others, probably because the other schools of thought also use vocabulary from economic geography.

In terms of citation patterns, three out of five schools of thought (REG, WCN and CSD) appeared as very coherent internally. It has to be noted that the very last period is characterised by a small average decrease of this insularity index, suggesting more integration between most schools of thought. But this is still insufficient evidence for a move from multidisciplinarity to interdisciplinarity in urban system research.

Chapter 3, published in the *International Journal of Urban Sciences*, investigates the feasibility of using the toponym co-occurrence method on web pages containing city names to identify patterns of relations between cities in a systematic way (RQ1). This approach builds urban systems on the basis of co-mentions of place names in documents of a text corpus. In this research, billions of web pages contained in the Common Crawl web archive were analysed. One of the innovations of this research is the inclusion of small settlements, that tend to be ignored in many studies. In addition, an attempt to classify these relations based on the content of the text document has been done. Because the city names were retrieved with simple string queries, attention is also paid to evaluating the impact of ambiguous place names in the resulting frequency of relations. In order to benchmark the method, gravity modelling is employed to assess the resulting spatial organisation of the Netherlands. It turns out that the gravity model fits the pattern of relationships between places as found in the corpus of web pages, which contributes to the assessment that the toponym co-occurrence method is an interesting proxy for relationships in real physical space. Using this method, it is established that the relationships in the Randstad region, considered by many as a coherent metropolitan entity, are actually somewhat less strong than expected. In contrast, historically important but currently small cities in the periphery, tend to have maintained their prominent position in the pattern of relationships. Relatively new suburban places in the shadow of a larger city tend to be weakly related to other places. The toponym co-occurrence method is widely applicable to many types of text archives. The accuracy of the results of the method, however, is also much determined by the quality of the underlying data. While using a gigantic web archive is probably better than using the highly varied results of a search engine like in previous studies, it is clear that there is little information to assess the representativeness of the data collected. The usability of the toponym co-occurrence method therefore hinges on a good quality dataset to which it is applied. Ideally, the dataset would allow for a better characterisation of the meaning of a co-occurrence of place names as they appear within a document, a paragraph, or a sentence. Rather than closing the discussion, this chapter leads to new challenges for improving the retrieval of geographical information from unstructured text data. Challenges that were approached in the two next chapters.

Chapter 4, published as a data paper in *Cybergeo, European Journal of Geography*, answers the second research question about the extraction of meaningful data on intercity relations from an archive of digitised historical newspapers (RQ2). It departs from the fact that Digital Humanities scholars have shown that such archives could be used to identify macroscopic trends related to historical and cultural changes, but that the wealth of geographical information from these sources has not been analysed very much. The conclusion was that the answer to the research question relies on four crucial steps that are necessary to build a dataset on intercity relations from this corpus of semi-structured text: the careful selection of a sub-corpus and of relevant geographical objects to work on, the creation of a meaningful ontology to go from a collection of news items to an origin-destination matrix, the identification of problems in the automatic identification of the city names and finally, the design of an algorithm in a trade-off between accuracy and computing time. These steps are described below.

The first important step in quantitative studies using a text archive is to select a relevant corpus and meaningful entities to extract. The content of a digital archive might be influenced by many factors such as digitalisation policies or copyrights issues. Carefully selecting the corpus can significantly reduce bias, and is necessary to create a dataset as representative as possible depending on the research question. Four criteria have been applied in the selection of newspapers: an initial and final time mark, the fact that it was available publicly, the presence of a publication place within the Netherlands in the metadata, and publications during at least two consecutive decades to be able to follow changes over time and to indicate that it had reached a readership. Concerning the geographical objects to extract, as identifying the terms that relate to cities in the common language was necessary, a 'placial' perspective was adopted. The terms people use to say where they live were used here. The "*woonplaatsnamen*", that can be translated as "names of places of residence", appeared as the most interesting concept. For the creation of a meaningful ontology, the research drew upon international news flow theory. The origin of the information flow corresponds to the place mentioned in the item and the destination is the place where the newspaper is published. This ontology is relevant when the area of readership is spatially bounded. For this reason, the empirical analysis of this data in Chapter 4 only deals with local newspapers. The next crucial step for building the dataset was to identify places that could be ambiguous in distant reading. This part benefited from the evaluation of under- and overestimations when doing simple counts based on word frequencies done in the previous chapter. The most important sources of errors leading to false positives were listed and methods from Natural Language Processing (NLP) to deal with them were presented. Most of the issues could be avoided by using Named-Entity-Recognition (NER) algorithms that aim to locate and classify entities from a

given text into pre-defined categories, including place names. Finally, because this research was dealing with a huge number of news items, and because NER takes time to perform, an algorithm was designed by the author to apply NER only on the 8% of ambiguous place names. The rest of the time, simple string queries were used. Thanks to this targeting approach, the computing time could be reduced significantly (from months to days) while maintaining highly accurate results.

The result of this research is DIGGER, a dataset that allows the study of the spatial diffusion of information on and between the Dutch cities from a corpus of 81 newspapers published in 29 different cities between 1869 and 1994, and that has been published as Open data.

Chapter 5, published in *Computers, Environment and Urban Systems*, aimed to answer the third research question about the usefulness of such data to study the evolution of the spatial organisation of a territory (RQ3). The overarching goal of the chapter was to test the potential of this novel data source for reconstructing the evolving urban geography of an entire country. In order to achieve this goal, the research looked for regularities in how flows of information develop over time and different hypotheses related to the evolution of systems of cities were tested. These hypothesis were previously identified by the literature but not always empirically tested.

The findings confirm previous research that shows that space matters greatly in the process of diffusion of information. For the period between 1869 and 1930, the size of cities emitting information, and the distance between cities are important and significant factors explaining the circulation of information. Moreover, the hampering dimension of provincial borders and the North-South divide that structure the Netherlands for a long time remained important over the entire period. In the longitudinal analysis of this data, interesting phenomena related to the evolution of the Dutch system of cities were highlighted. First, the use of a cross-temporal gravity model with a varying distant exponent revealed a decreasing spatial friction over the entire period. However, when taking into account that the parameter associated with city size could also vary, an increasing spatial friction happening at the same time as a rising impact of population size over the period could be observed. This initial contradictory result could later on be explained by descriptive analysis on the average distance over which information travels as well as the spatial distribution of the origins of information flows.

Indeed, while almost all newspapers started to present news from more distant cities, this process was not homogeneous through space. In many cases, increasing attention was paid to places in the immediate proximity of where a newspaper

was published, but at the same time, attention shifted to the 4 largest cities of the Randstad, most of the time at the expense of closer-by medium-sized cities. The main driving factors behind this increase of long-distance interactions is indeed related to a polarisation around the main economic, political, and demographic core of the Netherlands, which from 1938 onwards, would be referred to as the "Randstad". This process can be considered to be one of hierarchical selection within the urban system in the context of space-time contraction.

Based on these results, it is concluded that it is feasible to use a computational social science approach to construct completely novel geographically relevant data sets on interurban relations, which allow the reconstruction of the evolution of the spatial organisation of a territory over time (RQ3).

Chapter 6 answers the last research question about the potential of using microdata to build bottom-up definitions of cities and urban systems (RQ4). This research departed from the fact that in the context of highly urbanised, polycentric areas such as the Netherlands nowadays, intra- and interurban processes are difficult to differentiate, and that the use of single arbitrary definitions of spatial entities such as municipalities or 'woonplaatsen' is problematic. In order to build bottom-up empirical definitions of functional areas, a method that is applicable in the Dutch context where there is no exhaustive commuting data is presented. Data on the job and residential careers of the full Dutch population was extracted and used it to identify people moving without changing job ('stable workers'). Because these people continue working at the same place but commute from two different locations, it can be deduced that these two locations belong to the same labour and housing market area. One of the main advantages of this approach is that it is based on individual data, and allows for an exploration of movement patterns of different population groups. Descriptive statistics of the distance of the movements indicates that they stay more local than the ones of workers changing jobs. Differences in terms of migration behaviour of different subgroups of 'stable workers' are also highlighted. Younger stable workers earning more than the median salary of their age group are the ones that cover the longest distance on average. Older stable workers earning less than the median income of their age group being the ones staying close to their initial residential location. Using the Intramax procedure, functional regions from the network of residential relocations of stable workers could be created. The different steps of this procedure revealed meaningful functional regions, their nested hierarchy, and their integration at a higher scale. Out of the many iterations, the outcome of the analysis that resulted in 34 regions provided an alternative to the existing division of labour market areas that is currently based on perceptions of local stakeholders. While there were some overlaps when comparing both, there were also many inconsistencies. Using the areas resulting from this clustering procedure,

which have strong functional basis, could lead to better labour market matching. It was also investigated which age group of stable workers is the most inclined to migrate between functional regions, flows that indicate the integration of regions at a higher spatial scale. This brief analysis demonstrates that workers between 25 and 34 years old are the ones that are the most likely to move between two regions while maintaining their job.

From this exploration on the potential of microdata in urban system research, it is concluded that using microdata on job and residential careers of individuals provides a good alternative to commuting data for delimiting functional regions and patterns of interconnections between them (RQ4).

7.3 Scientific contributions, limitations and directions for further research

This PhD thesis began with the aim of developing alternative methods for measuring patterns of interrelationships between cities while evaluating their potential to advance our understanding of the organisation and evolution of the Dutch system of cities. The exploratory dimension of this aim, and the strategy of exploiting three different databases have led me to a situation where several doors have been opened without necessarily closing them. However, these explorations were characterised by success and the identification of limitations that further advance the debate on the data issue in urban system research. This section reviews the main contributions of this PhD thesis to the field of urban systems research. I will also discuss some shortcomings of the research and possible directions for further study.

The first year of this PhD research started with a review of the literature using quantitative methods (Chapter 2). Beyond the fact that this study was the first bibliometric analysis of the system of cities literature[32], it has some wider methodological relevance. First, the procedure that was designed to delineate the scientific landscape has proven to be very efficient. It was based on a series

32 Although there are other bibliometric studies covering a subset of the field such as the one from van Meeteren et al. (2015), or being more general such as the one from Pumain and Raimbault (2020)

of queries on the bibliometric database, improved with qualitative checks at each iteration of the process. Moreover, attention was given to identify the 'classics' of the field in order to integrate publications citing them. The result was a set of publications highly relevant for the subject of the dissertation. I believe that this robust way of delineating a scientific field can help researchers beyond geographic research willing to conduct a bibliometric analysis. Second, a methodology was developed to measure the interactions between the subfields of urban system research allowing to visualise and measure where collaborations are happening and where they need to be further developed. Of course, this work did not replace the extensive reading of actual papers, but it provided a good framework to structure the reading process, and find areas of the field one would not think of if adopting a snow-ball strategy. The data used for this study covered a period between the mid-1990s to 2017. At the end of the period, signs of opening were identified in the different subfields with a slight increase of collaboration between them. It would be very interesting to re-run the analysis four years later. Since this research, there have been efforts to bring together researchers from the different traditions of urban systems research during conferences and workshops (such as the 2017 Symposium in Ghent or the series of sessions 'Cities and Networks' at the AAG 2018 and 2019), special issues in journals (e.g. Derudder and Neal, 2018), and publications of a research agenda by influential researchers of the field (Lobo et al., 2020). It would be interesting to see whether these initiatives have already had an impact on the global structure of the scientific field of urban system research.

Establishing an overview of the field and its classics has been determinant in the research design of the first analysis of the Dutch urban system of this PhD research, featured in Chapter 3. The contribution to the field with this project has been to upscale a method developed by Tobler and Wineburg (1971) and apply it to a contemporary setting. The ambition of this research was to develop a synthetic indicator about relatedness between places based on the assumption that two cities mentioned together on the same web page are more likely to be related. This was motivated by the fact that many web pages that were reviewed with colleagues confirmed this assumption. These could, for instance, be pages of companies with different sites, web pages of transport authorities, or road planners indicating a train or bus line, news articles about events happening in different places, etc. Taken one by one, most of these co-occurrences could be interpreted as an actual relation. This method has the advantage of being scale free, and allows the investigation of interactions between settlements of any size and potentially across national borders. When looking at the overall structure of the network, a good fit with the gravity model seems to indicate that the method gives a reasonably accurate general picture of the system. The relative weak position of new towns in the vicinity of big ones, and the relative success of medium size historical cities in the network

echoes recent research on the European settlement system. However, the approach developed in this chapter also manifests some limitations. While the relations are interpretable when taken one by one, once aggregated, the actual meaning of the relation becomes blurred and hard to identify as it is very diverse. The approach provides a good aggregate view of the system, but it is difficult to disentangle the actual geographical mechanisms explaining the structure of the network. One of the reasons is that little is known about the representativeness of the underlying data that was used. However, it is important to note that it does not have to do with the method itself. When Tobler and Wineburg (1971) looked at co-occurrences of towns in merchant cuneiform tablets, they knew that having two of them on a single tablet would imply an actual trading relation. I am convinced that the method could have very relevant implementations on a corpus carefully selected with a better characterisation of the relation implied by a co-occurrence. More detail could also be obtained when not just looking for co-occurrences at the page or document level, but on a paragraph or a sentence level. Despite these limitations, this initial experiment has also allowed for an identification of important challenges related to using text as data for geographical research, and notably the issue of corpus selection and place name recognition that were explored in Chapter 4.

Chapters 4 and 5 benefited a lot from the initial experiment with the web archive. During the data collection phase, two important challenges raised by the previous research were solved: selecting a representative corpus and solving most of the issues with place name recognition. For this work, the research benefited from exchanges with historians and collections specialist from the Koninklijke Bibliotheek, where I was welcomed as a guest researcher for half a year, as well as from computer scientists specialised in enriching the digital collections of the library. Thanks to that, work could be done on a corpus that was still massive but representative of the informational landscape of the Netherlands for several decades. State of the art methods of information retrieval could then be applied to create geographical information from text data. The ontology chosen for defining the relations between cities was also slightly different, and maybe easier to interpret as a geographer. Because this research worked with local (not national) newspapers the readership of each newspaper could be located, allowing for the creation of an origin-destination matrix of information flows. Thanks to the robustness of this data, a theory-driven approach could be developed, where the hypotheses related to the evolution of a system of cities with novel data covering the entire territory of the Netherlands for six decades could be tested. This work gives a good hint of the possibilities offered by computational social sciences. It departed from the assemblage of a massive number of historical sources that was analysed with NLP methods, and a modelling approach rooted in quantitative geography was applied in order to reveal hidden patterns in the corpus related to the evolution of the territorial organisation of the

country. It was of course expected that a decreasing importance of distance over the period studied would be found, however, the fact that a process of hierarchical selection appeared with such clarity in the data was as a real surprise. This method has proven to be very efficient in describing the past dynamics of an urban system, and it would be interesting to see it applied on the digital archives of newspapers from other countries. This research is exclusively focused on the 'domestic' news flows, but looking deeper into the information field of rising international cities like Amsterdam and Rotterdam during this study period could be very informative on the genesis of the world city network. Of course, such an approach could be improved in order to better characterise the evolution of the relations between cities. An interesting direction would be to filter the different items in the database. During the qualitative checks of the underlying data, a lot of companies advertising their products or shops in another part of the country were identified. With a robust classification algorithm, such data could provide a very interesting take on firm networks or the geographical diffusion of innovations for example.

Finally, another contribution has been to work on the scale and boundaries of cities in the Dutch context. Over the last few years, the questions of the scale and boundaries of cities have become major issues in urban studies because of the identification of shortcomings in studies using singular and arbitrary choices of spatial scales. This scale question was considered since performing the analysis for the study in Chapter 2, as one of the reasons of the fragmentation of research on systems of cities that were identified was that the different schools of thought do not define cities and systems of cities in the same way. In the context of the Netherlands, this issue is especially important because of the high degree of urbanisation and the existence of big population centres located close to each other. The main contribution of Chapter 6 is to adopt a bottom-up approach that departs from basic interactions at the level of individuals. Thanks to that, a delineation could be made between functional regions without departing from predefined population centres, which is more suitable for a territory characterised by polycentricity. While most studies that discern functional regions use aggregated data, one of the main innovations of the approach developed in this chapter is that it is based on micro-level data, which allows for an exploration of how functional regions differ for different types of people, hence taking into account individual level heterogeneity. Considering this heterogeneity more seriously is a major direction for future research because the spatial organisation of cities and their functional linkages differ greatly depending on population subgroups. Moreover, the source used for this study (CBS microdata) contains data from the end of the 1990s and is periodically updated. It could then be used to map the evolving geography of functional regions.

7.4 Concluding remarks

This work has contributed to the line of research that seeks to approach cities in a relational way. Thanks to an historical approach this research demonstrates that while there are global variations in a system of cities (in this case an expansion of the information field of Dutch cities), this process does not affect all the places in the same way. Some cities benefit from increasing interactions while others do not. The hierarchical selection resulting in the rise of the Randstad happened at the expense of medium sized cities and provincial capitals. They were sort of skipped over, a mechanism comparable to the tunnel effect generated by fast transportation between two places. Such an historical hindsight allow for a reflection on the self-organisation of a system of cities, where the most central nodes manifest processes of cumulative growth. This goes beyond growth that is targeted by planning interventions or place based policies. In another experiment on residential moves of stable workers, the nested structure of functional regions in the Netherlands and the strong connections between them was also shown. This is especially true in the case of the Randstad where many residential moves of stable workers are happening between urban regions. These elements show that cities cannot and should not be seen as closed, bounded, coherent entities, but rather as open, complex and interconnected to ranges of other spaces and places. These investigations were limited to interactions at the regional and national scale, but they also exist beyond national borders at the continental and global scales.

With the deluge of spatial and relational data that we are experiencing, it is very likely that urban systems research will further expand in the coming years. However, its current data-driven dimension results in many studies describing very precisely what is happening in some layers of the urban systems but with little engagement with an overarching theory. I hope that this work has contributed to giving some insights on possible ways to observe overarching dynamics and processes in systems of cities. Connecting the different types of relations, and looking at the multitude of layers that characterised urban systems is a major direction for developing an integrated urban system theory.

Bibliography

Derudder, B., Neal, Z., 2018. Uncovering Links Between Urban Studies and Network Science. Netw Spat Econ 18, 441–446. https://doi.org/10.1007/s11067-019-09453-w

Lobo, J., Alberti, M., Allen-Dumas, M., Arcaute, E., Barthelemy, M., Bojorquez Tapia, L.A., Brail, S., Bettencourt, L., Beukes, A., Chen, W.-Q., Florida, R., Gonzalez, M., Grimm, N., Hamilton, M., Kempes, C., Kontokosta, C.E., Mellander, C., Neal, Z.P., Ortman, S., Pfeiffer, D., Price, M., Revi, A., Rozenblat, C., Rybski, D., Siemiatycki, M., Shutters, S.T., Smith, M.E., Stokes, E.C., Strumsky, D., West, G., White, D., Wu, J., Yang, V.C., York, A., Youn, H., 2020. Urban Science: Integrated Theory from the First Cities to Sustainable Metropolises (SSRN Scholarly Paper No. ID 3526940). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3526940

Pumain, D., Raimbault, J., 2020. Conclusion: Perspectives on Urban Theories, in: Pumain, D. (Ed.), Theories and Models of Urbanization: Geography, Economics and Computing Sciences, Lecture Notes in Morphogenesis. Springer International Publishing, Cham, pp. 303–330. https://doi.org/10.1007/978-3-030-36656-8_16

Tobler, W., Wineburg, S., 1971. A Cappadocian Speculation. Nature 231, 39–41. https://doi.org/10.1038/231039a0

van Meeteren, M., Poorthuis, A., Derudder, B., Witlox, F., 2015. Pacifying Babel's Tower: A scientometric analysis of polycentricity in urban research. Urban Stud 0042098015573455. https://doi.org/10.1177/0042098015573455

TOC

About the author

Antoine Peris was born in 1992 in Prades, France. After finishing high-school in Agen, he went to Toulouse for completing *Hypokhâgne* and *Khâgne* at Lycée Saint Sernin with a major in History and Geography. He then went to study Geography at Université Paris 1 Panthéon Sorbonne where he received his Bachelor and Master.

In August 2016 he started a PhD at Delft University of Technology under the supervision of Evert Meijers and Maarten van Ham. This PhD was part of the NWO VIDI project "Beyond Agglomerations: Mapping Externality Fields and Network Externalities" led by Evert Meijers. During the PhD Programme, Antoine presented his work at various international and European conferences and gave lectures and tutorials at TU Delft, Université de Paris and Leiden University. In 2018 he was researcher-in-residence at the Koninklijke Bibliotheek, the National Library of the Netherlands. In 2021, Antoine started working as post-doctoral researcher at UMR Espace in Avignon Université.

TOC

Publications

Peer-reviewed Journal Articles

Peris, A., Meijers, E., van Ham, M., 2018. The Evolution of the Systems of Cities Literature Since 1995: Schools of Thought and their Interaction. Networks Spatial Economics 18, 533–554. https://doi.org/10.1007/s11067-018-9410-5

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

Peris, A., Meijers, E., van Ham, M., 2021. Information diffusion between Dutch cities: Revisiting Zipf and Pred using a computational social science approach. Computers, Environment and Urban Systems 85, 101565. https://doi.org/10.1016/j.compenvurbsys.2020.101565

Peris, A., Faber, W.J., Meijers, E., van Ham, M., 2020. One century of information diffusion in the Netherlands derived from a massive digital archive of historical newspapers: the DIGGER dataset. Cybergeo : European Journal of Geography.

Other publications

Peris, A., Faber, W.J., 2019. DIGGER: a dataset built on Delpher, the digital archive of historical newspapers of the National Library of the Netherlands. https://doi.org/10.4121/uuid:a14a1607-dafe-4a8a-aebc-d1c5cd66a588

Meijers, E., Peris, A., 2020. De gerelateerdheid van plaatsen in Regio Oost: een verkenning middels de toponym co-occurrence methode, Rapportage Kracht van Oost.

Submitted Papers and Work in Progress

van den Berghe, K., Peris, A., Meijers, E., Jacobs W., (submitted to a peer-review journal). Friends with Benefits: The Emergence of the Amsterdam-Rotterdam-Antwerp (ARA) Functional Polycentric Port Region

Peris A. (work in progress) Mapping functional regions in the Netherlands by analysing individual residential and job histories

21#07

Cities in interaction

Analysing the Dutch system of cities with computational methods

Antoine Peris

Cities never function in isolation but as nodes in overarching systems characterised by flows of goods, people, and information. To fully understand the evolution of cities, a relational approach is needed, which investigates cities in relation to other cities and urban regions. While a significant part of urban system research has focused on aspects such as the concentration of populations and economic activities, the understanding of the actual networks connecting cities and their impact is still limited. However, the required data is notoriously difficult to obtain. This dissertation contributes to knowledge on the relationship between cities in the Netherlands by exploiting – in novel ways – three data sources: web pages mentioning cities, local historical newspapers, and administrative registers.

After providing an overview of the systems of cities literature, the toponym co-occurrences method is explored. This method aims at identifying patterns of relations between cities in a systematic way by looking at the co-mentions of cities in text documents (here in web pages). Using text as data appeared as a great direction for studying urban systems, and elements from this first exploration are used in the next section of the thesis where the past dynamics of the Dutch urban system is reconstructed using information flows retrieved from digitised historical newspapers. Finally in a last empirical part, the potential of information from individual-level registers about professional and residential trajectories for measuring relations between places at multiple spatial scales is investigated. This measure is then used to reveal the nested hierarchy of functional regions in the Netherlands.

A+BE | Architecture and the Built Environment | TU Delft BK