Revisiting Urban Dynamics through Social Urban Data

Methods and tools for data integration, visualization, and exploratory analysis to understand the spatiotemporal dynamics of human activity in cities

Psyllidis, Achilleas

DOI
10.7480/abe.2016.18

Publication date
2016

Document Version
Final published version

Citation (APA)
Psyllidis, A. (2016). *Revisiting Urban Dynamics through Social Urban Data: Methods and tools for data integration, visualization, and exploratory analysis to understand the spatiotemporal dynamics of human activity in cities*. [Dissertation (TU Delft), Delft University of Technology]. A+BE | Architecture and the Built Environment. https://doi.org/10.7480/abe.2016.18

Architecture and the Built environment

18
2016

Revisiting Urban Dynamics through Social Urban Data

Methods and tools for data integration, visualization, and exploratory analysis to understand the spatiotemporal dynamics of human activity in cities

Achilleas Psyllidis

Revisiting Urban Dynamics through Social Urban Data

Methods and tools for data integration, visualization, and exploratory analysis to understand the spatiotemporal dynamics of human activity in cities

Achilleas Psyllidis
Delft University of Technology, Faculty of Architecture and the Built Environment, Department of Architectural Engineering and Technology

Architecture
and the
Built environment

abe.tudelft.nl

Design: Sirene Ontwerpers, Rotterdam

ISBN 978-94-92516-20-6
ISSN 2212-3202

Revisiting Urban Dynamics through Social Urban Data

Methods and tools for data integration, visualization, and exploratory analysis to understand the spatiotemporal dynamics of human activity in cities

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben;
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 17 november 2016 om 12:30 uur
door

Achillefs PSYLLIDIS

Diplom-Ingenieur in Architectural Engineering,
Master of Philosophy in Architecture – Spatial Planning,
National Technical University of Athens, Griekenland
geboren te Athene, Griekenland

This dissertation has been approved by the

promotor: Prof. ir. K. Oosterhuis
copromotor: Dr. ir. N. M. Biloria

Composition of the doctoral committee:

Rector Magnificus	chairman
Prof. ir. K. Oosterhuis	Delft University of Technology
Dr. ir. N. M. Biloria	Delft University of Technology

Independent members:

Prof. dr. ir. A. van Timmeren	Delft University of Technology
Prof. dr. ir. I. S. Sariyildiz	Delft University of Technology
Prof. dr. ir. A. K. Bregt	Wageningen University
Prof. dr. E. A. van Zoonen	Erasmus University Rotterdam
Dr. C. Andris	Pennsylvania State University

This research was funded by a scholarship from the **A. S. Onassis Foundation** from 2012 to 2016 (F ZI 085-1), as well as by a scholarship from the **Greek State Scholarships Foundation (IKY)**, co-financed by the resources of the **European Social Fund** (ESF – Educational Program "Education & Lifelong Learning") and the **National Strategic Reference Framework (NSRF)** 2007 – 2013 of the European Union, from September 2012 to March 2015. The research was further financially supported with individual research grants from the **Foundation for Education and European Culture (ΙΠΕΠ)** from 2012 to 2016, and the **A. G. Leventis Foundation** from 2014 to 2016.

To my mother, Kaiti, my grandmother, Eleni, and my sister, Despoina,
who were, are, and will always be there for me

And in memory of my father, Konstantinos

Acknowledgments

After four years of intense work at Delft University of Technology, it is time for me to reflect on the people who have supported me while pursuing this PhD research. Although this dissertation would not have been completed without my perseverance and willingness to continuously explore and test new ideas, often outside my comfort zone, I cannot but acknowledge the invaluable contribution in many ways of a number of people while conducting this research.

First and foremost, I am deeply grateful to my scholarship providers and financial supporters, namely, the A. S. Onassis Foundation, the Greek State Scholarships Foundation (IKY), the European Social Fund, the Foundation for Education and European Culture (IPEP), and the A. G. Leventis Foundation for giving me the opportunity to discover new worlds of knowledge and for helping me pursue my dreams. "Without *you*, *I* would not have set out", paraphrasing a verse of C. P. Cavafy's *Ithaca* (1911).

I would like to thank my supervisors, Prof. ir. Kas Oosterhuis and Dr. ir. Nimish Biloria, for the excellent collaboration and for giving me the freedom to explore new topics and ideas. Prof. K. Oosterhuis, thank you for giving me the exciting opportunity to be part of the Hyperbody research group. You have been a source of inspiration for me through your books and work, since my undergraduate studies in architecture. Dr. N. Biloria, I am deeply thankful for your continuous support since day one of my PhD research, for the numerous discussions and exchange of ideas, for your excellent attitude, and for believing in me and my endeavors.

I would like to express my heartfelt gratitude to Prof. dr. ir. Geert-Jan Houben and to Dr. Alessandro Bozzon for introducing me to the exciting world of data science, for their mentorship and excellent collaboration, and for showing remarkable open-mindedness. I am deeply grateful to Prof. dr. G. J. Houben for giving me the opportunity to collaborate with the Web Information Systems group, as well as for his willingness to help me since the first day I contacted him, and for his trust and support over the past couple of years. Thank you also for the opportunity to continue this collaboration; this time as a member of your research group. I would also like to specifically thank Dr. Alessandro Bozzon for the fruitful and pleasant discussions, for the critical comments and always helpful suggestions, and for the steadfast enthusiasm that kept me motivated. Every time I left your office or had a discussion with you, I was richer in knowledge.

Also, many thanks to the SocialGlass team members, namely, Dr. Stefano Bocconi, Christiaan Titos Bolivar, and Jie Yang for the wonderful collaboration, the exchange of knowledge, the positive attitude and for the great moments we have spent together.

My sincere gratitude to the independent members of my doctoral committee, Prof. dr. ir. Arjan van Timmeren, Prof. dr. ir. Sevil Sariyildiz, Prof. dr. ir. Arnold Bregt, Prof. dr. ir. Liesbet van Zoonen, and Dr. Clio Andris for their willingness and availability to review this dissertation, as well as for honoring me by serving on my doctoral committee.

I would also like to express my gratitude to the members of the Ontology Engineering group at the Universidad Politécnica de Madrid, namely, Dr. Raúl García Castro, Prof. Asunción Gómez Pérez, María Poveda Villalón, Idafen Santana Pérez, and Filip Radulović, as well as Pieter Pauwels from Ghent University and Dr. Dimitrios Tzovaras from CERTH/ITI, for the warm hospitality in Cercedilla, Madrid and for your fruitful guidance in the fields of ontology engineering and linked data. Moreover, I would especially like to thank my teammates and friends, Matthew Horrigan, Oudom Kem, and Diarmuid Ryan for the excellent companionship and collaboration, as well as for the great time we had together in Spain. It has been really nice working with you.

My sincere gratitude goes to my Dutch language teachers, Liesbeth Nos and Helen van Boekhove, from the Foreign Languages Institute of the University of Athens and the Netherlands Institute at Athens respectively, not only for teaching me the Dutch language but also for introducing me to the Dutch culture, thereby making my time here in the Netherlands much easier.

I would also like to thank my students for their passionate involvement in the courses and workshops, as well as for the fruitful discussions we had together that helped me strengthen my ideas. Special thanks to my former student and Hyperbody's student assistant Marco Galli for always feeling "fantastic", thereby contributing to a cheerful atmosphere.

Many thanks to the fellow PhD candidates, Alejandro, Ali, Flavia, Gary, Han, Jiaxiu, Luz-María, Nurul, Pirouz, and Sina, as well as to our great Hyperbody programmer Vera László for the numerous discussions over lunches, dinners, coffees, and beers that helped me escape the stressful PhD routine.

I would especially like to thank my fellow PhD candidate Eleni Papadonikolaki for her friendship over the past four years here at TU Delft, as well as for the countless spirited research debates, discussions, peer reviews, and exchange of inspiring ideas that largely contributed to this research. Her companionship, useful feedback, and support have been invaluable. I have treasured every moment, comment, thought, idea, and discussion over these past four years.

I would like to give special thanks to my friend and former colleague Bas Kalmeyer for sharing great moments in and out of work, as well as for the advice and support at times needed. Thank you for making me feel at home here in the Netherlands.

Moreover, I would especially like to thank my best friend Michalis Bourgazas for his strong friendship since we first met in the army back in 2007, on his beautiful island Samos, and for all the great moments we then had together in Athens and, thereafter, here in Delft whenever he came to visit me. Thank you for reminding me that there is much more to life than work, PhD-related or otherwise.

This last paragraph I have reserved for my family – my mother, Kaiti, my grandmother, Eleni, and my sister, Despoina – whom I cannot thank enough for their unconditional love and support throughout my entire life. Words cannot express how grateful I am to you for being by my side at every stage of my life and career, either professional or academic. This dissertation is dedicated to you. Also, I would like to dedicate this thesis in memory of my father, Konstantinos, whose life was unfortunately too short to enable him to share this important moment in my life with me.

Thank you — Dank u wel — Σάς ευχαριστώ!

TOC

Contents

TOC

List of figures

List of tables

TOC

List of Abbreviations

Acronym	Description
ABM	Agent-Based Model
API	Application Programming Interface
CA	Cellular Automaton
CBD	Central Business District
CDR	Call Detail Record
CSR	Complete Spatial Randomness
CSV	Comma Separated Value
DBMS	DataBase Management System
DCAT	Data Catalog vocabulary
DSN	Distributed Sensor Network
DSS	Decision Support System
ESDA	Exploratory Spatial Data Analysis
E/R	Entity/Relationships model
FDR	False Discovery Rate
FOAF	Friend Of A Friend
FOV	Field of View
GIS	Geographic Information Systems
GML	Geography Markup Language
GPS	Geographic Positioning System
GTFS	General Transit Feed Specification
HCI	Human-Computer Interaction
HGC	Human-Generated Content (*sometimes UGC: User-Generated Content)
HTML	Hyper-Text Markup Language
HTTP	Hyper-Text Transfer Protocol
IoT	Internet of Things
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KML	Keyhole Markup Language
LBCS	Land-Based Classification Standards
LBSN	Location-Based Social Network
LDA	Latent Dirichlet Allocation
LISA	Local Indicators of Spatial Association
LOD	Linked Open Data
LoD	Level of Detail
LUCAS	Land Use/Cover Area Statistical survey

>>>

Acronym	Description
LUCC	Land Use/Land Cover Change
LUTI	Land Use Transportation Interaction
LUTM	Land Use Transportation Model
MAS	Multi-Agent System
MAUP	Modifiable Areal Unit Problem
OGC	Open Geospatial Consortium
OLS	Ordinary Least Squares
OTN	Ontology of Transportation Networks
OWL	Web Ontology Language
OWL DL	Web Ontology Language Description Logic
OWL2	Web Ontology Language 2 Existential Logic
OWL2	Web Ontology Language 2 Query Language
OWL2	Web Ontology Language 2 Rule Language
POI	Point of Interest
PSS	Planning Support System
RDBMS	Relational DataBase Management System
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RDSMS	Relational Data Stream Management System
RFID	Radio-Frequency Identification
ROUTE	Route Ontology of Urban Transportation Entities
SDI	Spatial Data Infrastructure
SN	Sensor Network
SPARQL	SPARQL Protocol And RDF Query Language
SQL	Structured Query Language
SSN	Semantic Sensor Network
TTL	Terse RDF Triple Language (Turtle)
UI	User Interface
UML	Unified Modeling Language
URI	Uniform Resource Identifier
VGI	Volunteered Geographic Information
VoID	Vocabulary of Interlinked Datasets
W3C	World Wide Web Consortium
WOB	Web Ontology Browser
XML	eXtensible Markup Language

Summary

The study of dynamic spatial and social phenomena in cities has evolved rapidly in the recent years, yielding new insights into urban dynamics. This evolution is strongly related to the emergence of new sources of data for cities, which have potential to capture dimensions of social and geographic systems that are difficult to detect in traditional urban data (e.g. census data). The majority of datasets that are generated from these new sources (e.g. sensors, mobile phones, online social media etc.) are spatially and temporally disaggregated, addressing short time intervals and individual locations of places and social agents. However, as the available sources increase in number, the produced datasets increase in diversity. Although the current capabilities of computing systems allow the storage, processing, analysis, and visualization of large-scale data, integration remains a challenge. In tackling the multifarious social, economic, and environmental challenges facing cities due to rapid urbanization, planners and policy makers need supporting frameworks to capitalize on the new possibilities given by emerging sources of social urban data.

To address the above challenge, this thesis proposes the design of a framework of novel methods and tools for the integration, visualization, and exploratory analysis of large-scale and heterogeneous social urban data to facilitate the understanding of urban dynamics. The research focuses particularly on the spatiotemporal dynamics of human activity in cities, as inferred from different sources of social urban data. The main objective is to provide new means to enable the incorporation of heterogeneous social urban data into city analytics, and to explore the influence of emerging data sources on the understanding of cities and their dynamics.

In association with the aim and objective of this thesis, the main research question is:

"How to integrate heterogeneous and multidimensional social urban data into the analysis of human activity dynamics in cities?"

The main question is further divided into five sub-questions. Accordingly, the research design is organized into five main parts, each one corresponding to one of the five sub-questions. The methods used to answer the research questions, along with the corresponding findings are presented in the following paragraphs.

What are the characteristics that distinguish emerging social urban data from traditional ones? — (Chapter 2)

After formulating the research aim, objectives, and scope, the concept of "social urban data" is introduced and defined (Chapter 2) to encompass data for cities that:

- are generated either directly or indirectly from people and their actions;
- derive from emerging sources such as sensors, mobile phones, geo-enabled social media, and LBSNs;
- are multidimensional in nature, meaning that they are spatially and temporally referenced;
- can be used to infer spatial, temporal, and social aspects of human movement, activity, and social connectivity;
- but are less structured and more semantically ambiguous than traditional urban data.

Following up on this definition, the characteristics of social urban data are described in comparison with traditional data for cities, by reviewing existing literature. The characteristics are namely: *diversity*, *scale*, *timeliness*, *structure*, *spatiotemporal resolution*, *semantic expressiveness*, *representativeness*, and *veracity*. Chapter 2 explores the extent to which each of the aforementioned characteristics typifies a certain data type or source and, further, investigates the strengths and weaknesses of social urban data as proxies for the analysis of urban dynamics. The identified strengths and weaknesses are used as a general basis for the design of the various methods and tools proposed by this research.

Social urban data do not comprise a unified category of data with common characteristics. In fact, according to the source that generates them (i.e. sensors, mobile phones, geo-enabled social media, and LBSNs), they may be characterized by varied levels of diversity, scale, timeliness, structure, spatiotemporal resolution, semantic expressiveness, representativeness, and veracity. However, it is argued that the eight aforementioned characteristics are not only inherent to emerging social urban data, but are also present – to a greater or lesser extent – in traditional data for cities.

The most distinguishing characteristic that differentiates emerging social urban data from traditional ones, is the purpose guiding their generation. Although conventional data for cities are created ad hoc, social urban data are generated organically and serve a variety of purposes. As such, they contain contextual, technological, geographical, demographic, and cultural biases, which in turn affect the overall data quality. In using social urban data as proxies for the analysis of urban dynamics, the identification of these biases is of critical importance to the interpretation of the obtained results. To leverage the intrinsic biases of social urban data and to extract unambiguous knowledge about the dynamics of cities, the integration of data from multiple sources is, therefore, deemed necessary.

How to transform heterogeneous data for cities into multidimensional linked urban data? — (Chapter 3)

Drawing on the necessity to employ different types of urban data in the analysis of cities and their dynamics, approaches to data integration are explored (Chapter 3). The fusion of data from multiple sources is hardly straightforward. What makes the assembly cumbersome, is in fact the inherent diversities of the sources from which the data stem. More specifically, the heterogeneities may pertain to differences in syntax (i.e. different data encoding), schemas (i.e. different structure and entity relationships), semantics (i.e. diverse contextual interpretations), or combinations of these three aspects.

In mitigating the various heterogeneities, a methodology for the transformation of heterogeneous data for cities into multidimensional linked urban data is designed and presented in Chapter 3. The methodology follows an ontology-based data integration approach and accommodates a variety of semantic (web) and linked data technologies. Overall, it comprises three main processes, namely: (a) urban data integration, (b) linked urban data generation, and (c) publication to the LOD cloud. In a nutshell, the proposed methodology consists of the following steps:

- Semantic integration:
  - Selection of data sources and data preprocessing
  - Data analysis and modeling
    - Schema extraction
    - Resource naming strategy definition
    - Ontology design and development
      - Terms extraction
      - Reuse of existing ontologies and external structured vocabularies
      - Terms hierarchy and ontology conceptualization
      - Ontology evaluation
    - Mapping source data to the ontology (data transformation)
- Transformation into multidimensional linked urban data:
  - Establishing links with other sources
- Publication to the LOD cloud:
  - Ontology and RDF dataset publication on the Web
  - Documentation accessibility (human-readable and machine-processable)
  - Registration into a Linked Data catalog and publication to the LOD cloud

The methodology is demonstrated through a use case, employing real-world data from multiple sources. In particular, nine large-scale spatiotemporal data sets are collected from three public transportation organizations and cover the entire public transport network of the city of Athens, Greece. As part of the data integration process, an ontology for public transportation systems is also designed and implemented.

The resulting integrated dataset is further linked to external resources to provide richer descriptions of the source data, and is eventually published to the LOD cloud.

The transformation of heterogeneous data for cities into multidimensional linked urban data has potential to provide richer descriptions of urban dynamics. Moreover, their publication to the LOD cloud facilitates their discovery and exploitation by stakeholders of different (city) domains. The methodology can be replicated and adapted to serve different types of (social) urban data, irrespective of the chosen sources. As it is based on ontologies, it also enables the semi-automatic iteration of the data mapping for any future updates of the source data, provided that the latter maintain their initial schemas.

How could urban planners, researchers, and policy makers leverage the potential of multidimensional linked data in city analytics? — (Chapter 4)

To encourage the consumption of linked urban data, as well as the incorporation of the above-described methodology (Chapter 3) into urban planning, research, and policy-making, a set of web-based tools for the visual representation of ontologies and linked data is designed and developed (Chapter 4). After reviewing existing approaches to and tools for ontology and linked data visualization, the identified limitations of related work set the basis and requirements for the design of the proposed tools. The tools – comprising the *OSMoSys* framework – provide graphical user interfaces for the visual representation, browsing, and interactive exploration of both ontologies and linked urban data. The use of different visualizations – in the form of interactive web documents and force-directed graphs – aim to support the adoption and consumption of linked urban data, without requiring extensive knowledge of the technology stack that underpins them. Therefore, the tools provide easy-to-use interfaces, accessible to a wide range of users, either experienced or amateur ones.

To further support the production of multidimensional linked urban data, an upper-level ontology is developed that formally describes and represents the relationships between the various elements of urban networks, pertinent to both the social and spatial sphere of urban systems. Individual datasets with heterogeneous attributes can be mapped to the aforementioned ontology and fused into a single dataset that combines the different attributes together.

The overall *OSMoSys* framework uses solely open software and standards, is provided under open licenses, and can be accessed through commonly-used web browsers. One of the aims of this framework is to assist in bridging, to some extent, the gap between linked data consumers and ontology engineers. Moreover, it can be used by domain experts as a basis to evaluate ontologies under development. Two ontologies and one large-scale linked dataset are used as benchmarks to test the potential and limitations of the framework.

What types of attributes can be derived from social urban data in relation to the dynamics of human activity? — (Chapter 5)

After introducing new methods (Chapter 3) and tools (Chapter 4) for the generation of linked data for cities that could offer richer descriptions of the urban environment than data from a single source, the attributes that can be derived from various social urban data are investigated (Chapter 5). Besides multidimensional linked urban data, it is also possible to derive several attributes of people and places from different geo-enabled social media content and LBSN data. To extract these attributes, a set of methods and techniques are described.

Prior to this, different approaches to measuring, modeling, and characterizing urban space are discussed, by reviewing existing literature. The focus is on the attributes – derived from both traditional and emerging sources of data – that have been used hitherto to measure and model urban systems and their dynamics. Next, the types of attributes in addition to the methods and techniques for extracting them, primarily from geo-enabled social media and LBSNs, are described. The derived attributes refer to characteristics of both the people who perform a certain (social) activity (e.g. socio-demographic characteristics, home location, individual trajectory, activity space, sentiments etc.) and the places where activities occur (e.g. land use, type of activity). The attributes are classified into four categories according to the nature of the feature they describe, namely: (1) *socio-demographic attributes*, (2) *functional attributes of places*, (3) *individual spatial movement patterns*, and (4) *topical attributes*. The first category refers to the approximated home location of individuals and characteristics such as gender, age range, and ethnicity. The second category refers to approximated land uses of POIs. The third category is about individual trajectories and activity spaces. The fourth category refers to the semantics and sentiments that can be derived from social media content. Further, Chapter presents how the derived attributes help measure the functional density and diversity of urban areas, as well as the geographical extents of activity spaces over different periods of time.

The incorporation of these attributes into urban analytics helps deviate from traditional approaches, in which people and places are usually perceived as aggregate uniform parameters within spatial subdivisions. The methods and techniques to extract disaggregate attributes from social urban data set the foundation for the design of a system that performs analyses on these attributes and provides insight into the dynamics of human activity in cities (Chapter 6).

How do different sources of social urban data influence the understanding of the spatiotemporal dynamics of human activity in cities? — (Chapter 6)

After introducing methods and tools for data integration (Chapter 3), visual exploration of linked urban data (Chapter 4), and derivation of various attributes of people and places from different social urban data (Chapter 5), it is examined how they can all be combined into a single platform and put to use in understanding spatiotemporal patterns of human activity in cities. To achieve this, a novel web-based system for the visualization and exploratory analysis of human activity dynamics is designed (Chapter 6). The system (coined *SocialGlass*) combines data from various geo-enabled social media (i.e. Twitter, Instagram, Sina Weibo) and LBSNs (i.e. Foursquare), sensor networks (i.e. GPS trackers, Wi-Fi cameras), and conventional socio-economic urban records, but also has the potential to employ custom datasets from other sources. Further, it accommodates a variety of visualization types and data filters to support the visual exploratory analysis of the spatiotemporal dynamics of human activity, as inferred from different social media.

A real-world case study is also analyzed and used as a demonstrator of the capacities of the proposed web-based system in the study of urban dynamics (Chapter 6). The case study explores the potential impact of a city-scale event (i.e. the Amsterdam Light festival 2015) on the activity and movement patterns of different social categories (i.e. residents, non-residents, foreign tourists), as compared to their daily and hourly routines in the periods before and after the event. The aim of the case study is twofold. First, to assess the potential and limitations of the proposed system and, second, to investigate how different sources of social urban data could influence the understanding of urban dynamics. To this end, a visual exploratory analysis is conducted on the collected data with the use of the *SocialGlass* system, in addition to a spatial autocorrelation analysis on 28 different variables of human activity, using global and local indices of autocorrelation along with statistical tests to assess the significance of the obtained results.

The findings of the case study suggested that it is necessary to consider different social categories of people, rather than aggregate populations, when studying the dynamics of human activity and movement behavior. Moreover, if social urban data – especially online social media – are used as proxies for the analysis of urban dynamics, the data collection period and the data source play a crucial role, when it comes to anomalies that could be reflected in the collected data, which could in turn lead to biased interpretations.

Conclusions and outlook

The increasing availability of data for cities that are generated by emerging sources, such as sensor networks, mobile phones, geo-enabled social media, and LBSNs have the potential to provide new insights into urban dynamics, but also create new challenges for urban planners, researchers, and policy makers. These data are mainly characterized by heterogeneity, owing to the variety of sources and the diversity of purposes they serve, and multidimensionality, meaning that the information they contain may simultaneously address spatial, social, temporal, and topical features of people and places. In addition, they offer new perspectives on how complex socio-spatial phenomena in cities change over shorter time intervals, compared to the sparsely updated conventional urban data. On the downside, though, is the muddled data structure, the ambiguous semantics of the contained information, and the several biases (of contextual, demographic, cultural, geographic, technological, or other nature).

The contribution of this doctoral thesis is the design and development of a framework of novel methods and tools that enables the fusion of heterogeneous data for cities and potentially fosters planners, researchers, and policy makers to capitalize on the new possibilities given by emerging social urban data. Having a deep understanding of the spatiotemporal dynamics of cities and, especially of the activity and movement behavior of people, is expected to play a crucial role in addressing the challenges of rapid urbanization. The adaptability of the methods and tools comprising the proposed framework enables them to serve scientific fields beyond urban science and spatial analysis, such as computational social science, urban geography, GIScience, and (human) mobility studies. Future research could focus on the development of multilayered urban models that connect the geographical with the social networks of cities, as well as on comparative studies of urban dynamics across several urban systems, in both developed and developing countries, using the developed tools. Overall, the framework proposed by this research has potential to open avenues of quantitative explorations of urban dynamics by employing a wide range of available data sources, contributing to the development of a new science of cities.

TOC

Korte Inhoud

Het onderzoek naar dynamische, ruimtelijke, en sociale fenomenen in steden is in de laatste jaren sterk ontwikkeld, hetgeen heeft geleid tot nieuwe inzichten in stedelijke dynamiek. Deze ontwikkeling is sterk gerelateerd aan het beschikbaar komen van nieuwe bronnen van data over steden (bv. sensoren, mobiele telefoons, online sociale media, etc.), die de potentie hebben dimensies van sociale en geografische systemen te duiden die moeilijk te beschrijven waren op basis van meer traditionele data (zoals volkstellingen). Omdat er echter steeds meer bronnen beschikbaar komen, zijn de resulterende datasets ook steeds meer divers. Behalve deze heterogeniteit, zijn nieuwe sociaal-stedelijke datasets ook multidimensionaal. Dit laatste houdt in dat ze tegelijkertijd informatie bevatten over zowel locaties, sociale aspecten, tijdsaspecten, en onderwerps-aspecten van personen en plaatsen. Daarom blijft het integreren en de geo-spatiële analyse van deze multidimensionale data een uitdaging. De vraag rijst daarom hoe dergelijke heterogene en multidimensionale sociaal-stedelijke data geïntegreerd kan worden ten behoeve van het analyseren van menselijke activiteit in steden.

Als antwoord op die vraag beschrijft dit proefschrift het ontwerp van een kader aan nieuwe methoden en middelen voor de integratie, visualisatie, en exploratieve analyse van grootschalige en heterogene sociaal-stedelijke data met als doel het begrip van stedelijke dynamiek te vergroten. Het onderzoekt richt zich met name op de spatio-temporele dynamiek van de menselijke activiteit in steden, zoals die is afgeleid uit verschillende bronnen van sociaal-stedelijke data. Het belangrijkste doel is om nieuwe middelen aan te reiken om heterogene sociaal-stedelijke data te betrekken bij het maken van stedelijke analyses en om onderzoek te doen naar de invloed van opkomende databronnen op het begrijpen van steden en hun dynamiek.

Daarom is er, om de verschillende soorten heterogeniteit te compenseren, een methodologie ontworpen voor het omzetten van heterogene data over steden in multidimensionale, gekoppelde stadsdata. Voor die methodologie wordt een benadering voor data-integratie op basis van ontologieën gehanteerd die ruimte biedt aan een veelheid aan technologieën op basis van semantische en gekoppelde (web) data. Er is een use case met onderlinge datakoppeling gebruikt om de voorgestelde methodologie te demonstreren. De use case maakt gebruik van negen grootschalige spatio-temporele datasets uit de praktijk van drie ov-organisaties, die samen het gehele ov-netwerk van de stad Athene (Griekenland) dekken.

Om het gebruik van gekoppelde stadsdata door planners en beleidsmakers nog verder te stimuleren, is er een set webtools ontworpen en ontwikkeld voor de visuele weergave van ontologieën en gekoppelde data. Deze tools – die samen het *OSMoSys*-kader vormen –

hebben grafische gebruikersinterfaces voor de visuele weergave, het doorbladeren en de interactieve verkenning van zowel ontologieën als gekoppelde stadsdata.

Na de introductie van methodes en hulpmiddelen voor data-integratie, visuele verkenning van gekoppelde stadsdata en de afleiding van verschillende kenmerken van mensen en plaatsen uit diverse sociale stadsdata, is onderzocht hoe deze allemaal kunnen worden gecombineerd tot één platform. Daarvoor is er een nieuw systeem (genoemd *SocialGlass*) op webbasis ontworpen voor het visualiseren en verkennend analyseren van de dynamiek van menselijke activiteit. Dit systeem combineert data uit verschillende sociale media met geofunctie (Twitter, Instagram en Sina Weibo) en LBSN's (Foursquare), sensornetwerken (gps-trackers, wificamera's) en de conventionele sociaaleconomische stadsadministratie, maar kan ook worden gebruikt voor andere datasets afkomstig uit andere bronnen.

Er is een casestudy gebruikt om de mogelijkheden van het voorgestelde websysteem voor het bestuderen van stedelijke dynamiek te demonstreren. In de casestudy zijn de potentiële gevolgen van een stadsbreed evenement (het Amsterdam Light Festival 2015) voor de activiteit en bewegingspatronen van verschillende sociale categorieën (bewoners, bezoekers, buitenlandse toeristen) in kaart gebracht en vergeleken met de dagelijkse en uurlijkse routine van die categorieën in de periodes voor en na het evenement. De casestudy heeft een tweeledig doel: in de eerste plaats het beoordelen van de mogelijkheden en beperkingen van het voorgestelde systeem, en in de tweede plaats het onderzoeken van de manier waarop verschillende bronnen van stadsdata onze interpretatie van stedelijke dynamiek kunnen beïnvloeden.

De bijdrage die dit proefschrift levert is het ontwerp en de ontwikkeling van een kader aan nieuwe methodes en middelen die de combinatie van heterogene, multidimensionale data over steden mogelijk maakt. Dit kader kan planners, onderzoekers en beleidsmakers stimuleren om gebruik te maken van de nieuwe mogelijkheden die in opkomst zijnde sociale stadsdata bieden. Een diepgaand inzicht in de spatio-temporele dynamiek van steden – met name de activiteit en bewegingen van mensen – zal naar verwachting cruciaal zijn om het hoofd te bieden aan de uitdagingen die snelle verstedelijking met zich meebrengt. In zijn algemeenheid maakt het in dit onderzoek voorgestelde kader een kwantitatieve verkenning van stedelijke dynamiek mogelijk en levert daarmee een bijdrage aan de ontwikkeling van een nieuwe wetenschap met betrekking tot steden.

Περίληψη

Η μελέτη της δυναμικής των κοινωνικό-χωρικών φαινομένων έχει εξελιχθεί ραγδαία τα τελευταία χρόνια, παρέχοντας έτσι νέες οπτικές σε ζητήματα αστικής δυναμικής. Η εξέλιξη αυτή είναι άμεσα συνδεδεμένη με την ανάδυση νέων πηγών χωρικών δεδομένων, τα οποία εμπεριέχουν γνωρίσματα που δύσκολα εντοπίζονται στα συμβατικά χωρικά δεδομένα (πχ στα απογραφικά δεδομένα). Στην πλειονότητά τους τα δεδομένα αυτά προέρχονται από πηγές όπως, για παράδειγμα, αισθητήρες, κινητά τηλέφωνα, και μέσα κοινωνικής δικτύωσης. Είναι μάλιστα χωρικά και χρονικά επιμερισμένα και, ως εκ τούτου, καλύπτουν μικρά χρονικά διαστήματα (πχ ανά λεπτό), ενώ παράλληλα αφορούν μεμονωμένες τοποθεσίες αντί για μεγαλύτερες χωρικές ενότητες, οι οποίες συνήθως συναντώνται στα συμβατικά χωρικά δεδομένα. Ωστόσο, όσο αυξάνεται ο αριθμός των διαθέσιμων πηγών, τόσο αυξάνεται και η ποικιλομορφία των παραγόμενων δεδομένων. Παρά τις τρέχουσες δυνατότητες των υπολογιστικών συστημάτων, όσων αφορά την αποθήκευση, επεξεργασία, ανάλυση, και απεικόνιση δεδομένων μεγάλης κλίμακας, το ζήτημα της ενοποίησης (integration) διαφορετικών δεδομένων παραμένει πρόκληση. Η αντιμετώπιση όμως των πολύπλοκων κοινωνικών, οικονομικών, και περιβαλλοντικών ζητημάτων των σύγχρονων πόλεων από πλευράς πολεοδόμων και φορέων χάραξης πολιτικής, ιδιαίτερα λόγω της ραγδαίας αστικοποίησης, καθιστά αναγκαία την ανάπτυξη υποστηρικτικών πλαισίων, τα οποία θα αξιοποιούν τις νέες δυνατότητες που παρέχονται από τη σύζευξη αναδυόμενων πηγών κοινωνικό-χωρικών δεδομένων.

Ανταποκρινόμενη στην παραπάνω πρόκληση, η παρούσα διατριβή προτείνει τον σχεδιασμό ενός πλαισίου καινοτόμων μεθόδων και υπολογιστικών εργαλείων για την ενοποίηση, απεικόνιση, και διερευνητική ανάλυση ανομοιογενών δεδομένων μεγάλης κλίμακας, με στόχο την κατανόηση της δυναμικής των σύγχρονων πόλεων. Η έρευνα εστιάζει συγκεκριμένα σε ζητήματα χωρικής και χρονικής μεταβολής της ανθρώπινης δραστηριότητας στις πόλεις, όπως αυτή συνάγεται από διαφορετικές πηγές κοινωνικό-χωρικών δεδομένων (social urban data). Στόχος της έρευνας είναι η παροχή νέων μέσων που επιτρέπουν την ενσωμάτωση ανομοιογενών κοινωνικό-χωρικών δεδομένων στη διαδικασία της χωρικής ανάλυσης, καθώς επίσης και η διερεύνηση του τρόπου με τον οποίο κάθε ένας από τους νέους τύπους δεδομένων επιδρά στην κατανόηση των αστικών συστημάτων και των δυναμικών τους.

Με βάση το παραπάνω αντικείμενο και στόχο της έρευνας, το κυρίως ερευνητικό ερώτημα είναι:

“Πώς καθίσταται εφικτή η ενσωμάτωση ανομοιογενών κοινωνικό-χωρικών δεδομένων στην ανάλυση της χωρικής και χρονικής κατανομής των ανθρώπινων δραστηριοτήτων στις πόλεις;”

Το παραπάνω ερώτημα επιμερίζεται σε πέντε υπο-ερωτήματα. Αντίστοιχα, το σχέδιο της έρευνας οργανώνεται σε πέντε βασικά τμήματα, κάθε ένα από τα οποία αντιστοιχεί σε ένα από τα πέντε υπο-ερωτήματα. Οι χρησιμοποιούμενες μέθοδοι για την απάντηση των ερευνητικών ερωτημάτων, καθώς και τα αντίστοιχα ευρήματα παρουσιάζονται στις επόμενες παραγράφους.

Ποια είναι τα χαρακτηριστικά γνωρίσματα που διακρίνουν τα αναδυόμενα κοινωνικό-χωρικά δεδομένα από τα συμβατικά χωρικά δεδομένα; — (Κεφάλαιο 2)

Έχοντας ήδη διατυπώσει το αντικείμενο, τη σκοπιμότητα, και το πεδίο της έρευνας, στη συνέχεια εισάγεται και ορίζεται η έννοια των «κοινωνικό-χωρικών δεδομένων» (Κεφάλαιο 2), ώστε να συμπεριλάβει τα δεδομένα εκείνα για τις πόλεις τα οποία:

- παράγονται άμεσα ή έμμεσα από τους ανθρώπους και τις δραστηριότητές τους,
- προέρχονται από αναδυόμενες πηγές, όπως αισθητήρες, κινητά τηλέφωνα, και μέσα κοινωνικής δικτύωσης που βασίζονται στη γεωγραφική θέση του χρήστη (geo-enabled social media & location-based social networks),
- είναι εκ φύσεως πολυδιάστατα, με την έννοια ότι εμπεριέχουν γνωρίσματα που αφορούν τόσο σε χωρικές όσο και σε χρονικές ιδιότητες,
- μπορούν να χρησιμοποιηθούν για την εξαγωγή χωρικών, χρονικών, και κοινωνικών πτυχών της ανθρώπινης κινητικότητας, δραστηριότητας, και κοινωνικής συνδεσιμότητας,
- αλλά υστερούν ως προς τη δομή και τη σημασιολογική ευκρίνεια σε σχέση με τα συμβατικά χωρικά δεδομένα.

Με βάση τον παραπάνω ορισμό, περιγράφονται τα χαρακτηριστικά γνωρίσματα των κοινωνικό-χωρικών δεδομένων σε σύγκριση με τα συμβατικά χωρικά δεδομένα, μέσα από ανασκόπηση της υπάρχουσας βιβλιογραφίας. Τα γνωρίσματα αυτά είναι: η ***ποικιλομορφία*** (diversity), η ***κλίμακα*** (scale), η ***χρονικότητα*** (timeliness), η ***δομή*** (structure), η ***χωρο-χρονική ανάλυση*** (spatiotemporal analysis), η ***σημασιολογική εκφραστικότητα*** (semantic expressiveness), η ***αντιπροσωπευτικότητα*** (representativeness), και η ***ειλικρίνεια*** (veracity). Το Κεφάλαιο 2 διερευνά τον βαθμό στον οποίο κάθε ένα από τα παραπάνω γνωρίσματα χαρακτηρίζει κάθε τύπο ή πηγή δεδομένων, ενώ παράλληλα εξετάζει τις δυνατότητες και τις αδυναμίες των κοινωνικό-χωρικών δεδομένων ως ενδιάμεσων (proxies) για την ανάλυση της αστικής δυναμικής. Στη συνέχεια, οι περιγραφόμενες δυνατότητες και αδυναμίες θέτουν τη βάση για τον σχεδιασμό των ποικίλων μεθόδων και εργαλείων που προτείνει η παρούσα έρευνα.

Ωστόσο, τα κοινωνικό-χωρικά δεδομένα δε συνιστούν μια ενιαία κατηγορία δεδομένων με κοινά χαρακτηριστικά. Ανάλογα με την πηγή από την οποία προέρχονται (αισθητήρες, κινητά τηλέφωνα, και μέσα κοινωνικής δικτύωσης)

μπορεί να χαρακτηρίζονται από ποικίλα επίπεδα ποικιλομορφίας, κλίμακας, χρονικότητας, δομής, χωρο-χρονικής ανάλυσης, σημασιολογικής εκφραστικότητας, αντιπροσωπευτικότητας, και ειλικρίνειας. Παρ' όλα αυτά, υποστηρίζεται ότι τα οχτώ προαναφερόμενα γνωρίσματα δεν είναι μόνο εγγενή χαρακτηριστικά των αναδυόμενων κοινωνικό-χωρικών δεδομένων, αλλά συναντώνται επίσης – σε μικρότερο ή μεγαλύτερο βαθμό – και στα συμβατικά χωρικά δεδομένα.

Το χαρακτηριστικότερο γνώρισμα που διαφοροποιεί τα αναδυόμενα κοινωνικό-χωρικά δεδομένα από τα συμβατικά, είναι ο σκοπός για τον οποίο παράγονται. Σε αντίθεση με τα τελευταία τα οποία παράγονται ad hoc, τα κοινωνικό-χωρικά δεδομένα μπορεί να εξυπηρετούν πολύ διαφορετικούς μεταξύ τους σκοπούς. Ως εκ τούτου, είναι δυνατόν να εμπεριέχουν στοιχεία μεροληψίας (biases) τεχνολογικής, γεωγραφικής, δημογραφικής, πολιτισμικής ή άλλης φύσεως, τα οποία με τη σειρά τους επιδρούν στη συνολική ποιότητα των παραγόμενων δεδομένων. Ο προσδιορισμός επομένως αυτών των στοιχείων είναι ζωτικής σημασίας για την ερμηνεία των αποτελεσμάτων, όταν στην ανάλυση της αστικής δυναμικής χρησιμοποιούνται κοινωνικό-χωρικά δεδομένα. Για να μετριαστούν αλλά και για να αξιοποιηθούν τα εγγενή στοιχεία μεροληψίας των κοινωνικό-χωρικών δεδομένων, με στόχο την εξαγωγή σαφούς γνώσης σχετικά με τη δυναμική των πόλεων, κρίνεται απαραίτητη η ενοποίηση (integration) δεδομένων προερχόμενων από διαφορετικές πηγές.

Πώς τα ανομοιογενή χωρικά δεδομένα δύνανται να μετασχηματιστούν σε πολυδιάστατα διασυνδεδεμένα χωρικά δεδομένα (multidimensional linked urban data); — (Κεφάλαιο 3)

Με βάση την ανάγκη χρήσης διαφορετικών τύπων χωρικών δεδομένων στην ανάλυση των πόλεων και των δυναμικών τους, διερευνώνται υπάρχουσες προσεγγίσεις σε ζητήματα ενοποίησης (Κεφάλαιο 3). Η σύζευξη δεδομένων από διαφορετικές πηγές δεν είναι μία απλή διαδικασία. Αυτό που καθιστά δύσκολη τη διασύνδεση, είναι οι εγγενείς διαφορές των πηγών από τις οποίες προέρχονται τα δεδομένα. Πιο συγκεκριμένα, αυτές οι ανομοιογένειες μπορεί να αφορούν συντακτικές (διαφορετική κωδικοποίηση), σχηματικές (διαφορετική δομή και συσχετίσεις οντοτήτων), σημασιολογικές (διαφορετικές ερμηνείες) διαφορές, ή συνδυασμούς αυτών.

Με στόχο τη μετρίαση των διαφόρων ανομοιογενειών, προτείνεται μια μεθοδολογία μετατροπής ανομοιογενών χωρικών δεδομένων σε πολυδιάστατα διασυνδεδεμένα χωρικά δεδομένα, η οποία παρουσιάζεται στο Κεφάλαιο 3. Η μεθοδολογία ακολουθεί την προσέγγιση της ενοποίησης δεδομένων με βάση οντολογίες (ontology-based data integration), ενώ παράλληλα βασίζεται σε τεχνολογίες σημασιολογικού ιστού (semantic web) και διασυνδεδεμένων δεδομένων (linked data). Συνολικά, αποτελείται από τρεις διαδικασίες: (α) την ενοποίηση χωρικών δεδομένων, (β) την παραγωγή διασυνδεδεμένων χωρικών δεδομένων, και (γ) τη δημοσίευση τους στο Linked Open

Data (LOD) cloud. Πιο αναλυτικά, η μεθοδολογία αποτελείται από τα ακόλουθα βήματα:

- Σημασιολογική ενοποίηση (semantic integration):
  - Επιλογή πηγών δεδομένων και προ-επεξεργασία
  - Ανάλυση και μοντελοποίηση δεδομένων
    - Εξαγωγή σχήματος
    - Καθορισμός στρατηγικής ονοματοθεσίας πόρων
    - Σχεδιασμός και ανάπτυξη οντολογίας
      - Εξαγωγή όρων
      - Επαναχρησιμοποίηση υφιστάμενων οντολογιών και δομημένων λεξιλογίων
      - Ιεράρχηση όρων και σύλληψη οντολογίας
      - Αξιολόγηση οντολογίας
    - Απεικόνιση δεδομένων στην οντολογία
- Μετασχηματισμός σε πολυδιάστατα διασυνδεδεμένα δεδομένα:
  - Δημιουργία δεσμών με εξωτερικές πηγές δεδομένων
- Δημοσίευση στο LOD cloud:
  - Δημοσίευση οντολογίας και RDF δεδομένων στον Παγκόσμιο Ιστό
  - Τεκμηρίωση και προσβασιμότητα
  - Εγγραφή σε κατάλογο Διασυνδεδεμένων Δεδομένων και δημοσίευση στο LOD cloud

Η μεθοδολογία παρουσιάζεται μέσα από μία μελέτη περίπτωσης, η οποία περιλαμβάνει τη χρήση πραγματικών δεδομένων από πολλαπλές πηγές. Συγκεκριμένα, συλλέγονται εννέα σύνολα χωρο-χρονικών δεδομένων μεγάλης κλίμακας, προερχόμενα από τρεις δημόσιους οργανισμούς μεταφορών, τα οποία καλύπτουν το σύνολο του δικτύου δημόσιων συγκοινωνιών της Αθήνας. Ως μέρος της διαδικασίας ενοποίησης των δεδομένων, σχεδιάζεται και αναπτύσσεται μια οντολογία για τα δημόσια συστήματα μεταφορών. Το παραγόμενα ενοποιημένα δεδομένα αναπτύσσουν, στη συνέχεια, δεσμούς με εξωτερικές πηγές και δημοσιεύονται στο LOD cloud.

Ο μετασχηματισμός ανομοιογενών χωρικών δεδομένων σε πολυδιάστατα διασυνδεδεμένα χωρικά δεδομένα δύναται να παρέχει πληρέστερες περιγραφές της αστικής δυναμικής. Επιπλέον, η δημοσίευση στο LOD cloud καθιστά ευκολότερη την εύρεση και αξιοποίηση τους. Η μεθοδολογία μπορεί να αναπαραχθεί και να προσαρμοστεί στις ανάγκες διαφορετικών τύπων (κοινωνικό-)χωρικών δεδομένων, ανεξάρτητα από την πηγή από την οποία προέρχονται. Καθόσον βασίζεται σε οντολογικά μοντέλα, δίνει επιπλέον τη δυνατότητα ημι-αυτόματης επανάληψης της διαδικασίας απεικόνισης των δεδομένων, ανεξάρτητα από τον ρυθμό ανανέωσης, με την προϋπόθεση τα δεδομένα αυτά να διατηρούν το αρχικό τους σχήμα (βάσης δεδομένων).

Με ποιον τρόπο μπορούν οι πολεοδόμοι και οι φορείς χάραξης αστικής πολιτικής να αξιοποιήσουν τις δυνατότητες των πολυδιάστατων διασυνδεδεμένων χωρικών δεδομένων στη χωρική ανάλυση; — (Κεφάλαιο 4)

Με σκοπό να ενθαρρυνθεί η ευρύτερη χρήση διασυνδεδεμένων χωρικών δεδομένων, καθώς και η ενσωμάτωση της παραπάνω αναφερόμενης μεθοδολογίας (Κεφάλαιο 3) στον αστικό σχεδιασμό, έρευνα, και χάραξη αστικής πολιτικής, σχεδιάζεται και αναπτύσσεται μια σειρά από διαδικτυακά (web-based) εργαλεία για την αναπαράσταση οντολογιών και διασυνδεδεμένων δεδομένων (Κεφάλαιο 4). Ύστερα από ανασκόπηση υπαρχουσών προσεγγίσεων και εργαλείων για την αναπαράσταση οντολογιών και διασυνδεδεμένων δεδομένων, εντοπίζονται οι σχετικοί περιορισμοί οι οποίοι με τη σειρά τους θέτουν τη βάση για τον σχεδιασμό των προτεινόμενων υπολογιστικών εργαλείων. Τα εργαλεία αυτά συνιστούν στο σύνολό τους την πλατφόρμα *OSMoSys*, η οποία παρέχει ένα γραφικό περιβάλλον εργασίας (graphical user interface) για την απεικόνιση, περιήγηση, και διερεύνηση οντολογιών και διασυνδεδεμένων δεδομένων. Η χρήση διαφορετικών απεικονίσεων – υπό τη μορφή δυναμικών γραφημάτων (force-directed graphs) – αποσκοπεί στην ευρύτερη υιοθέτηση και χρήση διασυνδεδεμένων δεδομένων, χωρίς να απαιτείται εξειδικευμένη γνώση οντολογικής μηχανικής ή τεχνολογιών σημασιολογικού ιστού. Ως εκ τούτου, τα εργαλεία είναι εύκολα προσβάσιμα από μια ευρεία γκάμα χρηστών, έμπειρων και μη.

Για την περαιτέρω υποστήριξη της παραγωγής πολυδιάστατων διασυνδεδεμένων χωρικών δεδομένων, αναπτύσσεται μια οντολογία ανώτερου επιπέδου (upper-level ontology), η οποία περιγράφει τη συσχέτιση μεταξύ των διαφόρων στοιχείων που συνθέτουν τα αστικά δίκτυα (urban networks), τόσο από την κοινωνική όσο και από τη χωρική σκοπιά των αστικών συστημάτων. Με αυτό τον τρόπο, δίνεται η δυνατότητα να απεικονιστούν δεδομένα από διαφορετικές πηγές με αρκετά ανομοιογενή γνωρίσματα στην προαναφερόμενη οντολογία και να ενοποιηθούν σε ένα και μοναδικό σύνολο δεδομένων, το οποίο συνδυάζει τα διαφορετικά γνωρίσματα μεταξύ τους.

Η πλατφόρμα *OSMoSys* κάνει αποκλειστική χρήση ανοιχτού λογισμικού, ενώ και η ίδια είναι προσβάσιμη από τους περισσότερους web browsers. Ένας από τους στόχους της πλατφόρμας είναι να συμβάλλει στη γεφύρωση του χάσματος μεταξύ των χρηστών διασυνδεδεμένων δεδομένων και των ειδικών στην οντολογική μηχανική. Για τον έλεγχο των δυνατοτήτων και των αδυναμιών της προτεινόμενης πλατφόρμας εργαλείων χρησιμοποιούνται δύο οντολογίες και ένα σύνολο διασυνδεδεμένων δεδομένων μεγάλης κλίμακας ως σημεία αναφοράς.

Ποια γνωρίσματα σχετικά με τη δυναμική της ανθρώπινης δραστηριότητας στο χώρο μπορούν να εξαχθούν από τα κοινωνικό-χωρικά δεδομένα; — (Κεφάλαιο 5)

Ύστερα από την εισαγωγή νέων μεθόδων (Κεφάλαιο 3) και υπολογιστικών εργαλείων (Κεφάλαιο 4) για την παραγωγή διασυνδεδεμένων χωρικών δεδομένων, τα οποία μπορούν να παρέχουν πληρέστερες περιγραφές του αστικού περιβάλλοντος σε σύγκριση με τα δεδομένα που προέρχονται από μία και μόνο πηγή, εξετάζονται τα γνωρίσματα που μπορούν να εξαχθούν από διάφορους τύπους κοινωνικό-χωρικών δεδομένων (Κεφάλαιο 5). Εκτός των διασυνδεδεμένων χωρικών δεδομένων, υπάρχει η δυνατότητα άντλησης γνωρισμάτων ανθρώπων ή τοποθεσιών από δεδομένα τα οποία παράγονται στα μέσα κοινωνικής δικτύωσης που βασίζονται στη γεωγραφική θέση του χρήστη (geo-enabled social media and LBSN data). Παρουσιάζεται, επομένως, ένα σύνολο μεθόδων και τεχνικών για την εξαγωγή αυτών των γνωρισμάτων.

Πριν από αυτό, το Κεφάλαιο 5 καταγράφει διαφορετικές προσεγγίσεις μέτρησης, μοντελοποίησης, και χαρακτηρισμού του αστικού χώρου, μέσα από ανασκόπηση της υπάρχουσας βιβλιογραφίας. Το κεφάλαιο εστιάζει στα γνωρίσματα εκείνα που μπορούν να αντληθούν τόσο από συμβατικές όσο και από αναδυόμενες πηγές δεδομένων, και τα οποία έχουν χρησιμοποιηθεί μέχρι σήμερα για τη μέτρηση και μοντελοποίηση των αστικών συστημάτων και των δυναμικών τους. Στη συνέχεια, περιγράφονται οι τύποι των γνωρισμάτων εκείνων που αφορούν στο είδος και την κατανομή της ανθρώπινης δραστηριότητας στο χώρο, τα οποία μπορούν να αντληθούν κυρίως από μέσα κοινωνικής δικτύωσης, ενώ παράλληλα παρουσιάζονται οι μέθοδοι και τεχνικές για την εξαγωγή τους. Τα εξαγόμενα γνωρίσματα αφορούν σε χαρακτηριστικά τόσο των ίδιων των ανθρώπων (π.χ. δημογραφικά χαρακτηριστικά, θέση κατοικίας, ατομικές τροχιές κίνησης στον χώρο, χώροι δραστηριότητας, συναισθήματα κλπ.) οι οποίοι επιτελούν μία συγκεκριμένη (κοινωνική) δραστηριότητα, όσο και σε χαρακτηριστικά των τοποθεσιών (π.χ. χρήσεις γης, τύπος δραστηριότητας κλπ.) στις οποίες πραγματοποιούνται οι εν λόγω δραστηριότητες. Τα γνωρίσματα ταξινομούνται σε τέσσερις κατηγορίες, σύμφωνα με τα χαρακτηριστικά τα οποία περιγράφουν. Οι κατηγορίες αυτές έχουν ως εξής: (1) ***κοινωνικό-δημογραφικά γνωρίσματα***, (2) ***λειτουργικά γνωρίσματα***, (3) ***γνωρίσματα σχετικά με την κίνηση των ατόμων στο χώρο***, και (4) ***θεματικά γνωρίσματα***. Η πρώτη κατηγορία αφορά την εκτιμώμενη θέση κατοικίας ενός ατόμου, καθώς επίσης και γνωρίσματα σχετικά με το φύλο, το ηλικιακό εύρος, και την εθνικότητα. Η δεύτερη κατηγορία αναφέρεται στις εκτιμώμενες χρήσεις γης των σημείων ενδιαφέροντος (points of interest – POIs). Η τρίτη κατηγορία αφορά τις τροχιές κίνησης και τους χώρους δραστηριότητας (activity spaces) ενός ατόμου. Η τέταρτη κατηγορία αναφέρεται στα συναισθήματα και στο περιεχόμενο των posts στα μέσα κοινωνικής δικτύωσης. Επιπροσθέτως, στο Κεφάλαιο 5 παρουσιάζεται ο τρόπος με τον οποίο τα παραπάνω εξαγόμενα γνωρίσματα μπορούν να συμβάλλουν στη μέτρηση της λειτουργικής πυκνότητας (functional density) και ποικιλομορφίας (functional diversity) αστικών περιοχών, καθώς

και στη μέτρηση του γεωγραφικού εύρους των ατομικών χώρων δραστηριότητας σε διαφορετικά χρονικά διαστήματα.

Η ενσωμάτωση των γνωρισμάτων αυτών στη χωρική ανάλυση συμβάλλει στην απόκλιση από τις συμβατικές προσεγγίσεις, στις οποίες οι άνθρωποι και οι χώροι δραστηριοποίησής τους αντιμετωπίζονται ως ενιαίες και ομοιόμορφα κατανεμημένες παράμετροι εντός προκαθορισμένων χωρικών ενοτήτων (π.χ. δημοτικά ή περιφερειακά διαμερίσματα). Οι μέθοδοι και τεχνικές άντλησης των επιμερισμένων (disaggregate) γνωρισμάτων από κοινωνικό-χωρικά δεδομένα θέτει τις βάσεις για τον σχεδιασμό ενός υπολογιστικού συστήματος, το οποίο αναλύει τα συγκεκριμένα γνωρίσματα και παρέχει περαιτέρω πληροφορίες σχετικά με τις μεταβολές της ανθρώπινης δραστηριότητας στο χώρο και τον χρόνο (Κεφάλαιο 6).

Πώς επιδρούν οι διάφορες πηγές κοινωνικό-χωρικών δεδομένων στην κατανόηση των χωρικών και χρονικών μεταβολών της ανθρώπινης δραστηριότητας στις πόλεις; — (Κεφάλαιο 6)

Έχοντας ήδη παρουσιάσει μεθόδους και υπολογιστικά εργαλεία για την ενοποίηση (Κεφάλαιο 3) και την απεικόνιση διαφόρων τύπων χωρικών δεδομένων (Κεφάλαιο 4), καθώς επίσης και για την άντληση ποικίλων γνωρισμάτων από διαφορετικές πηγές κοινωνικό-χωρικών δεδομένων (Κεφάλαιο 5), εξετάζεται πώς όλα τα παραπάνω μπορούν να συνδυαστούν σε μία κοινή υπολογιστική πλατφόρμα, ώστε να ενισχύσουν την κατανόηση των χωρικών και χρονικών μοτίβων της ανθρώπινης δραστηριότητας στις πόλεις. Με αυτό τον στόχο, σχεδιάζεται ένα καινοτόμο διαδικτυακό (web-based) σύστημα για την απεικόνιση και διερευνητική ανάλυση των μεταβολών της ανθρώπινης δραστηριότητας (Κεφάλαιο 6). Το σύστημα με την ονομασία *SocialGlass* συνδυάζει δεδομένα από ποικίλα μέσα κοινωνικής δικτύωσης (π.χ. Twitter, Instagram, Sina Weibo, Foursquare), δίκτυα αισθητήρων (π.χ. συστήματα εντοπισμού θέσης GPS, κάμερες με δυνατότητα σύνδεσης Wi-Fi), καθώς και συμβατικές πηγές κοινωνικό-οικονομικών δεδομένων. Μπορεί, ωστόσο, να υποστηρίξει και δεδομένα που προέρχονται από πηγές οι οποίες δεν περιλαμβάνονται στις αμέσως προηγούμενες. Παράλληλα, το σύστημα παρέχει μια ποικιλία εργαλείων για το φιλτράρισμα και την απεικόνιση δεδομένων.

Μέσω της ανάλυσης μιας πραγματικής μελέτης περίπτωσης εξετάζονται οι δυνατότητες του διαδικτυακού συστήματος, όσον αφορά την κατανόηση της χωρικής δυναμικής (Κεφάλαιο 6). Η μελέτη περίπτωσης διερευνά τον πιθανό αντίκτυπο ενός event μεγάλης κλίμακας (συγκεκριμένα, του Amsterdam Light Festival 2015) στον τρόπο με τον οποίο διαφορετικές κοινωνικές κατηγορίες ανθρώπων (κάτοικοι, μη-κάτοικοι, ξένοι τουρίστες) κινούνται ή επιτελούν συγκεκριμένες δραστηριότητες στον χώρο, σε σύγκριση με την καθημερινή και ωριαία ρουτίνα τους τις περιόδους πριν και μετά το event. Ο στόχος αυτής της μελέτης είναι διττός. Πρώτον, αφορά

TOC

στην αξιολόγηση των δυνατοτήτων και περιορισμών του προτεινόμενου συστήματος και, δεύτερον, εστιάζει στη διερεύνηση του τρόπου με τον οποίο κοινωνικό-χωρικά δεδομένα προερχόμενα από διαφορετικές πηγές επιδρούν στην κατανόηση της χωρικής δυναμικής. Για τον σκοπό αυτό, διεξάγεται αρχικά οπτική διερεύνηση (visual exploratory analysis) στα δεδομένα που έχουν συλλεχθεί μέσω της πλατφόρμας *SocialGlass*. Στη συνέχεια, πραγματοποιείται χωρική αυτοσυσχέτιση (spatial autocorrelation analysis) 28 διαφορετικών μεταβλητών που αφορούν στην ανθρώπινη δραστηριότητα, κάνοντας χρήση ολικών και τοπικών δεικτών αυτοσυσχέτισης (ολικός δείκτης Moran's I, τοπικοί δείκτες Moran's I_i και Getis-Ord G_i^*) σε συνδυασμό με ελέγχους στατιστικής σημαντικότητας των αποτελεσμάτων (έλεγχος τυχαιοποίησης και αναδειγματοληψίας στις τιμές των ολικών και τοπικών δεικτών).

Τα αποτελέσματα της μελέτης περίπτωσης υποδεικνύουν ότι για την ακριβέστερη κατανόηση των χωρικών και χρονικών μεταβολών της ανθρώπινης δραστηριότητας είναι αναγκαία η θεώρηση διαφορετικών κοινωνικών κατηγοριών ανθρώπων αντί ενιαίων πληθυσμών (aggregate populations), όπως συνηθίζεται μέχρι σήμερα. Επίσης, στην περίπτωση που η ανάλυση της χωρικής δυναμικής βασίζεται σε κοινωνικό-χωρικά δεδομένα – και κυρίως σε δεδομένα προερχόμενα από μέσα κοινωνικής δικτύωσης – τότε η περίοδος συλλογής αλλά και η πηγή από την οποία προέρχονται τα χρησιμοποιούμενα δεδομένα παίζουν καθοριστικό ρόλο, ιδίως όσων αφορά στην ανίχνευση πιθανών ανωμαλιών, οι οποίες με τη σειρά τους μπορούν να οδηγήσουν σε εσφαλμένες ερμηνείες των αποτελεσμάτων.

Συμπεράσματα και προοπτικές

Η ολοένα και αυξανόμενη παραγωγή κοινωνικό-χωρικών δεδομένων από αναδυόμενες πηγές, όπως οι αισθητήρες, τα κινητά τηλέφωνα, και τα μέσα κοινωνικής δικτύωσης δύναται να παρέχει νέες γνώσεις σχετικά με τη δυναμική των πόλεων. Ταυτόχρονα, όμως, δημιουργεί νέες προκλήσεις για τους πολεοδόμους και τους φορείς χάραξης αστικής πολιτικής. Τα δεδομένα αυτά χαρακτηρίζονται από ανομοιογένεια – λόγω της ποικιλίας των πηγών από τις οποίες προέρχονται αλλά και λόγω των διαφορετικών σκοπών που εξυπηρετούν – και πολυ-διαστατικότητα (multidimensionality), καθώς οι πληροφορίες που εμπεριέχουν μπορούν να αναφέρονται ταυτόχρονα σε χωρικά, κοινωνικά, χρονικά, και θεματικά γνωρίσματα ανθρώπων και τοποθεσιών. Σε σύγκριση, μάλιστα, με τα συμβατικά χωρικά δεδομένα τα οποία δεν ενημερώνονται τακτικά, οι νέοι τύποι κοινωνικό-χωρικών δεδομένων προσφέρουν νέες οπτικές σχετικά με τον τρόπο με τον οποίο τα πολύπλοκα κοινωνικά και χωρικά φαινόμενα των πόλεων μεταβάλλονται μέσα σε μικρά χρονικά διαστήματα. Ωστόσο, στα αρνητικά χαρακτηριστικά συγκαταλέγονται η συχνά συγκεχυμένη δομή, η διφορούμενη σημασία της πληροφορίας που εμπεριέχουν, καθώς επίσης και η εγγενής μεροληψία, η οποία μπορεί να σχετίζεται με ζητήματα δημογραφικής, πολιτισμικής, γεωγραφικής, τεχνολογικής ή άλλης φύσεως.

Η συμβολή της παρούσας διατριβής αφορά στον σχεδιασμό και την ανάπτυξη ενός πλαισίου, αποτελούμενου από καινοτόμες μεθόδους και υπολογιστικά εργαλεία, το οποίο καθιστά ικανή τη σύζευξη ανομοιογενών χωρικών δεδομένων, ενώ δύναται να συμβάλλει στην καλύτερη αξιοποίηση από τους πολεοδόμους και τους φορείς χάραξης αστικής πολιτικής των δυνατοτήτων που προσφέρουν τα αναδυόμενα κοινωνικό-χωρικά δεδομένα για την κατανόηση των πόλεων. Η σε βάθος κατανόηση των χωρικών και χρονικών μεταβολών των πόλεων, και ειδικότερα, των δραστηριοτήτων και της κινητικότητας των ανθρώπων, αναμένεται να διαδραματίσει ένα κρίσιμο ρόλο στην αντιμετώπιση των προκλήσεων που σχετίζονται με τη ραγδαία αστικοποίηση. Η προσαρμοστικότητα των μεθόδων και υπολογιστικών εργαλείων που αναπτύσσονται στην παρούσα διατριβή καθιστά ικανή την εφαρμογή τους σε επιστημονικά πεδία πέραν αυτού της χωρικής επιστήμης και ανάλυσης, όπως είναι για παράδειγμα οι υπολογιστικές κοινωνικές επιστήμες (computational social science), η αστική γεωγραφία, η επιστήμη γεωγραφικών πληροφοριών (GIScience), και η μελέτη της ανθρώπινης κινητικότητας (human mobility). Στο μέλλον, η έρευνα θα επικεντρωθεί στην ανάπτυξη πολυ-επίπεδων κοινωνικό-χωρικών μοντέλων (multilayered urban models), καθώς επίσης και στη συγκριτική μελέτη της δυναμικής πολλαπλών χωρικών συστημάτων, τόσο σε αναπτυγμένες όσο και σε αναπτυσσόμενες χώρες, με τη χρήση των υπολογιστικών εργαλείων που αναπτύχθηκαν στα πλαίσια αυτής της διατριβής. Το προτεινόμενο πλαίσιο έχει τη δυνατότητα να ανοίξει τον δρόμο προς μια ποσοτική διερεύνηση της δυναμικής των πόλεων με τη χρήση μιας ευρείας γκάμας πηγών δεδομένων, συμβάλλοντας έτσι στην ανάπτυξη μίας νέας επιστήμης για τις πόλεις.

TOC

TOC

1 Introduction

§ 1.1 Background

§ 1.1.1 From Location Theory to Urban Dynamics

Ever since the establishment of urban planning in the early 20th century, and increasingly after its institutionalization from the mid-1950s onwards, there has been growing interest in quantitative approaches to urban phenomena (Michael Batty, 2013b; Hall, 1988). Methods and tools were incrementally imported from social sciences, such as urban and quantitative geography, as well as economics, social physics, and mathematics, among others, and integrated into urban studies. The goal was to support planners to effectively comprehend and subsequently tackle the challenges facing cities. These challenges were traditionally pertinent to issues of land-use location and allocation, population distribution, economic growth, as well as to the improvement of the poor quality of life – an aftereffect of the first industrial city. In the early days of planning, these phenomena were approached in a rather static fashion (Hall, 1988). Forasmuch as changes in the urban environment were insignificant, dynamic parameters were largely disregarded. Therefore, the primary concern was about the physical structure of cities and the location of agglomerations (Michael Batty, 2013b). Location theory provided the main theoretical pillars and models about land uses (Alonso, 1964; Thünen, 1966), the distribution of central places (Christaller, 1933), and the location of industries (Weber, 1909). Although the concept of "urban dynamics" has been introduced already in the late sixties by (Forrester, 1969), focusing primarily on economic interactions and the growth of urban systems, the scarcity of frequently updated data posed challenges to the advancement of the field. However, as the demographic, social, and economic dynamics of cities were growing in unprecedented rates, the necessity for systematic planning support tools and techniques became ever more quintessential.

In cities of the 21st century, rapid urbanization processes have resulted in a demographic outburst that has not been experienced hitherto, in both developed and developing countries (Townsend, 2013; UNFPA, 2007). The subsequent tip in the ratio of urban-to-rural global populations poses additional challenges not only to

cities, but also to planners and policy makers. To a certain extent, these challenges are not different in nature when compared to the age-old issues of urban planning. Economic growth, social segregation, human migration, accessibility to services, aging populations, and transportation optimization, among others, continue to constitute the key problems. In addition, emerging urban issues come to the fore, relating to energy consumption, environmental sustainability, shrinking cities, and de-industrialization. Present-day planners and policy makers are, thus, confronted with a multiplicity of complex urban problems that change rapidly over time and, therefore, need to address demands in shorter time spans than in the past decades. Subsequently, the primary concern is not only on the physical structure of cities, but also on the dynamic interactions between the various components that comprise urban systems.

Following up on the previous observations, a contemporary notion of urban dynamics does not only concern economic interactions and growth processes of cities, but also pertains to human mobility, flows of individuals and goods, and the distribution of social activity over space and time. Therefore, it simultaneously addresses spatial, social, and temporal aspects of the urban environment. Understanding the dynamics of human activity in cities is essential to urban planning, policy making, and transportation planning. Quantitative measures of flows and the distribution of social activity over space and time have potential to facilitate the characterization of urban areas and the development of urban models (e.g. land-use transportation models, mobility models etc.) to simulate and, possibly, predict the use of urban space by individuals. In achieving this, conventional urban data such as census records and travel surveys, though reliable and accurate, have limited capacities to give insights into the spatiotemporal dynamics of cities, primarily due to their infrequent update rate. Therefore, emerging sources such as sensors, mobile phones, and social media could be used as proxies for human activity and mobility dynamics, in combination with traditional sources of urban data.

§ 1.1.2 Emerging Data Sources as Proxies for Urban Dynamics

The systematic use of urban data to understand morphological and functional aspects of cities has its roots already in the late fifties, when quantitative methods started to be applied to urban and regional studies (Kwan & Schwanen, 2009). Early attempts in modeling cities using mathematical abstractions employed data from population censuses, individual or household travel surveys, and economic surveys, which comprised the only available sources to calibrate model parameters. Drawing on location theory, one of the first systems to be mathematically represented were transportation networks and their relation to land uses (Michael Batty, 2009; Forrester,

1969; Hunt, Kriger, & Miller, 2005; Lowry, 1965). More recent models addressed spatial interactions (e.g. gravity models, spatial interaction models etc.), and the correlation between vegetation and urban coverage (e.g. *Land Use/Land Cover Change models (LUCC)*) (Michael Batty, 2007; Button, Haynes, Stopher, & Hensher, 2004; Fotheringham, Brunsdon, & Charlton, 2000). Gradually, dynamic simulation methods were developed, based on Cellular Automata (CAs), Agent-Based Models (ABM), and Multi-Agent Systems (MAS) (Michael Batty, 2009; Michael Batty & Torrens, 2005). Overall, the aforementioned attempts constituted the fundamental means to simulate and quantitatively assess the way cities function. However, the infrequent update rate and – especially in regards to surveys – the relatively small sample, due to high costs, in combination with the short time span covered by the data, posed significant constraints for the exploration of dynamic urban phenomena at the disaggregate level.

In recent years though, the increasing penetration of sensor resources (e.g. GPS trackers, RFID cards etc.) that are embedded in physical space or in handheld devices (i.e. cellphones), in combination with geo-enabled social media (e.g. Twitter, Instagram etc.) and location-based social networks (LBSNs) (e.g. Foursquare) provides an emerging set of data sources about cities. The majority of data that are generated from these new sources are tagged to space and time, have frequent update rate, therefore addressing short time spans, and allow for disaggregation at the level of individual location or person. In addition, data derived from geo-enabled social media and LBSNs are further enriched with human-generated – mainly textual – information, from which topical attributes of social activity (e.g. type of activity, sentiments etc.) may be extracted (Ciuccarelli, Lupi, & Simeone, 2014; Noulas, Scellato, Mascolo, & Pontil, 2011; Psyllidis, Bozzon, Bocconi, & Bolivar, 2015a; Quercia & Sáez-Trumper, 2014).

Although the data generated from these sources usually serve different purposes than those pertaining traditionally to urban and regional studies, they have potential to be used as proxies for the study of urban phenomena. More specifically, owing to the inclusion of spatial, social, and temporal dimensions, they offer new possibilities to the exploration of urban dynamics. This thesis defines the data produced from emerging sources (i.e. sensors, mobile phones, geo-enabled social media, and LBSNs) as *social urban data*. That is, data for cities that are spatially and temporally referenced, are generated either directly from people (e.g. social media data) or indirectly from people's actions (e.g. RFID data, call detail records etc.) and, as such, are less structured and more semantically ambiguous than traditional urban data. Further, this thesis refrains from using the term "big data" since it is lacking a clear definition, it is mostly focused on the volume of data, and it also constitutes a generic concept that is insufficient when it comes to addressing the inherent diversities of emerging urban data types. To some extent, social urban data may qualify as "big data" for cities, especially in regards to volume, but could also refer to relatively "small" data.

There is already a wealth of research employing different social urban data to infer human mobility behavior, inter- and intra-urban flows, and patterns of social activity over time and space (see Sect. 6.1). The methods used in these studies strongly deviate from traditional approaches to urban analysis, in order to cater to the distinctive characteristics of social urban data. This research focuses particularly on devising new computational methods and tools for the integration of heterogeneous social urban data into the analysis of urban dynamics.

§ 1.2 Problem Statement

The diversities between the various types of sources producing social urban data determine respectively the suitability of each source in addressing specific aspects of the urban environment. It is therefore important that these diversities are understood prior to incorporating social urban data into urban analytics. A framework focusing specifically on the distinguishing characteristics of social urban data is currently lacking.

For instance, sensor data such as GPS tracks and RFID records are generally characterized by high spatial and temporal resolution, are structured by a database schema, and have frequent update rate. On the downside, they lack any additional semantics on the socio-demographic attributes of the persons who produce them or the type of activity performed.

Similar limitations apply to call detail records (CDRs) from mobile phones. These particular characteristics make both sensor data and CDRs suitable for use in the analysis of human mobility – inter alia (Amini, Kung, Kang, Sobolevsky, & Ratti, 2014; Bazzani, Giorgini, Rambaldi, Gallotti, & Giovannini, 2010; Calabrese, Diao, Di Lorenzo, Ferreira, & Ratti, 2013; Gonzalez, Hidalgo, & Barabasi, 2008; Roth, Kang, Batty, & Barthelemy, 2011; Zhong et al., 2016) – and less appropriate for the study of social activity.

Conversely, data from geo-enabled social media and LBSNs are often tagged to specific venues in the city (e.g. restaurants, theaters, museums etc.) and are also accompanied by human-generated – primarily textual – content, which adds a certain level of semantic richness. However, the fact that these types of data can be generated by virtually everyone, deviating from the centralized and prescriptive rationale of database management systems (DBMSs) that structure data by a predefined schema, results in "noisy" data streams that could hinder the process of extracting meaningful knowledge from them.

Drawing on the above considerations, it is clear that the employment of social urban data deriving from a single source may yield biased conclusions and result in a fragmented understanding of the complexity characterizing the dynamics of cities. Therefore, it is necessary to combine different types of social urban data together and also to integrate them with the more reliable traditional urban data to mitigate the structural and semantic ambiguities of the former and, eventually, derive richer descriptions of the urban environment. This challenge calls for new approaches that deviate from the traditional ones used hitherto in urban analytics and are able to grapple with the dynamic nature of emerging sources of urban data. Although the current capabilities of computational systems allow the storage, processing, analysis, and visualization of large-scale data, *integration* remains a challenge. Moreover, the majority of studies on human mobility and activity patterns that make use of emerging data types, usually employ only one data source (Helbich, Arsanjani, & Leitner, 2015; Lenormand et al., 2014). In order for urban planners and policy makers to capitalize on the new possibilities given by social urban data, it is important to design new methods and develop tools that enable the integration of data from multiple sources to extract knowledge about the spatiotemporal dynamics of cities.

§ 1.3 Research Aim, Objectives and Scope

In addressing the above-mentioned challenge, the aim of this research is to design a framework of novel methods and tools for the integration, visualization, and exploratory analysis of large-scale and heterogeneous social urban data to facilitate the understanding of human activity dynamics in cities.

In association with this aim, the research has the following four objectives. The first objective is to investigate the distinguishing characteristics of social urban data and, subsequently, identify potential and challenges of these data in the analysis of urban dynamics. The second objective is to explore ways to create interoperable urban data from different sources and, drawing on these, to design methods and develop tools for data integration and interlinkage that could facilitate urban planners, researchers, and policy makers to extract meaningful knowledge from multiple datasets. The third objective is to investigate methods for the derivation of multidimensional (i.e. spatial, social, functional) attributes of people and places from social urban data, and explore how these could enrich metrics of human movement and activity. The fourth objective is to design easily accessible computational tools that incorporate all the aforementioned methods, allow for data integration, and facilitate the interactive exploration of urban dynamics.

The scope of this research focuses primarily on the dynamics of human activity in cities and to a lesser extent on human mobility, as inferred from different sources of social urban data. Therefore, the majority of socio-demographic attributes of individuals, as well as the types of activity they perform are extracted (and inferred) primarily from online social media. For the purposes of this study, the dynamics of human activity and mobility are investigated at the intra-urban level (see Chapter 6). However, the proposed framework can also be used for the exploratory analysis of human activity and mobility at the inter-urban level, or in remote urban systems simultaneously (see Chapter 6).

With regard to data, the research focuses predominantly on data deriving from various geo-enabled social media (i.e. Twitter and Instagram) and LBSNs (i.e. Foursquare), which offer public APIs. In addition, official large-scale data coming from publicly accessible governmental repositories are also used to illustrate examples of data integration. Moreover, for the purposes of data interlinkage, links are established only with external datasets that are already published on the Linked Open Data cloud (LOD cloud) (Schmachtenberg, Bizer, & Paulheim, 2014) and can be publicly retrieved and queried via dedicated endpoints. Therefore, the research does not incorporate data from sensor networks and mobile phones (i.e. CDRs), as these are usually stored in proprietary repositories and, hence, could not be acquired. However, the tools comprising the proposed framework can be adapted to also accommodate and integrate data from sources beyond the ones used in this thesis.

§ 1.4 Research Questions

In relation to the challenge addressed by this research, the main research question is:

How to integrate heterogeneous and multidimensional social urban data into the analysis of human activity dynamics in cities?

To answer this overarching question, the research further addresses five sub-questions in association with its objectives:

1 *What are the characteristics that distinguish emerging social urban data from traditional ones?*
Social urban data are only recently used in research around issues of human mobility and activity. Their potential is scarcely exploited by the urban planning practice and governance to date. Therefore, it is important to juxtapose emerging sources of urban data with traditional ones to identify the strengths and weaknesses of these new sources, prior to using them as proxies for the study of dynamic urban phenomena.

2 *How to transform heterogeneous data for cities into multidimensional linked urban data?*
After identifying the strengths and weaknesses of both traditional and emerging sources of urban data in relation to city dynamics, ways to create interoperable urban data are explored. Merging together data that are characterized by heterogeneous data formats, schemas, structure, resolutions, and naming conventions remains a challenge. Although there exist generic methods for data integration, there is a lack of a domain-specific methodology with a focus on urban analytics that further allows for semi-automatic data integration. Therefore, domain-oriented methods for transforming urban data from multiple sources in to linked urban data, need to be designed.

3 *How could urban planners, researchers, and policy makers leverage the potential of multidimensional linked data in city analytics?*
The adoption of integrated and linked data in urban planning research and practice is currently limited. User interfaces and visualizations could facilitate and potentially increase the consumption and employment of linked urban data in the study of cities. To further foster engagement, the user interfaces and visualizations need to allow for easy access, along with several possibilities for navigation and data filtering, especially when it comes to handling large-scale linked urban data.

4 *What types of attributes can be derived from social urban data in relation to the dynamics of human activity?*
Social urban data are multidimensional in nature, meaning that they are tagged to space and time and, further, contain additional information which could be used to infer attributes of the individuals who produce them. These could relate to a person's demographic characteristics, social ties, preferred places to socialize, type of activities performed in these places, and sentiments about particular activities. The extraction of these attributes is crucial for the analysis of human activity, its distribution over geographic space, and its evolution over time.

5 *How do different sources of social urban data influence the understanding of the spatiotemporal dynamics of human activity in cities?*
Although there exists a wealth of studies on urban dynamics employing emerging data types from single or – less often – multiple sources, the inherent diversities of social urban data and how these may influence the interpretation of the results are usually overlooked. Moreover, domain-oriented tools or platforms that simultaneously enable the collection, integration, visualization, and exploratory analysis of heterogeneous urban data are scarce. In order to create awareness about the dynamics of human activity, such tools need to address the diversities of social urban data and further mitigate their structural and semantic ambiguities. In designing and developing tools to support these processes, easy access and use, along with different types of visualization catering to the particularities of each data source and/or the issue in

question, also need to be considered. Moreover, the tools are necessary to be tested in real-world use cases to not only assess their capacities, but also evaluate the knowledge gained from different sources of social urban data.

§ 1.5 Research Design: Approach and Methods

In order to answer the research questions and to address the main aim and challenge, a research design is developed. The latter is organized into five main parts, each one corresponding to one of the five sub-questions formulated in the previous section. The overarching actions undertaken in each part are namely: (1) definition of social urban data, (2) design of integration and interlinkage methods, (3) design and implementation of linked data visualization tools, (4) exploration of methods for attributes extraction, and (5) design of a system for the analysis of human activity dynamics. Combined, these five parts provide the answer to the main research question.

The research follows a mixture of both deductive and inductive approaches, depending on the actions undertaken in each part of the research design. In particular, the design sequence first involves *exploration* of specific components of the proposed framework (i.e. social urban data characteristics, integration and interlinkage, attributes extraction), followed by the *design* of methods and tools (i.e. for urban data integration, linked urban data visualization, visual exploration of urban dynamics) that fill in gaps identified in the exploration phases. This sequence is not necessarily linear, meaning that not all outputs of each part are required in the next part. Instead, the exploration and design phases are organized around the two main components of the research aim, i.e. data integration and analysis of human activity dynamics, and may therefore overlap. For instance, the extraction of attributes from social urban data does not necessarily follow the integration phase, but could instead be extracted from individual sources. The following paragraphs provide an overview of the methods applied to each part of the research design, in order to address the corresponding research questions (Figure 1).

In the first part of the research design, the concept of "social urban data" is introduced and *defined*, to encompass data for cities that are generated by emerging sources, and their distinguishing characteristics are described. Existing literature on generic "big data" has already recognized some characteristics that are, however, considered typical of only large-scale datasets (Kitchin, 2014a; Mayer-Schönberger & Cukier, 2013; Zikopoulos, Eaton, Roos, Deutsch, & Lapis, 2012). Instead, in this research, emerging sources of urban data are juxtaposed with traditional ones, to explore the

extent to which each of the identified characteristics typifies a certain data type or source. To further define the opportunities and challenges of social urban data in the analysis of urban dynamics, existing literature is reviewed. The identified strengths and weaknesses (i.e. the main output of this part) are used as a general basis for the design of the various methods and tools proposed by this research.

The second part of the research design explores first the heterogeneities of urban data at different levels (i.e. syntactic, schematic, and semantic), and further investigates general approaches to data interoperability from the perspective of ontology engineering (Gómez-Pérez, Fernández-López, & Corcho, 2004), semantic web (Berners-Lee, Hendler, & Lassila, 2001), and linked data (Berners-Lee, 2006). Driven by the current lack of domain-oriented frameworks for data integration, a methodology for urban data integration and interlinkage is *designed*. The methodology follows an ontology-based data integration approach and accommodates a variety of semantic (web) and linked data technologies. Overall, the methodology covers issues of urban data integration, linked urban data generation, and publication to the LOD cloud for further exploitation in urban analysis.

To demonstrate the applicability of the proposed methodology to urban data, a real-world use case is presented, covering all three parts of the methodology. To conduct the use case, nine large-scale spatiotemporal datasets are collected from multiple sources, in particular, three public transportation organizations. As part of the data integration process, an ontology of public transportation networks is also designed and implemented. The resulting integrated dataset is further linked with external resources to provide richer descriptions of the source data, and is eventually published to the LOD cloud.

The third part comprises the *design* and *implementation* of tools that could foster the adoption of (parts of) the methodology proposed in the previous part by urban planners, researchers, and policy makers. Existing approaches to and tools for ontology and linked data visualization are first explored. The limitations of related work set the basis and requirements for the design of the proposed tools. In addition, an ontology of urban networks is also developed, to provide a knowledge model (i.e. a framework of formally-described domain concepts and their interrelationships, representing real-world entities in a machine-processable format) of the interactions between the various components comprising cities. This ontology, in combination with the outputs of the previous part (i.e. the linked dataset and the ontology of public transportation entities), are used as benchmark tests for the tools proposed in this part of the research design.

The second and third part of the research design together address the challenge of urban data integration and the generation of multidimensional linked urban data.

The fourth part of the research design first reviews the existing literature on the measurement of the physical urban structure and the modeling of spatial flows. It also reviews recent attempts in integrating (online) social networks into the physical structure of cities. Next, it investigates how these could be enriched with the new possibilities offered by social urban data. In particular, it *explores* and *describes* methods to extract attributes primarily from geo-enabled social media and LBSN data that are pertinent to individuals (e.g. socio-demographic variables, social contacts, trajectories etc.), places (e.g. function), and the interactions between the two (e.g. activity spaces, type of activity etc.). It also addresses how these approximated attributes help measure the functional density and diversity of urban areas, as well as the geographical extents of activity spaces over different periods of time. The outputs of this part are subsequently used in the design of a system for the spatiotemporal analysis of human activity, introduced in the next part.

In the fifth part, a novel system for the visualization and exploratory analysis of human activity dynamics is designed. Besides accommodating the methods and techniques of the previous part, the system also incorporates the integration methodology of the second part, to integrate emerging with traditional urban data. The system is tested in a real-world case study that investigates how different social categories of people appear to use urban space over time, as inferred from different sources of geo-enabled social media. The data used in the case study come from official census records that are integrated into the system, as well as from different geo-enabled social media and LBSNs that are collected by the system modules.

Besides employing the visual exploration capacities of the system, a spatial analysis on the collected data is also undertaken. In particular, global and local spatial autocorrelation analyses are performed to investigate potential clusters (i.e. activity patterns) in the distribution of the social activity of the different groups over space and time. Both analyses are undertaken for 28 different variables of the collected data in total. Moreover, statistical tests are performed to assess the significance of the obtained results.

The fourth and fifth part of the research design together address the challenge of using heterogeneous social urban data as proxies for the analysis of human activity dynamics.

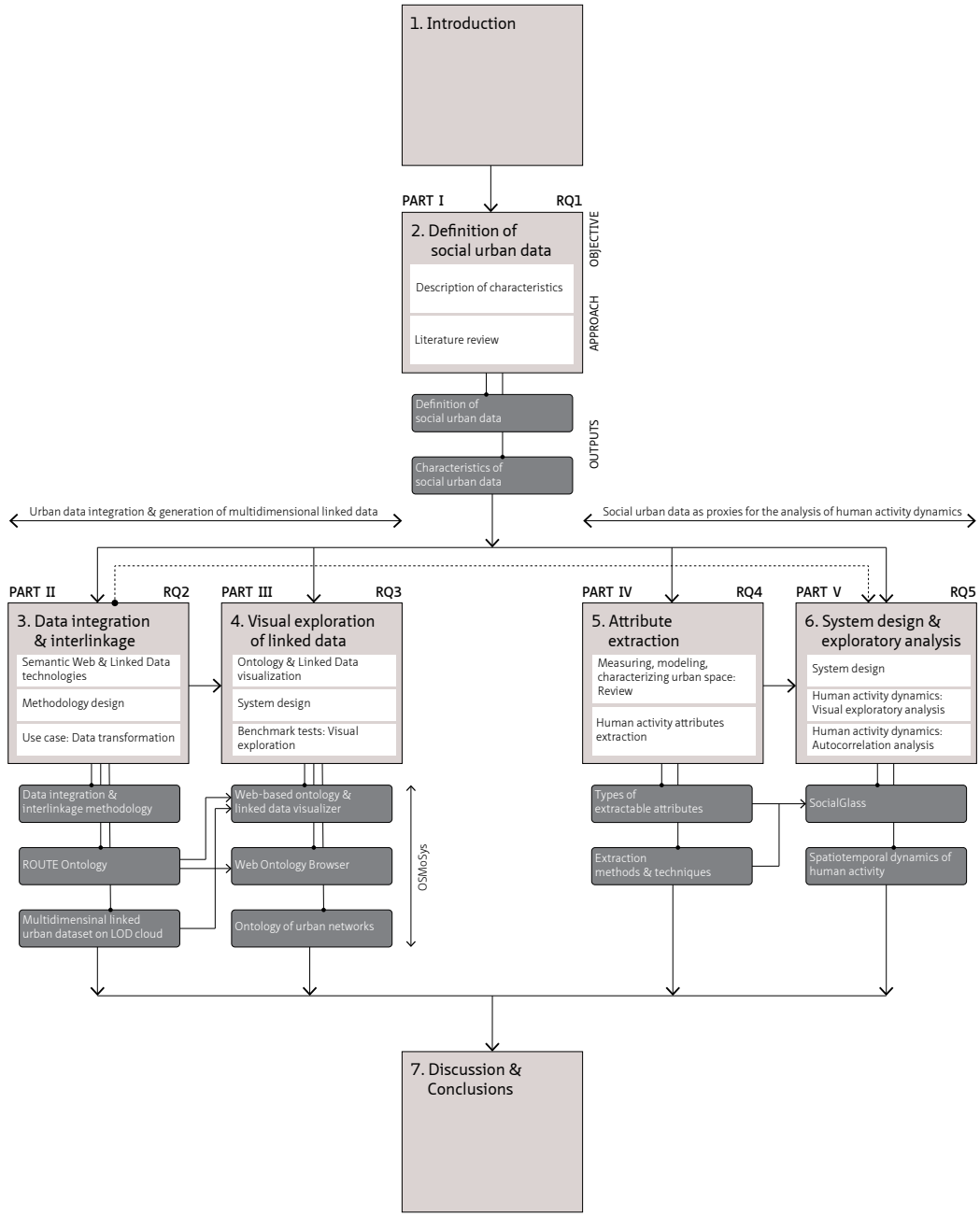


FIGURE 1 Schematic overview of the research design structure and thesis outline.

TOC

§ 1.6 Thesis Outline

Following up on the research design, each one of its five parts is associated with a chapter of the thesis. Accordingly, each chapter addresses one of the five research questions. The answer to the main question is provided in the final chapter (Chapter 7) of this thesis, along with a discussion on the findings of each sub-question.

- Chapter 2 provides a definition of social urban data and their distinguishing characteristics by juxtaposing them with traditional urban data. Drawing on existing literature and various datasets that are collected for the purposes of this research, it defines the strengths and weaknesses of each source and the data it generates, especially with regard to representing aspects of urban dynamics.

- Chapter 3 focuses on the semantic integration of heterogeneous urban data and on processes for transforming them into multidimensional linked urban data. The chapter first explores heterogeneities in urban data and various approaches to data interoperability. Next, it proposes a methodology for semantic integration, linked data generation, and publication to the LOD cloud, oriented specifically for urban analytics. The methodology is demonstrated through a use case, employing real-world datasets from multiple sources. The resulting linked dataset is used for testing the ontology and linked data visualization tools, demonstrated in Chapter 4.

- Chapter 4 presents a set of computational tools to potentially facilitate the adoption of linked data in urban analytics. Prior to presenting the proposed tools, the chapter reviews existing domain ontologies pertinent to cities and planning, as well as approaches to ontology visualization. To semantically enrich data coming from various city sectors or agencies, an upper-level ontology of urban networks is also presented, as a sharable and extendable knowledge model to which different urban data can be mapped. This ontology, along with the outputs of Chapter 3, are used as benchmark tests for the proposed tools.

- Chapter 5 focuses on attributes that can be extracted from various social urban data, in relation to individuals, places, and the interactions between them. First, the literature on the measurement of the physical urban structure is reviewed. Next, the chapter reviews measures and models of spatial interactions, as well as the recent work on the integration of (online) social networks into the physical structure of cities. Following up on these reviews, it explores new methods and techniques to extract attributes from social urban data, to subsequently support the characterization of urban areas and social activities. Metrics of individual trajectories, activity density and diversity, are further revisited. The methods and techniques presented in this chapter set the foundation for the design of the web-based system for the exploratory analysis of human activity dynamics that is presented in the following chapter.

– Chapter 6 presents the design of a novel web-based system that enables the exploratory analysis of human activity dynamics in cities, by integrating emerging with traditional urban data from multiple sources. The chapter first presents the rationale of the system as regards the approximation of attributes of social activity in urban space, based on the outputs of Chapter 5. Following up on this, it presents the various components and modules that comprise the system architecture and cater to data ingestion and analysis, semantic enrichment and integration, visualization and exploratory analysis, and real-time monitoring of urban dynamics. An instance of the system is put to use in a real-world case study to assess the potential and limitations of the proposed system, and to also investigate how different sources of social urban data could influence the understanding of urban dynamics.

– Chapter 7 discusses the findings of each chapter by revisiting the research questions formulated in the Introduction chapter. The limitations of the research are also highlighted. In addition, the overall conclusions are presented, first by answering the main research question and, then, by summarizing the major findings. Afterwards, the chapter presents potential application to practice and research and, finally, concludes with pointers to future research.

TOC

2 Defining the Characteristics of Social Urban Data

§ 2.1 Introduction

The key role of urban data in the study of cities is acknowledged already since the 1950s, when quantitative (i.e. statistical) methods for spatial analysis began gaining in popularity. In recent years, a considerable interest has been stimulated with regard to data-driven approaches to urban analytics, due to the increasing availability of new sources generating data about cities (Michael Batty, 2012, 2015; Birkin, 2009; Fischer, 2006; Fotheringham et al., 2000; Helbich et al., 2015; Kitchin, 2015; Kwan & Schwanen, 2009; Miller & Goodchild, 2014; Solecki, Seto, & Marcotullio, 2013). Recent literature specifically calls attention to the shift from a data-poor to a data-rich environment and the implications this has for spatial studies (Michael Batty, 2013c; Kitchin, 2013, 2014a; Miller, 2010; Miller & Goodchild, 2014). The prevalent concept used to define the emerging deluge of data sources is that of "big data". However, the lack of a shared definition of the concept (Kitchin, 2014a), in combination with its broad and generic nature, and its primary focus on volume or size could be rather misleading in the particular context of urban data. For instance, enterprise or biological data may well qualify as "big data", but are distinguished by entirely different characteristics than those suitable to address aspects of urban systems. Therefore, a new concept needs to be defined that focuses specifically on emerging data about cities and, by determining the boundaries of scope, to further help define opportunities and challenges with regard to urban analytics.

To address this gap, the concept of *social urban data* is defined in this chapter. Drawing on this definition, the characteristics of social urban data are further described. Existing literature to date has elaborated on the characteristics of emerging data sources, yet attributing them only to large-scale datasets. Instead, in this chapter, emerging sources of urban data are juxtaposed with traditional ones, either "big" or "small", to describe the extent to which each of the defined characteristics typifies a certain data type or source. Existing research employing different types of emerging urban data is reviewed to further describe the potential and limitations of social urban data in the analysis of urban dynamics.

The chapter is structured as follows. First, the definition of social urban data is provided. Next, a model is introduced to classify (social) urban data according to the source that generates them. Afterwards, the characteristics of social urban data are defined. Based on these characteristics, social urban data are compared to traditional ones to investigate the strengths and weaknesses of each source, with regard to the analysis of urban dynamics. Finally, Sect. 2.5 summarizes the conclusions.

§ 2.2 Defining Social Urban Data

Prior to defining the characteristics of social urban data, it is necessary to first define the concept's scope. From the term itself, two distinctive, yet interrelated, components are recognized, namely: the *social* and the *urban*. The following paragraphs describe the scope of each component individually, to then provide the overall definition of the concept.

The *social* component implies that the data are generated by people, either directly (explicit data generation) – as is e.g. the case with social media data – or indirectly (implicit data generation) through people's actions, such as via mobile phone activity (captured in CDRs), tap ins/outs of the public transport systems (captured in RFID card records), exchange of e-mails, GPS traces, among others (Bocconi, Bozzon, Psyllidis, Bolivar, & Houben, 2015; Weigend, 2009).

The primary emerging sources of explicitly-generated social data are geo-enabled social media (e.g. Twitter, Instagram, Flickr, Sina Weibo) and LBSN platforms (e.g. Foursquare). Data derived from these sources are usually enriched semantically with content, in the form of short texts (i.e. microposts), images, or videos that could in turn reflect activities, feelings, opinions, or experiences of the people who generated them. However, the interpretation of the semantics yields ambiguities that are pertinent to a variety of contextual, cultural, technological, and demographic biases. The content posted on social media is invaria bly tagged to time and, to a lesser extent, to space (either with exact geo-coordinates or in relation to a specific venue or region). People could thus act both as "human sensors" (Goodchild, 2007) who organically create data about their activities and as interpreters who generate social content on demand (e.g. through crowdsourcing) (Boulos et al., 2011). Further, the majority of these platforms enable people to develop online networks of social contacts (i.e. online "friendships"), from which social relationships could be inferred. Conversely, implicitly-generated social data are sourced from sensors, such as GPS trackers, RFID card readers, and cameras, as well as from mobile phones. Data generated from these sources are tagged to both space and time, are generally well-structured, but usually lack the semantic expressiveness characterizing data from online social media.

The *urban* component indicates that the datasets under consideration pertain to cities. This further entails that they are not only *for* cities, but are also generated *in* specific urban settings. Urban data are necessarily geo-referenced, which means that they are tagged to space, indicating certain locations in cities. However, as cities are not only about locations, but are also about interactions between these locations (Michael Batty, 2013b) that evolve dynamically over time, the urban data considered here are also tagged to time.

By combining the two components together, *social urban data* refer to data for cities that:

- are generated either directly or indirectly from people and their actions;
- derive from emerging sources such as sensors, mobile phones, geo-enabled social media, and LBSNs;
- are multidimensional in nature, meaning that they are spatially and temporally referenced;
- can be used to infer spatial, temporal, and social aspects of human movement, activity, and social connectivity;
- but are less structured and more semantically ambiguous than traditional urban data (Figure 2).

§ 2.3 Classifying Data for Cities

The multiplicity of both traditional and emerging sources from which urban data can be sourced, calls for a classification according to the source type. To this end, this research adopts the tripartite information model introduced by (Devlin, 2013), which classifies data into three general categories: *process-mediated*, *machine-generated*, and *human-sourced*. Though Devlin's model is originally created for business data, the general categories it identifies are quite generic and can also be used to classify (social) urban data (Table 1).

By adapting the model to the scope of this research, *process-mediated* data refer primarily to authoritative data for cities that are traditionally sourced from population censuses, individual or household travel surveys, land-use diagrams, and real-estate records, to name but a few.

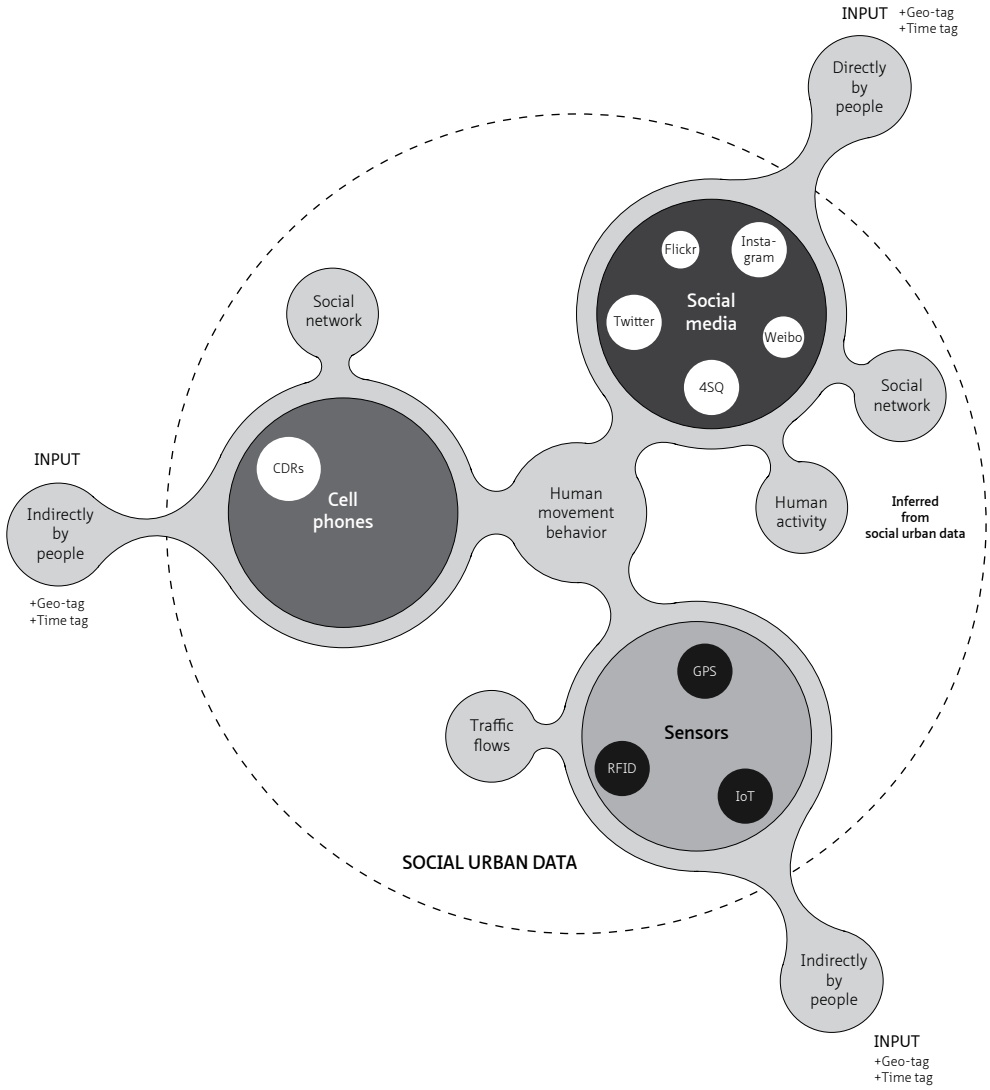


FIGURE 2 Schematic representation of social urban data.

Machine-generated data encompass measurements and observations coming from a growing amount of individual sensing devices and sensor networks that are distributed across the urban fabric. This category also refers to call detail records (CDRs) that are sourced from mobile phone activity.

Finally, the category of *human-sourced* data pertains to human-generated content (HGC) that is sourced from various geo-enabled social media and LBSNs, such as micro-blogging posts (e.g. Twitter, Sina Weibo), pictures (e.g. Flickr, Instagram), and videos (e.g. YouTube), among others.

TOC

TABLE 1 Categories of (social) urban data, major data types and sources, following the classification of (Devlin, 2013).

Data category	Type of (social) urban data	Source
Process-mediated	Census records Migration/traveling data Land uses Real-estate records	Census bureaus Household/individual travel surveys, tourism bureaus, airline statistics Planning organizations Housing organizations
Machine-generated	RFID records GPS traces Sensor records (e.g. traffic counts, pedestrian counts CDRs	RFID devices GPS trackers Urban sensing devices Cell phones
Human-sourced	Geo-tagged social media content (e.g. posts, pictures, videos, check-ins etc.) E-mails Crowd-sourced data	Twitter, Instagram, Sina Weibo, Foursquare, Flickr E-mail providers Crowd-sourcing platforms (e.g. Mechanical Turk)

§ 2.4 Defining the Characteristics

Although social urban data are increasingly employed by researchers in performing urban analytics, there is a lack of understanding as to what their strengths and weaknesses (i.e. biases) of each source are, in relation to the datasets that are traditionally used in the study of cities. To define the characteristics of social urban data, emerging sources – as described in the definition given in Sect. 2.2 – are juxtaposed with traditional ones, following Devlin's classification (see Sect. 2.3), to investigate the extent to which each characteristic typifies a specific source. An additional objective of this juxtaposition is to identify the strengths and weaknesses of each data type – focusing on one characteristic at a time – in understanding how cities function over time. Drawing on existing literature and research employing different types of emerging and traditional urban data to address aspects of urban systems and dynamics, the following paragraphs define eight characteristics and examine what distinguishes emerging social urban data from traditional ones.

The eight characteristics that are defined here, are namely: *diversity*, *scale*, *timeliness*, *structure*, *spatiotemporal resolution*, *semantic expressiveness*, *representativeness*, and *veracity*.

§ 2.4.1 Diversity

The diversity of urban data is a multifaceted characteristic, in particular, of a threefold nature. In the first place, it addresses the different sources that presently generate data for cities. In the second place, it refers to the variations in terms of format, quality, resolution, structure, and semantics. In the third place, pertaining specifically to geo-enabled social media and LBSNs, diversity refers to the different demographics of people who contribute content to these platforms (see also Chapter 6).

With regard to the first facet of diversity, the current availability of multiple sources from which data about cities can be derived is both an opportunity and a challenge for urban planning and policy-making. The opportunity lies in the possibility to employ data from different sources simultaneously. Although this could help overcome the bias of a single source (e.g. incompleteness, small sample coverage, lack of time tags etc.) and potentially provide richer descriptions of urban systems and human behavior, it also raises significant problems of interoperability and integration. These problems relate to the second facet of diversity, i.e. the different file formats, resolution, structure, naming conventions, that hinder the fusion of data from multiple sources (see also Chapter 3).

Traditional *process-mediated* urban data comprise many different types and continue to be the most reliable sources of information with regard to spatial analytics and planning (Michael Batty, 2013c). They are essentially characterized by their relatively high quality in terms of accuracy, completeness, validity, and general truthfulness of the content (Psyllidis et al., 2015a). Although the majority of these data increasingly come in structured digital formats (e.g. shapefiles, tabular data in the form of logistic sheets or comma separated values etc.), there still exist several authoritative data in formats that are not machine-readable (e.g. images, raster maps etc.). Diversity is further intensified by the existence of disparate data silos, in which the majority of authoritative urban data are stored. The use of different naming conventions, models, and schemas across sectors poses several challenges to interoperability (Métral, Falquet, & Cutting-Decelle, 2009; Psyllidis, 2015).

Following up on the diversities of traditional (process-mediated) data, the role of diversity in *social urban data* is examined.

At present, sensors and sensor networks are embedded in the physical infrastructure of cities (e.g. road networks, water management systems, energy consumption monitoring systems, street lighting, smart grids etc.), producing a proliferating amount of *machine-generated* data (Verdone, Dardari, Mazzini, & Conti, 2008; Vinyals, Rodriguez-Aguilar, & Cerquides, 2008). The streams of data they generate (i.e. measurements and observations) provide information about pedestrian movement

and transport flows, environmental conditions, air quality, electricity usage, weather, and sound levels, to name but a few (Fernandez, Marsa-Maestre, Velasco, & Alarcos, 2013; Psyllidis & Biloria, 2014). Therefore, they constitute invaluable sources of information, in regards to the inner-workings of city infrastructure. However, the diversity of sensing devices leads subsequently to the generation of heterogeneous measurements, observations, and data representations, which pose challenges to interoperability and compatibility (Compton et al., 2012).

Aside from distributed and embedded sensing devices, machine-generated data further encompass call detail records (CDRs), sourced from mobile phone devices. Though depending on the mobile phone provider, these records usually come in tabular formats, containing attributes about the type of activity (i.e. phone call, SMS etc.), caller and receiver IDs (encoded), caller and receiver location, timestamp, and total duration (Blondel et al., 2012). Therefore, data from mobile phones do not present significant diversities. Yet, the nature of the attributes they contain enables them to be used as proxies for individual human movement (Calabrese et al., 2013; Gonzalez et al., 2008), mobility patterns in and across cities (Amini et al., 2014; Grauwin, Sobolevsky, Moritz, Gódor, & Ratti, 2015; Kang, Ma, Tong, & Liu, 2012), and human activity (Diao, Zhu, Ferreira, & Ratti, 2015; Wang, Kang, Bettencourt, Liu, & Andris, 2015).

Human-generated data derived from geo-enabled social media (e.g. Twitter, Instagram, Flickr, Sina Weibo) and LBSNs (e.g. Foursquare), constitute increasingly important sources of information about the interaction of people with the urban environment (Balduini, Bozzon, Valle, Huang, & Houben, 2014; Cranshaw, Schwartz, Hong, & Sadeh, 2012; McKenzie, Janowicz, Gao, & Gong, 2015; Psyllidis et al., 2015a; Steiger, Westerholt, Resch, & Zipf, 2015). The proliferation of these sources has been made possible firstly through the growing penetration of broadband connection, secondly through the advent of the Social Web (otherwise called Web 2.0), and lastly through the advances in database systems, which currently allow the decentralized production of datasets that do not adhere to particular models or schemas. Data crawled from these particular sources are usually tagged to space (either with exact geo-coordinates or in relation to a specific venue) and time, and are further enriched with semantic information about the users who generate them (e.g. age, gender, online social contacts etc.), the places of activity (e.g. geo-location, category, popularity etc.), and the type of activity performed (e.g. topical information extracted from microposts, sentiments etc.) (Jiang, Alves, Rodrigues, Ferreira, & Pereira, 2015; Noulas et al., 2011; Psyllidis et al., 2015a; Quercia & Sáez-Trumper, 2014; Shelton, Poorthuis, & Zook, 2015). Owing to these characteristics, human-generated data have capacity to be used in the study of spatiotemporal urban dynamics.

What particularly distinguishes human-generated data from other types of social urban data, is that diversity does not only refer to differences in format, quality, or structure, but also implies differences in the population demographics across platforms

(Yang, Hauff, Houben, & Bolivar, 2016). These intrinsic diversities may lead to biases of contextual (e.g. type of activity, role of individual users etc.), geographic (e.g. urban or rural locations), cultural (e.g. popularity of social media in different countries), demographic (e.g. age, gender, socio-economic status etc.), or technological (e.g. penetration rate) nature. Although these intrinsic diversities and biases pose several challenges to the interpretation of human-generated social urban data, they also have potential to provide richer descriptions of activities, if accommodated in the study of urban dynamics (see also Chapter 6).

§ 2.4.2 Scale

Scale refers to the size or volume of data. This particular characteristic is widely considered the prevalent one in the contemporary data landscape, especially in existing literature on "Big Data" (Boyd & Crawford, 2012; Dutcher, 2014; Kitchin, 2014b; Laney, 2001; Mayer-Schönberger & Cukier, 2013; McGovern, 2015). Although the current data supply is indeed rather voluminous, the issue of size is definitely not new, especially with regard to urban data (Michael Batty, 2013a; Miller & Goodchild, 2014). In fact, the characterization of data as either large- or small-scale is to a great extent dependent on the processing capacities of the available computational resources. Thereby, what nowadays is considered to be "large-scale data", it could possibly be characterized as rather small-scale in the near future.

The following paragraphs describe the characteristic of scale and investigate the extent to which it constitutes an influencing factor of both traditional and social urban data.

Although the scale of *traditional process-mediated* data is generally considered to be rather small – compared, for instance, to real-time streams from sensors and social media – certain certain types can be rather voluminous (Michael Batty, 2015). For instance, complete population censuses may, in some cases, comprise hundreds of millions of individual records and, therefore, processing with conventional computational techniques could prove to be cumbersome (Miller & Goodchild, 2014). However, as the update rate of population censuses is rather infrequent (usually once every ten years), the dataset will not increase in size, before the next census. Similarly, data about flows of people (e.g. migration, commuting), goods, and information are usually large in scale, yet their size remains unchanged for a long period of time, as travel and migration surveys are conducted rather infrequently.

Scale is a distinguishing characteristic of social urban data, either *machine-* or *human-generated*. Data generated from sensors, mobile phones, and social media usually come in large volumes, are updated in (near) real time and, therefore, require new

methods and tools to be handled, processed, stored, and analyzed. Nevertheless, larger volumes of data do not necessarily imply better quality data (Boyd & Crawford, 2012). In addition, although access to large-scale social urban data is increasingly becoming available, full access to the entire data stream is usually restricted. For example, in the case of mobile phone data, providers usually grant access only to a limited sample of CDRs at a high cost. Similarly, in certain social media platforms such as Twitter, access to the entire set of public microposts (called "firehose") is only allowed at a high cost. Therefore, the majority of Twitter data stem from the publicly-available "Streaming API" that allows access to a limited sample (i.e. approximately 1%) of the entire feed (Morstatter, Pfeffer, Liu, & Carley, 2013). Moreover, access to various social media APIs are generally subject to change, according to changes in the data sharing policies of the social media platforms.

§ 2.4.3 Timeliness

Timeliness refers to the update frequency of a data source. Traditional process-mediated urban data (e.g. population census, travel surveys etc.) are generally characterized by low update rates, ranging from several months to decades. On the contrary, emerging social urban data are transmitted continuously and can further be collected in (near) real time. In general, timeliness closely relates to the characteristic of scale, since the higher the frequency update the higher the volume of data produced.

The high update frequency of emerging social urban data enables observations on shorter time intervals than it has been possible with traditional ones (Michael Batty, 2013a). Therefore, the employment of social urban data in city analytics entails a deviation from the traditional understanding of urban phenomena.

Urban planning, research, and policy-making have hitherto been dealing with long time periods (e.g. years or even decades). This, of course, was a direct consequence of the then available datasets that were updated rather infrequently. In principle, the collection methods of most traditional urban data, which involve questionnaires, surveys, and on-site observations, allow for very limited updates and are rather demanding in term of – human or other – resources. Therefore, traditional urban data become available only every few months, years, or even decades. On the contrary, streaming social urban data open new avenues in understanding how cities function in short time horizons. In some sense, this implies a high degree of immediacy to the analysis and planning of urban space, marking an entirely new condition for planners, researchers, and policy makers (M. Batty et al., 2012).

Provided that real-time social urban data are invariably collected, it could be anticipated that a large amount of data for cities will be available in the future at very fine temporal resolution. This could mark an entirely new condition for urban planning and policy-making, in which different city stakeholders will have at their disposal datasets that would cover a very broad range of timescales, instead of just a limited sample of them.

§ 2.4.4 Structure

The characteristic of structure refers to the way data are stored and organized. Although traditional urban data are generally well-structured, social urban data are characterized by several intrinsic variations. In particular, machine-generated data (i.e. sensor records and CDRs) are generally better-structured and easier to process than human-generated data from social media, as the latter are created spontaneously and, as such, are unstructured.

In general, data can be classified into structured, semi-structured, and unstructured. *Structured* datasets are accompanied by a clearly defined model, which not only describes the different data types (e.g. Booleans, characters, integers, arrays, lists, frames etc.), but also the relationships between them. Further, it determines the way in which they will be stored and accessed. Thereby, such datasets can be better understood, processed, and queried by computing systems. Traditionally, structured datasets are stored and managed in Relational Database Management Systems (RDBMSs – e.g. MySQL, PostgreSQL etc.).

On the other hand, *unstructured* data do not follow specific models and, therefore, require additional handling to be further read and processed by computing systems. The majority of real-time streams generated from social media resemble unstructured data.

At the intersection of the two previous categories lie the *semi-structured* data. The latter do not follow a specifically defined model, but do contain attributes, metadata, and other markups in a structured format (e.g. XML, JSON etc.). For example, an image – which individually constitutes an unstructured data object – accompanied by metadata, such as geo-location (i.e. latitude and longitude coordinates), timestamp, and keywords about its content in a structured format, resembles a semi-structured data object.

Process-mediated authoritative data usually comprise a mixture of the three aforementioned categories. A large number of open municipal data is increasingly becoming available in machine-readable and non-proprietary formats (e.g. CSV). In addition, the majority of transport data are generally well structured, yet different

models may be used depending on each sector's needs and usage purposes. However, there still exists a large number of unstructured authoritative data such as raster maps, aerial photos, and other analog documents.

Data streams from sensing devices and sensor networks are generally characterized by heterogeneous device types, communication protocols, and models (Barnaghi, Wang, Dong, & Wang, 2013). Sensor data are well-structured, accompanied by several attributes and markups pertinent to the specific measurements and observations (Golab & Özsu, 2003). Recent advances in relational data stream management systems (RDSMSs) and NoSQL databases (e.g. MongoDB, Apache Cassandra, HBase etc.) have played an important role in enabling the storage and performance of different operations on streaming data.

On the other hand, human-generated data from social media are highly unstructured, consisting of microposts expressed in natural language (e.g. Twitter, Sina Weibo), pictures (e.g. Instagram, Flickr), videos or sound files. The structural heterogeneities characterizing social urban data forms significant barriers to interoperability and, therefore, to integration and interlinkage (see Chapter 3).

§ 2.4.5 Spatiotemporal Resolution

Drawing on what has been discussed already, emerging social urban data cover a wide range of spatial and temporal scales. Different data sources are characterized by diverse spatiotemporal granularity and, therefore, the data they produce are at different levels of detail (LoD). For instance, a geo-referenced record referring to the location of a specific venue in a city may include, besides the latitude and longitude coordinates, metadata about its address, which in turn contains attributes such as street name, number, floor, postal code, area, city, state, country, and others. If the aforementioned attributes constitute individual fields – instead of being aggregated into a single field (e.g. column) – the respective data object is considered fine-grained.

Typically, traditional urban data are aggregated into predetermined spatial units such as districts, neighborhoods, or municipalities. These aggregations into arbitrary spatial divisions unavoidably lead to the long-lasting issue of spatial analysis, called the Modifiable Areal Unit Problem (MAUP). MAUP refers to the interpretation biases that could result from the aggregation of spatial data into various predefined districts (modifiable areal units) (Openshaw, 1984). Changing the scale of the districts (i.e. the size of areal units) will automatically change the data aggregation into them, which will eventually lead to a different spatial distribution of the studied variables, thus influencing the observations and the ways in which urban phenomena are understood.

Although contemporary geo-referenced social urban data allow for spatial analysis at the disaggregate level (e.g. individual locations), MAUP remains a significant challenge when it comes to the characterization of urban areas, where disaggregate data (e.g. point-based observations) need to be aggregated into arbitrary spatial units, determined by the person who performs the analysis (Steiger et al., 2015).

As regards sensor data, the spatial and temporal resolution is dependent on several parameters, such as the device features (i.e. the measurement capacity), the Field of View (in the case of remote sensing data) and the spatial coverage (when it comes to a sensor networks).

In the case of geo-referenced human-generated data from social media the LoD of the metadata differs substantially across platforms. For instance, in Sina Weibo – a popular social media platform to Chinese populations, equivalent to Twitter – enables metadata about the device type that generated the micropost to be crawled (e.g. cellphone type, desktop or laptop type etc.) (Q. Gao, 2013). This high LoD is not found in other popular platforms, such as Twitter, Flickr, and Instagram. However, high spatiotemporal resolution also raises issues of privacy (see Chapter 7).

§ 2.4.6 Semantic Expressiveness

Extracting meaning from social urban data is crucial for understanding the nature of social activities and their dynamics. This could lead to more qualitative insights into how cities function and, also, how they are experienced by people. Thus, besides the spatial and temporal attributes, the semantics of urban data play a significant role, in this regard. In order to (semi)-automatically extract the meaning of (large-scale) data, new specialized methods are required, leveraging semantic (web) technologies, knowledge representation tools, and linked data principles that will further be elaborated in Chapter 3. However, not all social urban data types are equally characterized by semantically rich content. Social urban data are in fact distinguished intrinsically by different degrees of semantic expressiveness.

The majority of traditional urban data are semantic by design. That is because they are created ad hoc and, therefore, contain several accompanying attributes that add reliable contextual information to each data record. For example, household travel surveys contain comprehensive information about trip-making activities at the disaggregate level such as the travel origin and destination, the purpose, the travelling time, as well as demographic and socio-economic attributes of the surveyed individuals or households (Zhong et al., 2015).

On the other hand, data from mobile phones (CDRs) and sensor resources present limited semantic expressiveness. For instance, as it was mentioned in Sect. 2.4.1, the majority of CDRs contain information about the start and end times of a call, the total duration, the geo-locations of callers and receivers, the type of interaction (e.g. phone call, SMS etc.), but they lack any contextual information. The same applies to sensor-generated data. Although there is a wealth of research on deriving information about human mobility and social interactions from mobile phone data (Andris, 2016; Calabrese, Smoreda, Blondel, & Ratti, 2011; Gonzalez et al., 2008; Ratti, Frenchman, Pulselli, & Williams, 2006; Toole, Herrera-Yaque, Schneider, & Gonzalez, 2015; Wang et al., 2015), the semantic quality of both CDRs and sensor data is poor.

In contrast, the content of human-generated data from social media has potential to reveal valuable contextual and topical information about the users and the type of activities they perform (Jiang et al., 2015; Llorente, Garcia-Herranz, Cebrian, & Moro, 2015; McKenzie et al., 2015; Psyllidis et al., 2015a; Quercia & Sáez-Trumper, 2014). Although social media data possess a high degree of semantic expressiveness, the extraction of reliable knowledge about the dynamics of human activity is hampered by several cultural, personal, geographical, technological, demographic, and contextual biases that could result in ambiguous interpretations (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011; Olteanu, Castillo, Diakopoulos, & Aberer, 2015; Yang et al., 2016).

§ 2.4.7 Representativeness

A crucial aspect in the use of social urban data as proxies for urban dynamics is to examine how representative these sources are, with regard to different population groups and their activities. However, the issue of representativeness is not new, as regards urban data in general. In spatial analysis and planning the use of data derived from sample populations is rather common. In the past, the limited storage and processing capacities of early computing systems posed several constraints to the analysis of large-scale urban data. Thereby, sampling strategies were developed in order to handle the data volumes (Miller & Goodchild, 2014). Such strategies are still in use nowadays for the collection of urban data (e.g. surveys, on-site observations etc.). Defining a sample is subject to many parameters that influence how well certain population groups are represented, as well as the extent to which generalized conclusions can be reached.

With the advent of cloud computing systems that allow for distributed storage and processing of large data volumes on computer clusters, it is currently possible to analyze entire populations, or considerably bigger samples than it has been possible hitherto (Mayer-Schönberger & Cukier, 2013). In this way, traditional large-scale

urban data, such as complete censuses, real-estate records, transport data, and datasets about interactions (e.g. migration, interpersonal relations etc.) can be handled more efficiently.

As regards social urban data, the degree of sample coverage is dependent on the source. Sensors or sensor networks cover limited regions in a city or represent specific population groups (e.g. those who use public transport to commute). In the case of mobile phone data, sample coverage depends on the penetration of the provider in question. In existing research that makes use of CDRs to study human mobility and interactions, data usually derive from a single telecom provider, raising question about the representativeness and the generalization of the obtained results (Goodspeed, 2013). In addition, calls made through online platforms (e.g. Skype, Viber, Whatsapp etc.) are hardly ever considered.

Similarly, social media are still used by a limited – yet perpetually increasing – amount of people. As a consequence, the data they generate represent, by definition, a sample of population groups with specific demographics. Usually, they portray younger populations, people from more affluent regions with access to broadband services, or certain ethnicities based on platform popularity (e.g. Twitter is popular in Europe, North America, and Australia, but not in China, where Sina Weibo is used instead) (Hargittai, 2007). As a result, a number of population groups are over-represented while others are completely excluded.

In addition, as mentioned previously, not all social media streams are publicly available for crawling. Platforms such as Facebook, do not offer public APIs, whereas Twitter allows only a small sample of public microposts to be crawled for free, through the "Streaming API". From these publicly available streams, only small amount consists of geo-referenced data. According to (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013) around 1.4% of microposts are tagged with a place indicator and an exact geo-location. Intrinsic demographic diversities across social media platforms (Mislove et al., 2011) and different usage patterns play a significant role in the representativeness of the collected data. The disregard of these diversities and sampling limitations in analyzing urban dynamics, is a major threat to validity and generalization.

§ 2.4.8 Veracity

Veracity relates to the trustworthiness of the information included in the data. In understanding how cities function, traditional process-mediated data are still considered the most reliable sources of information about cities and socio-economic activities (Michael Batty, 2013c). They are generated to serve a specific purpose,

therefore resulting in high quality data. Machine-generated data are also characterized by high levels of reliability, although the accuracy of the data is dependent on the resolution of the sensing device, that is, the precision with which the measurements are made.

Unlike the aforementioned data categories, social media data are rather ambiguous sources of information, owing to their unstructured nature and the several biases they contain. Identifying the degree of veracity in social media data is a challenging issue. Aspects that could affect the trustworthiness of the data are pertinent to language, user behavior, context, periodicity, and personal biases (Derczynski & Bontcheva, 2014). An example of data ambiguity may refer to the collection period of social media data. Large-scale events or celebration taking place in a city could create data anomalies and, therefore, result in false interpretations of the observed activity patterns, if the events are neglected in the data analysis (see also Chapter 6).

To improve the level of veracity in social media data, the strategies of crowdsourcing and human computation are increasingly gaining in popularity. These two approaches rely on the role people can play both as producers and as interpreters of data (Balduini et al., 2014; Boulos et al., 2011; Burke et al., 2006). Whereas people create data on social media in an organic fashion, the information generated through crowdsourcing and human computation procedures is entirely on demand. Crowdsourcing allows general public and professionals – through the use of particular platforms and/or devices – to contribute their own additional information, metadata, and content interpretations (by e.g. annotating maps with comments, photos, videos etc.; by sharing information about weather phenomena; by collectively editing articles, as is the case of Wikipedia etc.). Human computation, on the other hand, relates to requests sent by a particular piece of software to (a group of) people, so as they evaluate, comment, or contribute to an issue in question.

In the context of urban data, in particular, the most prominent and widespread application of crowdsourcing relates to the collective contribution of volunteered geographic information (VGI) to platforms such as Google Earth and OpenStreetMap. However, certain biases are still present to some extent. As is the case with social media, the majority of VGI contributors appear to originate from more affluent regions, that is, places with access to broadband services, higher technology penetration levels etc. Subsequently, these areas are characterized by an abundance of information, whereas in less affluent regions information is still scarce (Goodchild & Li, 2012; Haklay, 2010; Townsend, 2013). A potential way to minimize the level of bias is by evaluating different users on the basis of their contributions and their respective levels of accuracy (Goodchild & Li, 2012; Miller & Goodchild, 2014). Nevertheless, for both crowdsourcing and human computation, irrespective of the techniques applied, the most challenging factor is to provide incentives for user engagement.

TABLE 2 Traditional and emerging social urban data: overall comparison of characteristics.

		Diversity	Scale	Timeliness	Structure	Spatiotemporal resolution	Semantic expressiveness	Representativeness	Veracity
TRADITIONAL URBAN DATA	Process-mediated (e.g. census records, land uses, migration data etc.)	+/−	+/−	—	+	+/−	+	+	+
EMERGING SOCIAL URBAN DATA	Machine-generated (e.g. RFID records, GPS traces, CDRs etc.)	+/−	+	+	+	+	—	+/−	+
	Human-sourced (e.g. social media data)	+	+	+	—	+/−	+	—	—
		+ High	+/− Medium	— Low					

§ 2.5 Summary and Conclusions

This chapter provided a definition of social urban data and described their distinguishing characteristics by juxtaposing them with traditional ones. Further, it investigated their strengths and weaknesses as sources for the analysis of urban dynamics (Table 2), by reviewing existing literature. Arguing that the concept of "big data" is insufficient to address the specificities of emerging data for cities, it introduced the concept of "social urban data" and defined its scope.

Social urban data do not comprise a unified category of data with common characteristics. In fact, according to the source that generates them (i.e. sensors, mobile phones, geo-enabled social media, and LBSNs), they may be characterized by varied levels of diversity, scale, timeliness, structure, spatiotemporal resolution, semantic expressiveness, representativeness, and veracity. However, it is argued that the eight aforementioned characteristics are not only inherent to emerging social urban data, but are also present – to a greater or lesser extent – in traditional data for cities.

The most distinguishing characteristic that differentiates emerging social urban data from traditional ones, is the purpose guiding their generation. Although conventional data for cities are created ad hoc, social urban data are generated organically and serve a variety of purposes. As such, they contain contextual, technological, geographical, demographic, and cultural biases, which in turn affect the overall data quality. In using social urban data as proxies for the analysis of urban dynamics, the identification of these biases is of critical importance to the interpretation of the obtained results. To leverage the intrinsic biases of social urban data and to extract unambiguous knowledge about the dynamics of cities, the integration of data from multiple sources is, therefore, deemed necessary.

TOC

3 Transforming Heterogeneous Data for Cities into Multidimensional Linked Urban Data

§ 3.1 Introduction

In measuring and analyzing the complex dynamics of urban systems, it is required that data from more than a single source are considered in conjunction. However, the combination of heterogeneous data is hardly straightforward. What makes the assembly cumbersome, is in fact the inherent diversities of the sources from which the data stem. More specifically, these heterogeneities may pertain to differences in syntax (i.e. different data encoding), schemas (i.e. different structure and entity relationships), semantics (i.e. diverse contextual interpretations), or combinations of these three aspects (Cruz & Xiao, 2009). As a matter of fact, the diversities of various datasets are proportional to the amount of sources. Thereby, it comes as no surprise that contemporary urban analytics are faced with increasingly heterogeneous data, forasmuch as the range of available sources that provide information about the city also expands rapidly. Chapter 2 specifically addressed this particular issue. Thus, the challenge is to enable the fusion of different urban data by alleviating the various heterogeneities. In other words, it is essential to explore data integration methods and techniques, so as to allow for complex urban models and simulations to be generated, and for demanding analytical questions to be answered.

The process of generating interoperable data requires that all three types of heterogeneity (i.e. syntactic, schematic, and semantic) are addressed. Nevertheless, the majority of existing standards for data integration concentrate on a single type. In particular, standards for syntactic interoperability provide guidelines for representing data in a common way and format. Conversely, schematic interoperability methods engage in matching different schemas, which are mainly used for describing how records in a database relate to one another. In this regard, they are mostly appropriate for highly structured datasets. As a consequence, syntactic and schematic interoperability standards are ill-suited to dealing with unstructured data, such as those generated from social media.

On the other hand, semantic interoperability efforts focus on conceptual models for building consensus among different disciplines that use varied naming conventions to describe the same data record or real-world entity. This is particularly crucial for urban analysis and planning, as not only the combination of heterogeneous data is required, but also several disciplines from various scientific domains are involved, making use of different terminologies. Semantic analysis is significant for interpreting the content of social urban data and for assessing its relevance to real-world urban dynamics. Concomitant with the semantic integration efforts are the recent advances in the Semantic Web and Linked (Open) Data. These endeavors focus on providing machine-processable descriptions of heterogeneous datasets, while further supporting their publication, retrieval, and reuse on the Web (Domingue, Fensel, & Hendler, 2011). Therefore, they open up new possibilities in studying simultaneously the relations between different aspects of urban systems.

This chapter focuses on the integration of heterogeneous urban data and on processes for generating links with external datasets. It first explores heterogeneities in urban data and outlines various approaches to data interoperability. Besides presenting the various standards for common formatting and representation, it further focuses on existing methods and technologies for semantically annotating urban data to allow for shared interpretation by both humans and computational systems. Driven by the current lack of domain-oriented frameworks for data integration, the chapter proposes a methodology for urban data integration and interlinkage. The designed methodology follows an ontology-based data integration approach and accommodates a variety of semantic (web) and linked data technologies. The methodology addresses issues of urban data integration, linked urban data generation, and publication to the Linked Open Data (LOD) cloud for further exploitation in urban analytics.

Finally, the proposed methodology is demonstrated through a use case, employing real-world, large-scale spatiotemporal data from multiple sources. In particular, it employs data from three different public transportation organizations that cover the entire transport network of the city of Athens, Greece. The data contain information about the origin and destination locations, stop times, daily, monthly, and yearly schedules, route descriptions, and geospatial features of routes and stop points, among others. Following the proposed methodology, the resulting integrated dataset is published to the LOD cloud and linked to other available geo-data on the Web.

§ 3.2 Background

§ 3.2.1 Urban Data Heterogeneities and Approaches to Interoperability

In Chapter 2, and specifically in Sect. 2.4.1, the diversity of data for cities has been described. Sensors, mobile phones, and social media have recently been placed next to the traditional sources used in urban analysis and modeling. Individually, each of these sources generates data with specific capacities, which can be exploited for measuring or inferring aspects of the urban environment. Reasonably, the larger the amount of data from different sources combined together, the wider the range of city aspects that can be covered. But to achieve such fusion, the inherent differences characterizing each data source have first to be overcome. These differences may relate to format, unit of measurement, level of accuracy, scale, degree of veracity, and naming conventions, to name but a few. In general, data heterogeneities can be classified into three categories: (a) *syntactic*, (b) *schematic*, and (c) *semantic* (Table 3).

Syntactic heterogeneity primarily refers to differences in file format or encoding. It constitutes the most basic and frequent type of heterogeneity among the three categories. In tabular representations of data (i.e. data of different types organized into rows and columns), a difference in syntax could also imply the use of different value separators (e.g. commas, semi-colons etc.) in tuples (i.e. organized sets of values) included in the datasets. To overcome this discrepancy and to further achieve interoperability across systems, the most common way is to convert the data in question into a shared structured format and/or representation, using standards. Such standards have been developed by organizations dedicated to spatial data interoperability, such as the Open Geospatial Consortium (OGC), or others focusing on data exchange over the Web, such as the World Wide Web Consortium (W3C). The majority of these standards concern the development of machine-readable data formats, such as the XML (eXtensible Markup Language), JSON (JavaScript Object Notation), the KML (Keyhole Markup Language), or the GML (Geography Markup Language) which allow datasets to be exchanged between computing systems or over the Web.

Schematic heterogeneity relates to the use of different schemas among structured datasets, stored in database management systems (DBMSs). Generally, a schema determines the objects that are allowed to be stored in a database. In the case of relational databases, schemas further specify the relationships between these objects. Thereby, schematic diversities can solely be encountered in structured datasets. Sensor data and CDRs constitute exemplary cases hereof. Frequently, schematic

heterogeneities refer to the use of different naming conventions to describe the same piece of information, which makes them analogous to semantic heterogeneities. In integrating datasets characterized be different schemas, the prevailing solution to date is that of schema matching. The latter entails the identification of semantically related objects, which are subsequently matched. Besides the differences in object definitions, other challenges include diversities in the types or units of measurement, at the data level. In the specific context of web-enabled sensor resources and networks, the Sensor Web Enablement (SWE) initiative by OGC is particularly dedicated to tackling such heterogeneities (Botts, Percivall, Reed, & Davidson, 2007).

Lastly, semantic heterogeneity concerns the differences in the meaning of data values, as well as in the interpretation of these values, which is largely influenced by the context. Unlike schematic heterogeneity, which is primarily detected in structured datasets, semantic diversities can also be found in unstructured and semi-structured data. The rapid increase of unstructured data, such as those generated from social media, reinforces the issue of semantic uncertainty and vagueness. Examples of such uncertainties or diversities may pertain to synonymous terms, such as "urban fabric" and "city fabric"; homonymous terms, such as the word "point" which might refer to a geographical entity denoting a specific location or to a measurement unit; and similar terms expressed in different languages. Encoding the meaning associated with the data, in such a way that it is machine readable and processable can be a rather intricate process. Especially in the domain of urban analytics and planning, which involves a wide range of geospatial and spatiotemporal data, as well as a broad spectrum of disciplines using varied terminologies, semantic integration is essential in the exchange and reuse of cross-sector datasets. The prevailing approach for addressing semantic integration to date, is based on ontology engineering techniques. Ontologies are conceptual models that formally describe a set of real-world entities and explicitly define the relationships between them (Gruber, 1993; Guarino & Giaretta, 1995; Mars, 1995; Studer, Benjamins, & Fensel, 1998; Zhu, 2014). Their aim is to provide a shared vocabulary or knowledge model among the different stakeholders of a domain, or address wider communities covering multiple domains. As such, ontologies are usually the outcome of joint effort among domain experts. This Chapter is specifically concerned with issues pertinent to semantic integration.

TABLE 3 Types of data heterogeneity and corresponding approaches to interoperability.

Type of heterogeneity	Description	Approach to interoperability
Syntactic	Difference in file format or encoding (e.g. different value separators, different identifiers)	OGC & W3C standards
Schematic	Difference in database schema (e.g. different naming conventions)	Schema mapping
Semantic	Difference in the meaning of data values (e.g. synonymy, polysemy, different abbreviations etc.)	Ontology engineering

§ 3.2.2 Ontology Engineering for Urban Data Integration

In the process of semantic integration, ontologies play a pivotal role, by accounting for shared vocabularies and formal definitions of domain concepts and their interrelationships. As it was mentioned in the previous section, these concepts represent real-world entities (e.g. a point of interest, an urban block, a building etc.), which are expressed in a machine-processable format, meaning that they can be further interpreted by computing systems. Despite the different knowledge representation languages used, ontology concepts are most frequently represented as *classes*. These are subsequently organized into *hierarchies* or *taxonomies*, connecting the different concepts together through explicitly defined *relationships*. Other essential features of ontologies include *attributes*, which enrich concepts with data types (e.g. integers, strings, Booleans etc.) and values, as well as *axioms*, which constitute logical statements (e.g. a class is *kind of* another class, or a class is *different from* another class etc.) enriching the knowledge about the domain in question. At the data level, ontologies are further characterized by *instances* – also referred to as *individuals* – which constitute specific objects belonging to more generic classes (e.g. a train station is an instance of a generic class representing all types of buildings). Relations and attributes may also contain several instances (Table 4). Being built on these components, ontologies constitute prominent repositories for sharing, interpreting, and reusing knowledge about a domain.

TABLE 4 Ontology elements.

Ontology Elements	Description
Class	A description of a concept in a domain (e.g. city, urban fabric etc.)
Individual (or Instance)	A specific element that is member of a certain Class (e.g. Amsterdam is an instance of the class "City", an urban block is an instance of the class "Urban Fabric" etc.)
Property (i.e. object, datatype)	A relationship between two ontology elements (e.g. *object properties* relate individuals to other individuals, *datatype properties* relate individuals to data values)
Datatype	A type of data value (e.g. literal, Boolean, string, integer etc.)
Annotation	Additional information to a Class or Property (e.g. version, label, comment, creator etc.)
Axiom	A logical assertion (or statement) (e.g. a *building* [i.e. a Class representing all buildings] *is a sub-class of an urban block* [i.e. a Class representing all urban blocks])

The main difference between ontologies and other conceptual models used in computing systems – such as the Entity/Relationships Model (E/R Model), the Unified Modeling Language (UML), or database schemas – is that the former describe an existing knowledge domain, whereas the latter prescribe a system that is about to be built (Grimm, Abecker, Völker, & Studer, 2011). Thereby, computational conceptual models precede a certain information system, whereas ontologies come after the establishment of a domain (Bishr & Kuhn, 2000; Métral et al., 2009).

In formally expressing ontologies, several languages have hitherto been developed, with the Web Ontology Language (OWL) being the leading one. The latter has been specifically designed to enable web-related applications of ontologies, as it will be further discussed in the following sections. However, OWL does not constitute a single knowledge representation language, but rather a family thereof, characterized by varied degrees of semantic expressiveness. In particular, OWL Full has the maximum amount of knowledge representation features; OWL DL (Description Logic) is also highly expressive, yet entailing certain component restrictions; and OWL Lite is the least expressive one, comprising only basic constructors (Antoniou & van Harmelen, 2009). The latest advancement of OWL, namely OWL 2, also contains three variants, specifically OWL 2 EL (Existential Logic), OWL 2 QL (Query Language), and OWL 2 RL (Rule Language) (Figure 3).

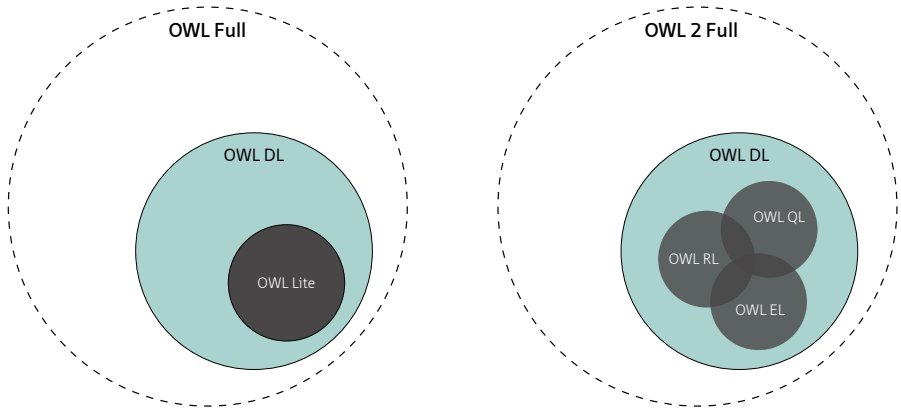


FIGURE 3 Family of OWL and OWL 2 languages.

In addition to the various languages, ontologies may be used to represent knowledge about a specific domain (e.g. urban planning, geographic information science etc.), or very generic concepts that cover multiple domains. In the first case, ontologies are called *domain ontologies*, whereas in the second case they are often referred to as either *top-* or *upper-level*, or even *foundational ontologies* (Grimm et al., 2011). Domain ontologies may incorporate concepts, or be established upon, upper-level ontologies through alignment processes, which will be discussed later.

In developing a domain ontology to which local heterogeneous datasets can be mapped, several actions need to be performed, together comprising what is referred to as *ontology engineering* methodology (Gómez-Pérez et al., 2004). At first, the *requirements* that the ontology has to fulfill are defined. These may include its general purpose and goal, as well as the intended end-users and its specific usability, to name but a few. Drawing on these requirements, relevant *terms* are *extracted* either directly from the local datasets or generally from the knowledge domain, so as to be later formally represented in the ontology corpus. The collected terms, which in fact represent various ontological entities, are subsequently structured and organized into hierarchies, comprising upper and sub-classes. At the same stage, which is generally referred to as *ontology conceptualization*, the relationships among the various terms (or classes) are also defined, in addition to relevant axioms and attributes. Besides the initially extracted terms, relevant concepts or entities stemming from existing upper-level ontologies, in addition to domain-specific concepts and terms from external structured vocabularies can be incorporated into the hierarchy. This allows for more efficient resource allocation and time management, since not all terms have to be defined from scratch, while supporting interoperability among diverse knowledge models. To this end, *ontology search* and *selection* processes need to be carried out, so as to respectively identify and choose the most appropriate ontologies or concepts to be incorporated. A frequent, yet significant, impediment to this process is the issue of

multilingualism, that is, when concepts are expressed and defined in different natural languages, subsequently influencing the degree of integration. At the *implementation* stage, following the selection and integration procedures, the definitive structure of the knowledge model is determined. This is achieved by specifying concrete axioms and relationships among the classes, deciding on the extent to which elements from external ontologies will be reused and/or aligned, as well as by introducing potential instances. Finally, the quality of the developed ontology has to be *evaluated*, not only in terms of domain coverage, but also for possible modeling and reasoning inconsistencies. This is achieved through the use of reasoners or other frames of reference. An implementation of the procedures mentioned in this paragraph will be presented in the applied example (Sect. 3.3.2.3).

In the context of urban planning, analysis, and modeling, the need for ontologies increasingly gains momentum, as the landscape of available data sources is progressively becoming complex and the range of involved disciplines is becoming broader (Falquet, Métral, Teller, & Tweed, 2011). The simultaneous consideration of spatial, social, and temporal aspects of urban systems, in addition to the diversity of urban data, provide a reasonable motivation for the development of shared knowledge models. However, the major obstacle in the development of such models lies in the existing and growing amount of city-related terms with vague meaning (e.g. the terms "place", "event", "downtown", "function", "land cover", "smart grid", "interactions" etc.), making it difficult to reach consensus among the disciplines involved. In tackling this issue, *ontology design patterns* can be particularly helpful, by providing a set of best practices and reusable strategies that can be further applied to building urban-related ontologies (Gangemi & Presutti, 2009).

The majority of related work to date has been conducted in the field of geographic information systems (Janowicz, Scheider, Pehle, & Hart, 2012). Examples include research on the formal definition of vague concepts, such as the term "place", in an attempt to provide a shared interpretation among the various stakeholders in geography and planning (Abdelmoty, Smart, & Jones, 2007; Goodchild, 2011; Jones, Alani, & Tudhope, 2001; Lutz & Klien, 2006). Such knowledge models could also be valuable to the discovery of place-related content from human-generated data in social media (McKenzie et al., 2015; Purves & Hollenstein, 2010). Drawing on the increasing significance of the varied temporal dimensions of urban processes, ontologies have also been developed to formally represent spatiotemporal properties of geographic data (Bittner, Donnelly, & Smith, 2009; Christakos, Bogaert, & Serre, 2001). Despite these efforts, the development of domain ontologies for urban analytics and planning is still at a nascent stage (Zhu, 2014).

§ 3.2.3 Data Integration on the Semantic Web

The methods and techniques that have been discussed thus far primarily concern approaches to data integration at the local level. However, the capacity of reusing and sharing semantically enriched datasets can be further extended by publishing and integrating these data on the Web. This can specifically be achieved through the principles and technologies of the *Semantic Web*.

According to its definition, the Semantic Web does not resemble an independent Web, but rather an extension of the existing one (Berners-Lee et al., 2001). What differentiates it from the current Web, is that the latter was essentially designed for linking documents that can be read and understood by people, whereas the Semantic Web is about linking data represented in a machine-processable way, so as to be easily used by both computers and people. Although the present state of the Web – frequently referred to as the *Social Web* or *Web 2.0* – is largely different from its initial stage – usually referred to as *Web 1.0* – in that it allows users to actively contribute content without following a centralized and prescriptive schema, it nevertheless continues to be about hyperlinks between documents. Conversely, the concept of the Semantic Web, which aspires to evolve into *Web 3.0*, aims to strengthen the reuse and integration of heterogeneous data that are not necessarily integral parts of certain documents. In this way, data can serve purposes different than the ones for which they were initially generated (Janowicz et al., 2012). Therefore, it offers a promising ground for the integration of disparate urban data and their exchange among the various city actors at a larger scale.

In achieving these goals, the Semantic Web architecture relies on ontologies and ontology-related technologies (Domingue et al., 2011). In the previous section, OWL and its various subcategories have been described as the predominant family of languages for expressing ontologies to be further used in a web context. However, OWL is built on top of a much simpler standard for describing and linking metadata, namely the Resource Description Framework (RDF).

RDF constitutes in fact the cornerstone of Semantic Web technologies and is further recommended by W3C. As its name indicates, it refers to a framework – in particular a data model – for semantically describing resources on the Web (Gandon, Krummenacher, Han, & Toma, 2011; Schreiber & Raimond, 2014). These resources may represent any real-world entity (e.g. a place, a person, an urban block, a building, a sensor, an administrative region, an organization etc.), each of which is uniquely described by using Uniform Resource Identifiers (URIs). Relationships – which in the RDF terminology are called properties – among the resources are also URIs. The properties link resources to other resources that function as property values, also described through URIs. Links between properties may also be established. Together,

the resources, the properties, and the property values comprise the fundamental RDF structure that is widely known as RDF triple. A triple further operates as a statement, with resources being the subjects, properties the predicates, and property values the objects. Combined, they represent a directed, labeled graph, where the subjects and the objects are the nodes and the predicates the connecting lines (i.e. edges). Therefore, RDF is a graph-based data model that enables the semantic description of any real-world entity on the Web (Schreiber & Raimond, 2014) (Figure 4).

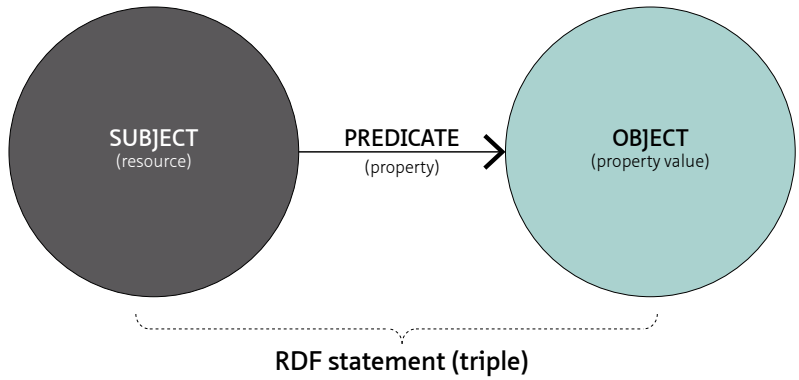


FIGURE 4 Graph-based structure of the RDF triple.

By mapping datasets to a domain ontology, a semantically rich RDF serialization is produced, in which each entity is identified by a unique URI. This further allows data elements to be easily discovered by machines and to also establish links with other data. In this way, new (linked) datasets are created that extend the application capacities of the source data (e.g. correlating weather data with transport information and fuel consumption records). In enabling such interoperability and exchangeability capacities, RDF statements need to be expressed in a machine-processable format. The prevalent RDF serialization standard to date is XML, yet statements can further be expressed as N-Triples, TTL or Turtle (Terse RDF Triple Language), and increasingly as JSON-LD (JavaScript Object Notation for Linked Data).

Finally, an integral part of the Semantic Web technology stack is the possibility to discover the datasets through specific queries, so as to be further exchanged or used in other applications. To this end, the SPARQL (SPARQL Protocol And RDF Query Language) language has been developed to query and retrieve RDF statements from the Web. An extension of it, named GeoSPARQL, is particularly interesting for urban analytics and comprises a specialized query standard established by OGC to allow the query and exchange of geospatial RDF data (Battle & Kolas, 2012). Such queries can be performed through dedicated services, known as SPARQL Endpoints.

§ 3.2.4 Generation and Publication of Linked Urban Data

Besides establishing internal links among the data that comprise an integrated dataset, the publication of the latter on the Web would further enable the creation of links with other integrated datasets from different domains. This could be beneficial to the analysis of urban dynamics that require the combination of various social and spatial attributes. The processes described in the previous sections facilitate the integration of heterogeneous data and the creation of links *within* datasets. This could suffice for a small-scale, local application, in which several datasets stemming from diverse sources need to be integrated to serve, for instance, the purposes of a complex urban model. To further ensure that these data resources are discoverable, accessible, and reusable in other applications, additional links have to be established *between* datasets from different domains on the Web. Interconnected data allow different stakeholders (e.g. planners, decision-makers, public organizations etc.) to explore and exploit datasets spanning several domains. To this end, data that are locally integrated through the use of semantic (web) technologies (e.g. ontologies, RDF statements, semantic matching etc.) further require the adoption of Linked Data principles to assure integration with external data on the Web, as well as discovery and full access to all the resources they describe (Heath & Bizer, 2011a). The resulting interconnected datasets are then available to be linked to other external datasets pertinent to similar or different domains (Heath & Bizer, 2011b).

In generating explicit links among structured data on the Web, Tim Berners-Lee has introduced a set of principles that form the underpinnings of data interlinkage (Berners-Lee, 2006). According to these principles:

- URIs have to be used to name and, therefore, identify any real-world entity;
- These URIs further need to be HTTP URIs to allow people discover these entities;
- When a certain URI is discovered, additional useful information has to be provided by using semantic standards (e.g. RDF, SPARQL);
- Additional links with external URIs also have to be included to enable people discover related entities.

The first two of the proposed guidelines draw, in fact, on fundamental web technologies, which enable the creation of hyperlinks through the HyperText Transfer Protocol (HTTP). Yet, the main difference lies in the URIs, which in this case are used for identifying any real-world entity or relationships between entities, rather than just web-based documents (see also Sect. 3.2.3). This further stresses the fact that the Semantic Web or Web of Data is actually an extension of the classic Web, built on top of its fundamental technologies, instead of being an independent type of Web. Relevant to this, is the proposed adherence to a certain standardized data model (i.e. RDF) and query language (i.e. SPARQL) for representing and retrieving structured data, which

are integral to the Semantic Web technology stack. Finally, the data resources – each of which is identified by a URI – and their explicit RDF links should be used as a basis to connect with other data resources, which also represent real-world entities and are identified by URIs (e.g. a function connected to a building in an urban block, a POI linked to an activity type carried out by a certain person etc.).

In addition to the above principles, Tim Berners-Lee provided a set of guidelines, in the form of a rating system, to evaluate the quality of the generated Linked Data. These guidelines – known as the *5-star Linked Data deployment scheme* – are mainly intended to foster the creation of high quality Linked Open Data (LOD), but can also be used for assessing the quality of proprietary Linked Data (Berners-Lee, 2006). According to the LOD deployment scheme:

- [***1-star*** dataset] Data available on the Web, in a machine-readable or non-machine-readable format, under an open license (for the case of open data);
- [***2-star*** dataset] Data available in a machine-readable format (e.g. XLS instead of JPEG);
- [***3-star*** dataset] Data that adopt the above principles in addition to being available in a non-proprietary format (e.g. CSV or JSON);
- [***4-star*** dataset] Data that comply with all the above and further make use of URIs and RDF statements to identify resources and explicitly describe their interconnections;
- [***5-star*** dataset] Data that adopt all the above principles and create additional links with external structured datasets.

The increasing amount of publicly available urban data generated by city-related organizations can be further enhanced by adopting the LOD principles. It could specifically increase the exploitation potential of open data, as it will be shown in the applied example later in this Chapter, as well as in Chapter 4.

The publication of integrated data on the Web involves the adoption of the Linked Data principles, as described above. To this end, data have first to comply with the original source's license attribution. The main issue hereof is privacy. Proprietary data without an open license and non-anonymized records cannot be published as LOD on the Web. This is particularly crucial when it comes to linking emerging urban data sources, such as CDRs and social media data, in which user anonymity has to be preserved and protected. Following this, both the locally generated RDF dataset and the ontology to which the original data were mapped need to be made accessible, by means of RDF repositories. Finally, to allow both people and computing systems to discover and exploit the data, registration to existing and well-established data catalogues is required (to enable discovery by people), in addition to the creation of semantic sitemaps (to enable discovery by search engines) (Bauer & Kaltenböck, 2012). Open data complying with the Linked Data principles, can be registered to the

LOD cloud[1], which accommodates all the publicly available linked datasets that have been published on the Web. This could enable better reuse and exploitation of the linked dataset by different stakeholders. The most recent state of the LOD cloud (in 2014) includes a total of 1,014 linked datasets, classified into seven different domains, comprising more than 85 billion RDF triples (Schmachtenberg et al., 2014). Relevant to the focus area of this research, linked data pertinent to user-generated content consist of 48 datasets (yet mostly from blogs, rather than social media platforms), while those related to geographic data comprise 21 datasets, altogether composing about 7 billion RDF triples.

§ 3.3 Designing a Methodology for Urban Data Integration and Interlinkage

Although the approaches and methods described in the previous sections set the foundation for the integration and interlinkage of data from multiple sources, they comprise generic standards and guidelines that are not necessarily applicable to every domain (Radulovic et al., 2015; Villazón-Terrazas, Vilches-Blázquez, Corcho, & Gómez-Pérez, 2011). Driven by the lack of a domain-oriented framework for data integration, the following paragraphs present the design of a methodology for the transformation of heterogeneous urban data into multidimensional linked urban data. The methodology follows an ontology-based data integration approach and accommodates a variety of semantic (web) and linked data technologies. Overall, it comprises three main processes, namely: (a) urban data integration, (b) linked urban data generation, and (c) publication to the LOD cloud. In a nutshell, the proposed methodology consists of the following steps (see also Figure 5):

1 http://lod-cloud.net

- Semantic integration:
  - Selection of data sources and data preprocessing
  - Data analysis and modeling
    - Schema extraction
    - Resource naming strategy definition
    - Ontology design and development
      - Terms extraction
      - Reuse of existing ontologies and external structured vocabularies
      - Terms hierarchy and ontology conceptualization
      - Ontology evaluation
    - Mapping source data to the ontology (data transformation)
- Transformation into multidimensional linked urban data:
  - Establishing links with other sources
- Publication to the LOD cloud:
  - Ontology and RDF dataset publication on the Web
  - Documentation accessibility (human-readable and machine-processable)
  - Registration into a Linked Data catalog and publication to the LOD cloud

The methodology is demonstrated through a use case[2], employing real-world data from multiple sources. In particular, nine large-scale spatiotemporal data sets are collected from three public transportation organizations and cover the entire public transport network of the city of Athens, Greece. As part of the data integration process, an ontology for public transportation systems is also designed and implemented. The resulting integrated dataset is further linked to external resources to provide richer descriptions of the source data, and is eventually published to the LOD cloud.

2 The methodology and use case presented in this chapter (in combination with the ontology and linked data visualization tools presented in Chapter 4) have been awarded the 1st Prize for Linked Open Data for Smart Cities. The applied example was initially developed in the context of the 1st Summer School on Smart Cities and Linked Open Data (LD4SC–15) [June 7 – 12, 2015, Cercedilla, Madrid, Spain], organized by the Ontology Engineering Group (OEG) of the Universidad Politécnica de Madrid (UPM) and the Information Technologies Institute (ITI). The project is included in the outcomes catalogue of the READY4SmartCities, part of the EU FP7 Coordination and Support Action (Birov et al., 2015).

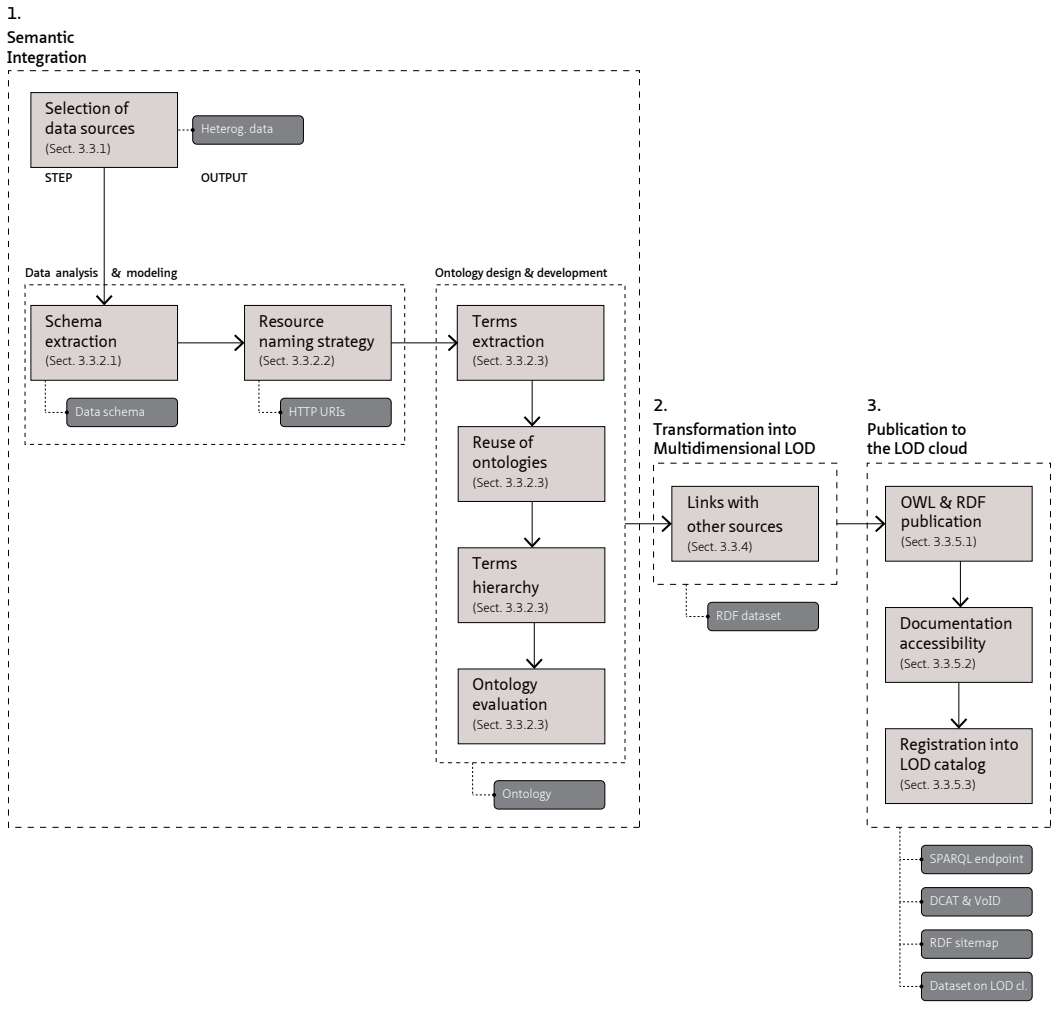


FIGURE 5 Diagram of the proposed methodology for transforming heterogeneous data for cities into multidimensional linked urban data.

§ 3.3.1 Data Sources

At the first step of the methodology, a set of requirements is initially specified, as regards the selection of the data sources. Bearing in mind that the goal of the final integrated dataset, resulting from the combination of various source data, is to become publicly accessible on the LOD cloud, the foremost essential requirement is that the initial data are also under open license and can be accessed from a public domain. Overall, the specified requirements for the demonstrated use case are as follows:

- Availability of the source data in a public domain, under an open license that allows further publication and reuse;
- Inclusion of both spatial and temporal parameters (e.g. geo-locations of origins, destination, and intermediate stops; different means of transport; time intervals; frequency etc.);
- Representation in a machine-readable and, preferably, non-proprietary format (e.g. CSV);
- Inclusion of entities that can be connected with generic entities from other domains to perform more complex socio-spatial measurements;
- Relatively large-scale dataset, in comparison with the average scale of the already published datasets to the LOD cloud;
- Reference to a real-world urban system.

On the basis of these requirements, the collected data of the presented example comprise 9 data sets derived from three different sources and covering the entire public transport network of the city of Athens, Greece. In particular, the source data originate from OASA (Athens Urban Transport Organization), OSY (Road Transportation), and STASY (Urban Railways), and are provided under an open license (Creative Commons Attribution 3.0) through a publicly accessible governmental web-based repository, based on an Open Government initiative. The different data sets contain records about the origin and destination points of each transportation service, as well as the entire network of intermediate stops for busses, trolleys, tram lines, metro, subway, and commuter rail, accompanied by geo-coordinates. Additional metadata include the full names of every stop, stop codes, route IDs, route types, directions, service frequency, timetable changes, as well as arrival and departure times. In total, the 9 different data sets comprise 2,100,000 tuples that are represented in the CSV (Comma Separated Value) format, which is both machine-readable and non-proprietary.

The first source of data, namely OASA[3], provides generic records about the service frequency, the timetables, the arrival and departure times, in addition to the route IDs and full names of origins and destinations, concerning all aforementioned means of public transport. The datasets provided by the second source, OSY[4], contain tuples about the bus and trolley network at a more disaggregate level than the ones collected by OASA. More specifically, they include the names (code and head sign) of each bus and trolley, as well as all intermediate stops between each origin and destination with precise geo-coordinates (i.e. pairs of latitude and longitude, using the WGS84 geodetic system). Besides latitude and longitude pairs, each stop is further described by a unique ID and code number, as well as by the full name of the street on which it

3 http://oasa.gr

4 http://www.osy.gr

is located. Finally, the third source, STASY[5], provides data about the entire network of metro, tram lines, suburban railway, and commuter rail. As is the case with the previous data source, each stop is accompanied by precise geo-coordinates, code numbers, full names, and nearby streets or major squares (Table 5).

TABLE 5 Data sources and data sets.

Source	Observations (attribute categories)	Mode of transport	Number of tuples
OASA (Athens Urban Transport Organization)	– Service frequency – Timetable – Arrival and departure times – Route ID – Origin & destination points	Bus Trolley Tram Metro Subway Commuter rail	51,872
OSY (Road Transportation)	– Code and head sign of buses and trolleys – Intermediate stops (geo-referenced) – Stop ID – Stop code number – Street network	Bus Trolley	1,983,955
STASY (Urban Railways)	– Intermediate stops (geo-referenced) – Stop code number – Stop full name – Streets and squares nearby each stop	Tram Metro Subway Commuter rail	64,173

As mentioned above, the datasets are publicly available and were retrieved from a web-based governmental repository[6]. The license accompanying the data is a Creative Commons Attribution 3.0 license[7], which allows data to be freely shared, transformed, adapted, reused, and republished. Thereby, the datasets meet the Open Data requirements, which further enable their publication to the LOD cloud (see Sect. 3.3.5) and their exploitation by third parties. In combination with the fact that all 9 data sets are provided online in a non-proprietary and machine-readable format, they can be rated as 3-star data, based on deployment scheme by Tim-Berners Lee (see Sect. 3.2.4). The goal is thus to combine them into an integrated 5-star dataset, using semantic technologies.

5 http://www.stasy.gr

6 http://labs.geodata.gov.gr

7 The license is available at: http://labs.geodata.gov.gr/en/dataset/urban-transportation-routesathens

§ 3.3.2 Data Analysis and Modeling

§ 3.3.2.1 Schema Extraction

In the previous section, it was mentioned that the collected data are distributed across 9 different subsets, each of which addresses specific parameters of the Athenian public transport network. More specifically, these subsets are: the *Calendar* (containing data about timetables); the *Routes* (comprising data about route IDs, in addition to arrival and departure times); the *Trips* (referring to origins and destinations); the *Stops_OSY* (with data about bus and trolley intermediate stops and their geo-coordinates); the *Stops_STASY* (as previously, yet only for metro, tram, suburban railway, and commuter rail); the *Stop_times_OSY*; the *Stop_times_STASY*; the *Agency* (containing metadata about each source); and the *Feed_info* (comprising data about service frequency). The structured format in which the data are provided (i.e. CSV) allows the extraction of the local schemas, by analyzing their particular features, types, and values.

Although the data originate from three different sources, common elements are identified across them, subsequently leading to the establishment of local links between the elements that are semantically equivalent. In particular, the element *route_id*, referring to an identifier encoding a certain route, is detected in both the *Routes* and the *Trips* subset. Similarly, the element *trip_id*, which encodes a particular trip type, is identified in the *Trips*, the *Stop_times_OSY*, and the *Stop_times_STASY* subsets. The *service_id* element, containing descriptions about the service frequency and iterations, is found in the *Calendar* and the *Trips* subset. The values of the three aforementioned elements are represented as strings. Finally, the element *stop_id*, indicating a unique code number for each intermediate stop, respectively exists in the *Stops_OSY*, *Stops_STASY*, *Stop_times_OSY*, and *Stop_times_STASY* subsets. Its values are represented as integers.

Other important elements include the *stop_lat* and *stop_lon*, respectively indicating the latitude and longitude coordinates of each stop, based on the WGS84 geodetic system and expressed as floats. These elements appear in the *Stops_OSY* and *Stops_STASY* subsets, but refer to different means of transport, depending on the data source. Also in the *Calendar* subset, there exist individual elements indicating each day of the week (i.e. *monday* up to *sunday*), the values of which are represented as Booleans. The latter signify whether or not a service is functional on a particular day. The complete schema of the dataset, containing the entire range of attributes and their values (i.e. data types), is illustrated in Figure 6.

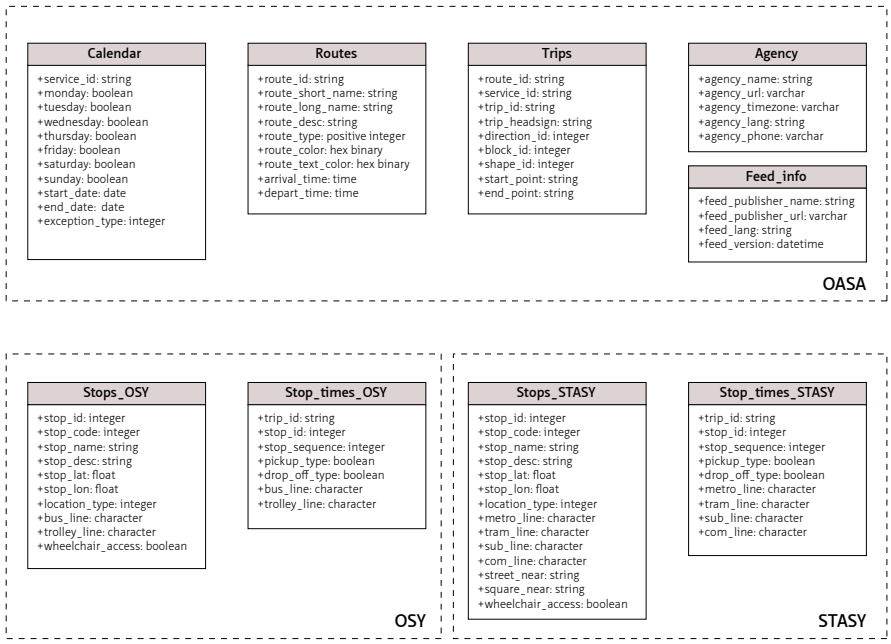


FIGURE 6 Data schema of the OASA, OSY, and STASY data sets.

§ 3.3.2.2 Resource Naming Strategy

Given that the goal is not only to generate an integrated dataset from the various sources, but also to publish it as Linked Open Data to strengthen its exploitation potential, HTTP URIs are used for identifying each resource in the dataset and their relationships. This decision specifically draws on the first and second principles of Linked Data (see Sect. 3.2.4). As has been discussed previously, resources in the context of the Semantic Web may represent any real-world entity, whether this refers to the entity itself or to a web-based document that describes what this entity is about. To this end, there generally exist two strategies to enable people discover these resources, namely *303* or *slash URIs* and *hash URIs* (Sauermann & Cyganiak, 2008). Both strategies allow the disambiguation between a real-world entity and a document that describes it. Moreover, they both give the opportunity to people or machines (both may operate as clients, sending requests to a server) to retrieve a representation of a resource in the format that best satisfies the criteria of readability – respectively, HTML and RDF (Heath & Bizer, 2011a).

Prior to defining which of the two resource naming strategies is more suitable for a dataset, the first step is to determine a persistent domain address. This will subsequently ensure that the resulting URIs will also be persistent. Changing URIs can have a negative effect on third-party applications that rely on them for retrieving relevant resources. The chosen URI form determines the path structure that follows the domain address. Generally, the main difference between slash and hash URIs lies in the way in which a client (human or machine) requests a certain representation of a resource from a server, as well as in the way a server responds to that request. In HTTP terminology, this process is known as *content negotiation* (Fielding et al., 1999). In the case of slash URIs, two HTTP requests are needed so that the client obtains a description of a real-world entity in the desired format (usually HTML for human clients and RDF for machine ones). Conversely, hash URIs are characterized by a distinctive part – called *fragment* –that is disconnected from the rest of the URI by a hash symbol. This part is excluded from a client's request and, thereby, a server returns all the available resources that share the same non-hash part (Sauermann & Cyganiak, 2008). As a consequence, in cases of large-scale sets of resources, the hash URI strategy will return to the client an extensive amount of unnecessary resource descriptions. Nevertheless, this strategy is more straightforward than slash URIs, as single descriptions can be obtained with only one request instead of two.

In the example presented here, a custom URI domain on GitHub is first created, in particular *https://route-owl.github.io/*. Given that the initial data from the different sources contain about 2,100,000 tuples, which are also frequently updated, and the resulting RDF statements will be much larger in volume, the chosen resource naming strategy is that of slash (303) URIs. Thus, the URI path form has the following general pattern:

https://route-owl.github.io/resource/<resource_type>/<resource_name>

For instance, the URI that returns the RDF statements of the various routes has the form:

https://route-owl.github.io/resource/route/route.rdf

The ontology model (described in the following section), to which the source data are mapped, uses a separate subfolder in the base URI domain. This is to disambiguate between the ontology model per se and its instances. Since the ontology comprises a much smaller and rather stable set of resources, the hash URI strategy is chosen in this case. Therefore, the corresponding URI path for the ontology model has the following generic pattern:

https://route-owl.github.io/ontology#

Accordingly, the generic URI paths for class names and properties (relationships) respectively follow the form:

https://route-owl.github.io/ontology#<ClassName>

and

https://route-owl.github.io/ontology#<propertyName>

§ 3.3.2.3 Ontology Design and Development

In order for the diverse source data to be integrated into a coherent dataset, a domain-specific ontology is developed that operates as a semantic model to which the initial data and their schemas are mapped. The ontology further assists in resolving the semantic discordance between concepts inherent to the various data sources, in addition to preventing term redundancy. In this example, it also aims to be used by planners, decision-makers, and other city stakeholders for performing queries and extracting information from the integrated data.

To this end, the developed ontology, named ROUTE, which stands for Route Ontology of Urban Transportation Entities, formally describes concepts of multi-modal public transportation systems and the relationships between them. It particularly comprises classes about transport services, pick-up and drop-off types, geospatial concepts about stops and routes, as well as temporal concepts describing time intervals, frequency, duration, among other related entities. It further enables formal representation of human agents, agencies (e.g. transportation companies), transfer types, and additional features, such as fare and payment methods, accessibility, and zoning. Classes are accompanied by several object and data properties, which define the relationships between classes, allow full statements (axioms) to be built, and specify values and units of measurement. By mapping data to the ROUTE ontology, third-party stakeholders can extract and infer combined information from various sources and, further, incorporate them into more disaggregate models of urban flows, interactions, or simulations of public transport systems. The following paragraphs describe the steps of the ontology design and development process.

Terms extraction

The extracted schema from the source data (Figure 6), presented in Sect. 3.3.2.1, serves as the basis for the terms that mainly correspond to classes in the ROUTE ontology. Further, the structure of the various schema elements and their attributes

assists in building the, later described, term hierarchy. Examples include terms related to route concepts, such as *start point*, *end point*, *stop*, *station*, *route type*, among others which are directly extracted from the data schema. Others pertain to modality concepts, such as *transportation type*, *pickup* and *drop-off type*, *transfer type*, and *wheelchair accessibility*. Geographic entities are represented by terms such as *place*, *administrative area*, *city region*, *point*, and *zone*. Moreover, time-related concepts include terms such as *temporal unit*, *date-time interval*, *time zone*, *stop time*, *trip duration*, among others.

Wherever possible, synonyms of the aforementioned terms were also included in the ontology vocabulary. This allows a wider range of semantics to be covered that would enable further links with relevant external datasets to be established. In achieving this, the ROUTE example draws on equivalence relations; an inherent functionality to OWL–DL. For instance, *temporal entity* is defined as equivalent to *instant* or *interval*; *temporal unit* is equivalent to *day*, *hour*, *minute*, *second* etc. To further increase the ontology's degree of versatility and, thereby, its potential to create external links, the majority of the terms were described in four languages, namely English, French, Greek, and Irish Gaelic. The extracted terms are then classified into concepts that operate as classes, object properties (i.e. relationships between classes), data properties (i.e. relationships between classes and datatype values), or instances (i.e. individuals by class). For instance, the terms *bus*, *metro*, *subway*, and *tram* operate as instances to the class *route type*, which is one of the main schema elements.

Reuse of existing ontologies and external structured vocabularies

The reuse of already existing ontologies, standards, and structured vocabularies in semantically integrating heterogeneous data enables the latter to relate and interact with datasets and applications relying on established knowledge models. It further assists in preventing semantic redundancy between similar or closely related concepts included in different ontologies. Drawing on this approach, the ROUTE ontology reuses concepts, relationships, and axioms from three existing ontologies and 15 external structured vocabularies.

Concerning existing ontologies, ROUTE firstly imports concepts and relationships from the *GTFS* (General Transit Feed Specification) ontology[8]. GTFS is in fact a direct translation of the general transit feed specification into an ontology, so that it can be used in a Semantic Web framework. It comprises a well-established standard for describing concepts pertinent to routes and route types, transfer types, trips, stops, and

8 Namespace: http://vocab.gtfs.org/terms#;
Homepage: https://raw.githubusercontent.com/OpenTransport/vocabulary/master/gtfs/gtfs.ttl

service availability, for several modes of transport. Its use in structuring transport data in numerous applications worldwide, in addition to its relevance to several schema elements of ROUTE, make it an excellent model upon which the ROUTE ontology can be built. The GTFS model is directly imported into ROUTE, yet many of its elements are extended by means of additional axioms.

To further capture both geospatial and temporal concepts, elements from the *WGS84 Geo Positioning* (*WGS84_pos*) vocabulary[9] and *Time* (*owl-time*) ontology[10] are respectively reused. The former comprises concepts about general geographic entities such as *wgs84_pos:Location* and *wgs84_pos:Point*, among other entities for formally representing latitude, longitude, and altitude information, using the WGS84 geodetic reference system. Conversely, the *Time* ontology incorporates several concepts pertinent to time intervals, date-time description, frequency, and duration description, and their relationships. In the case of these two vocabularies, element integration into ROUTE is carried out by means of referencing resource URIs, instead of directly importing them.

In relation to the several external structured vocabularies, ROUTE reuses elements from 15 different ones, describing various concepts. The majority of these vocabularies are widely-recognized and used in several applications thus far. Thus, by adopting part of their terms and by integrating them into the ROUTE ontology ensures commonly accepted descriptions of concepts across various domains and, hence, increased interoperability potential in future applications. Closely related to ROUTE's scope, the *otn* (Ontology of Transportation Networks) provides more specialized concepts about road networks, land cover and land use, as well as traffic. To capture formal knowledge about agents, organizations, and their relationships, it references elements from the *foaf* (Friend of a Friend) vocabulary. Other terms derive from vocabularies such as *dbpedia-owl*, *dc* (Dublin Core), *dct* (Dublin Core Terms), *schema*, *terms*, and *vann* (Vocabulary for Annotating vocabulary descriptions). Within the LOD framework, it reuses concepts from the *ns* vocabulary, for describing Creative Commons rights in RDF format. Finally, it complies with the formalities of *owl*, *owl2xml*, *rdf*, *rdfs*, *xml*, and *xsd*. The set of external ontologies and structured vocabularies is extracted from the Linked Open Vocabularies (LOV) repository[11], so as to ensure that all models are provided under an open license, which additionally allows further processing and reuse (Table 6).

9 Namespace: http://www.w3.org/2003/01/geo/wgs84_pos#;
Homepage: http://www.w3.org/2003/01/geo/

10 Namespace: http://www.w3.org/2006/time#;
Homepage: http://www.w3.org/TR/owl-time

11 http://lov.okfn.org/dataset/lov/

TABLE 6 Direct (i.e. complete) or partial reuse of ontologies and structured vocabularies.

	Ontology | Vocabulary	Prefix	URI (Namespace)	Import
ONTOLOGIES	GTFS (General Transit Feed Specification)	gtfs	http://vocab.gtfs.org/terms#	direct
	WGS84 Geo Positioning	wgs84_pos	http://www.w3.org/2003/01/geo/wgs84_pos#	direct
	Time ontology	owl-time	http://www.w3.org/2006/time#	direct
	Ontology of Transport Networks	otn	http://www.pms.ifi.lmu.de/rewerse-wga1/otn/OTN.owl	partial
STRUCTURED VOCABULARIES	Dbpedia-owl	dbpedia-owl	http://dbpedia.org/ontology#	partial
	Dublin Core	dc	http://purl.org/dc/elements/1.1/	partial
	Dublin Core Terms	dct	http://purl.org/dc/terms/#	partial
	Friend Of A Friend	foaf	http://xmlns.com/foaf/0.1	partial
	NameSpace vocabulary (Creative Commons)	ns	http://creativecommons.org/ns#	partial
	Web Ontology Language vocabulary	owl	http://www.w3.org/2002/07/owl#	direct
	Web Ontology Language 2 vocabulary	owl2xml	http://www.w3.org/2006/12/owl2-xml#	direct
	Resource Description Framework	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	direct
	Resource Description Framework schema	rdfs	http://www.w3.org/2000/01/rdf-schema#	direct
	Schema vocabulary	schema	http://schema.org/#	partial
	Terms vocabulary	terms	http://purl.org/dc/terms/	partial
	VANN vocabulary	vann	http://purl.org/vocab/vann/	partial
	Extensible Markup Language vocabulary	xml	http://www.w3.org/XML/1998/namespace	direct
	Extensible Markup Language schema	xsd	http://www.w3.org/2001/XMLSchema#	direct

Terms hierarchy and ontology conceptualization

Besides acquiring appropriate terms for semantically describing schema elements and related concepts, a hierarchy is created, defining super-classes and sub-classes and the properties that connect them together. The various external ontologies and structured vocabularies, mentioned in the previous section, do not supply the entire set of concepts needed to describe the source data. Therefore, a set of new classes, object and data properties, and instances are introduced to serve the purpose of ROUTE. These new ontology elements draw, firstly, on the particularities of the source data at hand and, secondly, on its potential future application in urban models for the study of (human) flows and multi-modal mobility. As a result, the approach followed in conceptualizing the ROUTE ontology is both data-driven and competence-driven.

Based on these principles, the entire ROUTE conceptualization comprises 271 entities, classified into 51 classes, 166 object, data, and annotation properties, as well as 54 individuals (instances) and datatypes, implemented with 1,140 axioms. The chosen coding formalism is that of OWL2- for its high expressiveness and for being recognized by W3C as one of the standard languages to represent ontologies in the Semantic Web context (Table 7). In developing the ontology, the Protégé platform is used (Knublauch, Fergerson, Noy, & Musen, 2004; Stanford University).

TABLE 7 ROUTE Ontology metrics, types of correspondence, and annotations.

	Metrics, types of correspondence & annotations	Counts / Annot.	Examples
ONTOLOGY METRICS	Total number of ontology entities	271	Classes, properties etc.
	Classes	51	*RouteType, StartPoint, EndPoint, Agent, StopTime* etc.
	Object properties	67	*hasStartPoint, isLocatedIn* etc.
	Data properties	54	*lat, lon, day, timezone* etc.
	Annotation properties	45	*language, prefix, creator* etc.
	Individuals	40	*"Bus line 040", "Syntagma square"* etc.
	Datatypes	14	*string, integer, Boolean* etc.
AXIOMS	Axioms	1,140	Logical statements
	Subsumption correspondences (*subClassOf* axioms)	93	Municipality is a *subClassOf* administrative area
	Mereology correspondences (*partOf* axioms)	182	End point is *partOf* route
	Assertion correspondences (*isA* axioms)	712	Syntagma square *isA* start point
	Equivalence correspondences (*equivalentTo* axioms)	3	Day is *equivalentTo* temporal unit
	Disjointness correspondences (*disjointWith* axioms)	1	Instant is *disjointWith* proper interval
	Domain axioms	59	The domain of *hasStation* property is the class *Stop*
	Range axioms	90	The range of *hasTransferType* property is the class *TransferType*
ANNOTATIONS	Namespace prefix	route	
	URI	https://route-owl.github.io/ontology#	
	Languages	EN, FR, GR, GA	
	Coding formalism	OWL2-	

In structuring the hierarchy of the ontology components, the first type of correspondence is that of class subsumption, formally defined by rdfs:subClassOf relationships. This organizes the extracted terms into super-classes and sub-classes. Super-classes correspond to key domain terms that are included in the data schema and also represent related concepts, which together will form the main links with external datasets. Given that public urban transport is the core domain of ROUTE, the main parameters captured by the ontology relate to the following:

- Modes of transport: the class *gtfs:RouteType* describes the type of transportation system used on a route. Thereby, it is accompanied by several instances, such as *Subway*, *Metro*, *Bus*, *Tram* etc.;
- Route features: the general *gtfs:Route* concept, which marks a route followed entirely or partly by *gtfs:Trip*, is further enriched with the newly introduced concepts *route:StartPoint*, *route:EndPoint*, and *route:Stop* to respectively represent the origin, destination, and intermediate stops across a certain transport route;
- Trip features: additional attributes are described by classes such as *gtfs:PickupType*, *gtfs:DropOffType*, *gtfs:PaymentMethod*, *schema:PriceSpecification*, and *gtfs:WheelchairBoardingStatus*, among others;
- Service attributes: relevant classes include *gtfs:Service*, *gtfs:Transfer*, *gtfs:TransferType*, *gtfs:TransferRule* etc.;
- Agents: since the scope of the ontology is to also be used in the study of human activity patterns, ROUTE incorporates classes such as *foaf:Agent*, subsumed by *gtfs:Agency*, to respectively describe people and organizations in general and, more specifically, transportation providers.

Geographic and spatial attributes are represented by classes such as *schema:Place* – which subsumes *schema:AdministrativeArea* and *dbpedia:City* – *wgs84_pos:Point*, subsumed by *wgs84_pos:SpatialThing*. ROUTE also reuses *gtfs:Shape*, which describes a route's polygon, and *gtfs:Zone*, which represents the different urban zones crossed by a route. To formally represent various types of land uses, the *route:PointOfInterest* class is introduced, accompanied by several instances such as *Restaurant, Museum, College, Hospital* etc. For its importance in the transportation domain, the *gtfs:Station* consists an individual class, separate from the rest of POIs.

Temporal attributes are described through *time:DayOfWeek*, *time:DurationDescription*, *time:TemporalUnit*, *time:TimeZone*, with various sub-classes and instances. In addition, *gtfs:Frequency* is used for describing how often a certain mode of transport operates, while the *time:DurationDescription* identifies the travel time between a *route:StartPoint* and a *route:EndPoint*. In further disaggregating the generic *Duration* concept, the *route:StopTimes* is introduced to represent arrival times at intermediate stops. The semantic network of the ROUTE ontology hierarchy is (partially) shown in Figure 7.

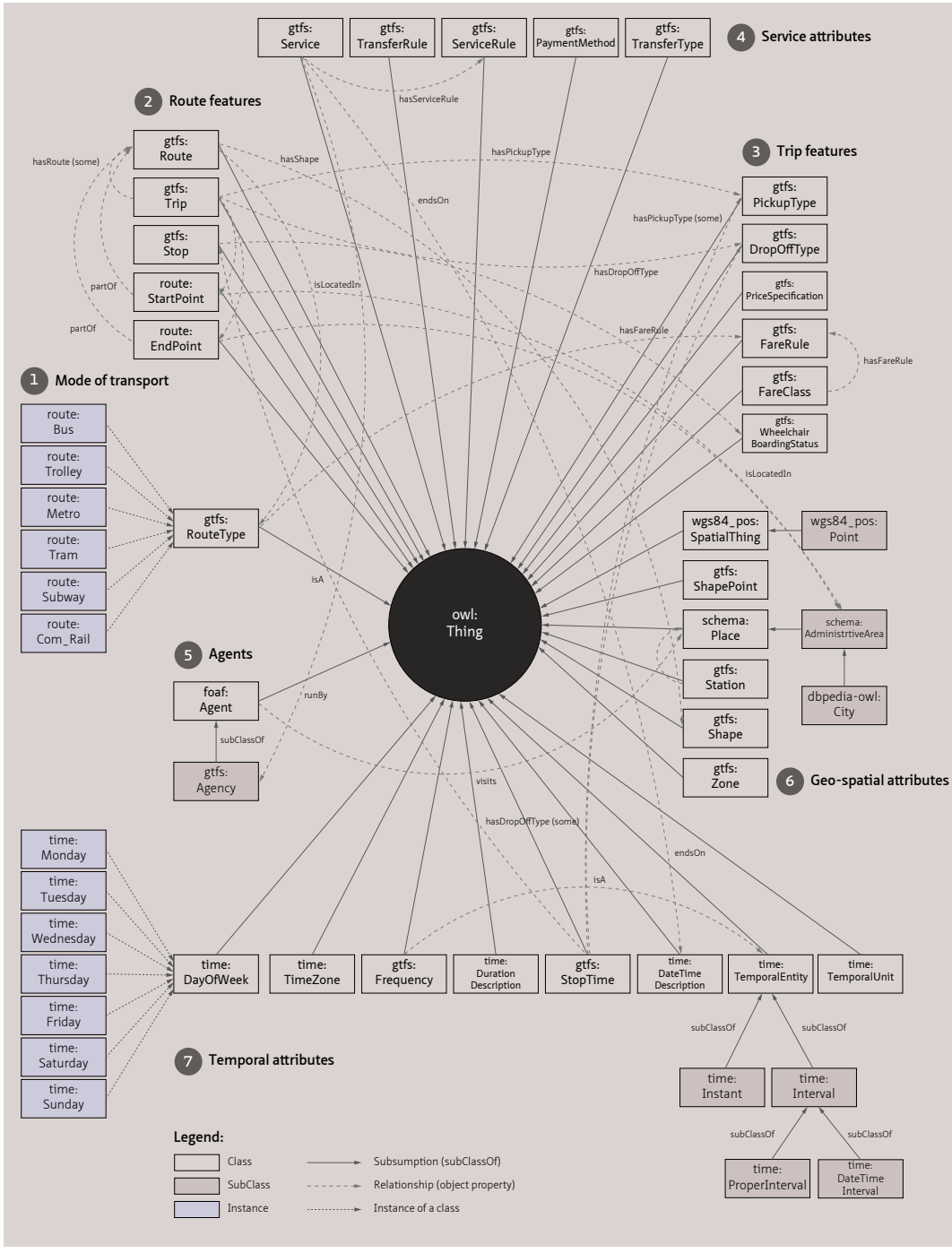


FIGURE 7 ROUTE Ontology. Semantic network representation of class hierarchy and indicative relationships (i.e. object properties).

Besides subsumption, ROUTE also incorporates mereology correspondences between classes, represented by the generic *partOf* and *hasPart* relationships. The latter is an inverse of the former, drawing on inverse relations supported by OWL. Upon these generic relationships, a set of object properties is built to specify how different super- and sub-classes are connected to each other. Subsumption relationships, described in the previous paragraphs, are necessary for logical reasoning, whereas mereology captures the knowledge of the ontology's domain. To further specify constraints in the way classes are related to each other, existential restrictions are established. The latter are denoted by the keyword *some* and indicate that the individuals belonging to one class have at least one type of relationship (object property) to the individuals of another class. For instance, the class *gtfs:Trip* represents the accumulation of *gtfs:StopTimes* of a certain type of vehicle following a particular route. The individuals belonging to *gtfs:Trip* are in fact various time intervals, characterizing e.g. a bus line. Besides arrival times at intermediate stops, a *gtfs:Trip* is also characterized by a *route:StartPoint* (origin) and a *route:EndPoint* (destination). Therefore, the individuals belonging to *gtfs:Trip* have *some route:hasStartPoint* and *some route:hasEndPoint* relationship with the individuals of *route:StartPoint* and *route:EndPoint* respectively. Similarly, each individual of these classes has *some route:isLocatedIn* relationship to the individuals of *schema:AdministrativeArea*, which is in turn *partOf* a *schema:Place*. Additional features characterizing the individuals of *gtfs:Trip* may refer to particular pickup and drop-off types. Hence, *gtfs:Trip* individuals have *some gtfs:hasPickupType* and *some gtfs:hasDropOffType* individuals respectively. Moreover, a *route:stopsAt* property is introduced to correlate the individuals of *gtfs:Trip* to those of *gtfs:StopTime*.

Given that the ROUTE dataset incorporates both spatial and temporal elements, spatial and temporal correspondences between classes are also established. To this end, new object properties are introduced and integrated with reused ones from external ontologies. Examples of spatial correspondences include properties such as *route:isLocatedIn*, *route:inDirection*, *route:hasLocation*, *route:hasStartPoint*, and *route:hasEndPoint*, among others. These relationships are used to construct axioms (statements) such as: A tram line (i.e. individual of *gtfs:RouteType*) is *route:inDirection* to a particular destination (i.e. individual of *route:EndPoint*), which in turn *route:belongsTo* a region in the city (i.e. individual of *schema:AdministrativeArea*). Examples of temporal object properties are *time:hasDurationDescription*, *time:inDateTime*, *route:startsOn*, *route:startsFrom* etc. With these properties, axioms such as the following can be described: A bus (member of *gtfs:RouteType*) of the line L1 (formally: *route:hasID some* string – that is a value of the class *route:ID*) *route:startFrom some ti* and reaches its final destination (*route:EndPoint*) at *some tj*.

In addition to existential restrictions described previously, ROUTE makes use of several cardinality restrictions to specify the *minimum*, *maximum*, and *exact* number of relationships that a class individual is allowed to engage in. For instance, a day – which is an individual of *time:DateTimeDescription* – can have a maximum

(*owl:maxCardinality*) of 1 literal (e.g. Monday). The aforementioned axiom actually uses a combination of a cardinality restriction (i.e. *maximum*) and a qualified cardinality restriction, by stating the amount of objects within the restriction (Bechhofer et al., 2004). Equivalence correspondences use the built-in OWL property *owl:equivalentClass* to signify which classes contain the same individuals, whereas the indication of individuals having different URIs, yet referring to the same real-world entity, is carried out by an *owl:sameAs* property. Finally, with the exception of reused elements, the URIs of the entire set of classes and properties introduced into ROUTE comply with the ontology's resource naming strategy, described in Sect. 3.3.2.2. The implemented ROUTE ontology is available online at the following link: https://github.com/ROUTE-OWL/ROUTE- OWL.github.io/blob/master/ontology.owl. The code expressed in RDF/XML syntax is also presented in Appendix A.

Ontology evaluation

Following the conceptualization and implementation of the ROUTE ontology, an evaluation process is carried out, so as to examine its technical quality and performance against various dimensions. Gómez-Pérez (Gómez-Pérez, 2004) identifies several criteria for evaluating ontologies, the most significant of which are completeness, consistency, conciseness, expandability, and sensitiveness. Potential logical inconsistencies, which relate to the existence of contradictory statements in the ontology conceptualization, are tested by means of reasoners. To this end, three consistency tests are performed, using a different reasoner each time, respectively FaCT++[12], HermiT[13], and Pellet[14]. No logical inconsistencies were detected in all three tests.

In further evaluating additional dimensions of the ontology, such as completeness (i.e. the entire set of ontology components is explicitly defined or can be inferred through reasoning) and conciseness (i.e. absence of redundancies and unused components), the OOPS! pitfall scanner[15] is used. OOPS! (OntOlogy Pitfall Scanner) is a web-based tool for detecting frequently appearing pitfalls (e.g. unconnected ontology elements, missing annotations, wrong equivalent properties etc.) in implemented ontologies (Poveda-Villalón, Gómez-Pérez, & Suárez-Figueroa, 2014). Several evaluations were performed during the ontology development process, so as to timely identify potential pitfalls prior to establishing the complete ontology conceptualization.

12 http://owl.man.ac.uk/factplusplus/ (Accessed on: February 19, 2016).

13 http://hermit-reasoner.com (Accessed on: February 19, 2016).

14 http://clarkparsia.com/pellet (Accessed on: February 19, 2016).

15 http://oops.linkeddata.es (Accessed on: February 19, 2016).

In the implemented ROUTE ontology only minor pitfalls were discovered by the web-based platform, all of which pertain to missing annotations in reused classes and object properties. However, this has no significant effect on the conciseness of the ontology. Besides, the set of newly introduced elements was entirely concise and complete. To further allow evaluation by domain experts, a web-based ontology browser and viewer have been specifically developed and will be presented in Chapter 4.

§ 3.3.3 Data Transformation and Integration: Mapping Source Data to the Ontology

To complete the semantic integration process of the proposed methodology, one of the most crucial tasks is to map each element of the heterogeneous source data to the developed ontology. The mapping ensures that the various original data, which may initially be represented in diverse formats, are eventually transformed into an integrated dataset that is expressed in a unified data format (i.e. RDF). The resulting RDF triples – as described in Sect. 3.2.3 – represent complete statements that connect the various data elements in a subject-predicate-object structure. Usually, subjects consist of the individuals belonging to a class (e.g. in the initial data, these could be the elements under a header, which is mapped to a class in the ontology), predicates contain the object properties, and objects may also comprise individuals of classes (e.g. attributes of the source data) or values.

Prior to mapping the nine source data sets to the ontology and, subsequently, transforming them into an integrated dataset, the type of RDF serialization is first defined. In Sect. 3.2.3, it has been described that RDF triples can be described in several machine-processable serializations, such as RDF/XML, N-triples, TTL (Turtle), and JSON-LD. In the ROUTE example, the chosen serialization is that of RDF/XML, which is also the most frequent expression of RDF triples. Although, this particular type of serialization is less human-readable, compared to the other formats, it is nevertheless easily processed by computing systems and can, thereby, serve semi-automatic or automatic Linked Data applications in urban analytics.

Following the selection of the RDF serialization, the mapping of the initial data to the ROUTE ontology is carried out by using the OpenRefine[16] platform and, specifically, its LODRefine[17] distribution, which contains extensions particularly built for Linked Data transformation purposes. OpenRefine is an open-source tool for cleaning,

16 http://www.openrefine.org/index.html (Accessed on February 21, 2016).

17 https://github.com/sparkica/LODRefine (Accessed on February 21, 2016).

normalizing, and transforming data from one format into another, while it is particularly effective with tabular data. Given that the initial data are in CSV format and having as a requirement the use of open-source tools and licenses throughout the integration and interlinking procedure, OpenRefine appropriately conforms to these prerequisites. Prior to merging the data together, the initial source subsets are cleaned from empty columns and missing elements that hinder the generation of complete RDF statements. After the initial data are cleaned and their values are normalized, each element is transformed into an RDF instance, by mapping it to the corresponding ontology component (hence, it becomes an individual of a certain class) and by, further, specifying its URI, based on the chosen resource naming strategy (see Sect. 3.3.2.2). For example, the class that formally describes a certain public transport route in the ROUTE ontology (i.e. *gtfs:Route*) has the following hash-based URI:

https://route-owl.github.io/ontology#Route

An instance of this class, e.g. the bus line with ID "040–20", which travels from Piraeus (member of the *route:StartPoint* class) to Syntagma square (member of the *route:EndPoint* class), is mapped to *gtfs:Route* and eventually obtains the following slash-based URI:

https://route-owl.github.io/resource/route/040-20

Following the mapping of an element to a class (i.e. subject), the object properties (i.e. relationships/predicates) of this element and the subsequent values (i.e. objects) of the properties are further specified. In the aforementioned example, the complete RDF description of the bus line "040–20" and its attributes is given as follows:

```
<rdf:Description rdf:about="https://route-owl.github.io/resource/route/040-20">
    <rdf:type rdf:resource="http://vocab.gtfs.org/terms#route"/>
    <route:shortName>40</route:shortName>
    <route:longName>PIRAEUS – SYNTAGMA SQUARE</route:longName>
    <gtfs:color>153CE0</gtfs:color>
    <gtfs:textColor>FFFFFF</gtfs:textColor>
    <route:hasType rdf:resource="https://routeowl.github.io/resource/route_
    type/3"/>
    <route:ID>040-20</route:ID>
</rdf:Description>
```

By iteratively applying the above-described mapping and transformation process to all data elements, the nine initial data sets are merged together, eventually generating the (single) integrated ROUTE dataset, which comprises 4,593,531 RDF triples in total.

The resulting RDF dataset has been validated for both syntax and representational consistency with the ROUTE ontology. Syntax validation has been carried out by means of the W3C RDF Validator[18]; a publicly available web-based syntax validation service for RDF datasets. The validation for representational consistency of the RDF dataset has been performed by means of reasoners, in particular Pellet and FaCT++. Both validation procedures showed that the obtained RDF data are fully consistent with the ROUTE ontology and are further expressed in a syntactically correct way.

The mapping and transformation procedure from the source data to the ontology components that is described in this section can be automatically performed for future updates, provided that the source data maintain their initial schemas. This is particularly important for spatiotemporal urban data that are characterized by frequent refresh rates. Therefore, urban models or simulations that require the use of integrated datasets from heterogeneous and dynamic sources can be automatically updated, once the first mapping has been established.

At this point, the obtained RDF dataset constitutes a 4-star dataset, as opposed to the initial 3-star data (see Sect. 3.2.4). The processes that have been described thus far comprise the first part (i.e. semantic integration) of the proposed methodology. The following sections elaborate on the second (i.e. linked data generation) and third part (i.e. publication to the LOD cloud) of the methodology.

§ 3.3.4 Establishing Links with Other Sources

Besides locally integrating heterogeneous data into a dataset, establishing links with resources of other integrated datasets enables the combination of information from various domains. In turn, this combination increases the potential for use in domains, other than those covered by the initial datasets. In the linking process, entities (e.g. terms, instances, properties etc.) of one dataset are connected with similar ones of the other dataset(s). To define and assess the similarity between two entities, there exist several methods and measures that are based on the label or comments of entities, the string structure, or the content, all of which are founded upon textual similarity aspects. Conversely, entities can be compared on the basis of their internal structure, i.e. their properties, the domain and range of these properties, the data types, and the cardinality restrictions (Euzenat & Shvaiko, 2013).

18 https://www.w3.org/RDF/Validator/ (Accessed on February 21, 2016).

In the specific context of Linked Data, links between datasets can be established by creating correspondences between instances. In the case where two or more datasets have been mapped to the same ontology, individuals belonging to the same classes can be connected with each other. In the case where the datasets have been mapped to different ontologies, an ontology matching process needs to be carried out first and, thereafter, the instances belonging to the aligned (equivalent) classes can be linked together (Nikolov, Ferrara, & Scharffe, 2011; Scharffe & Euzenat, 2011).

The process of linked data generation is further demonstrated in the presented use case.

The integrated dataset obtained in the previous phase establishes links with external geo-datasets that are already available on the LOD cloud. In particular, links are established with resources from DBPedia[19] and GeoNames[20]. DBPedia comprises a knowledge base which stores structured information extracted from Wikipedia pages in RDF format. Conversely, GeoNames contains geospatial semantic information about place names and their relations, also expressed in RDF format. At the current state of the DBPedia dataset, more than 526,000 places are represented as URIs. By generating links with resources of these particular datasets, the aim is to take advantage of the large amounts of external interconnections that these two datasets have already established and, thereby, enrich the integrated dataset with information that is not inherent to the source data.

In establishing links with external geographic datasets, the classes that have potential to be connected are first specified. Subsequently, the instances belonging to these classes are linked together, by means of the *owl:sameAs* property. In particular, the individuals of the classes *schema:AdministrativeArea*, *gtfs:Agency*, and *dbpedia-owl:City* are connected with those belonging to the equivalent classes of DBPedia, whereas the instances of *wgs84_pos:SpatialThing* are linked to the individuals of the same class in the GeoNames ontology. The LODRefine distribution of OpenRefine, along with its RDF extension, are used again (see also Sect. 3.3.3), yet in this case for performing the data interlinking. The links are included in the ROUTE dataset and are identified by *owl:sameAs* relationships. For example, the following RDF code describes the establishment of a link between the administrative region of "Glyfada" – in its Romanized format – and the corresponding instance of the same resource in DBPedia (in Greek):

19 https://datahub.io/dataset/dbpedia (Accessed on February 22, 2016).

20 https://datahub.io/dataset/geonames-semantic-web (Accessed on February 22, 2016).

```
<rdf:Description rdf:about="https://route-owl.github.io/resource/ad_area/
GLYFADA">
    <rdf:type rdf:resource="http://schema.org/AdministrativeArea"/>
    <rdfs:label>GLYFADA</rdfs:label>
    <owl:sameAs>http://el.dbpedia.org/resource/Γλυφάδα</owl:sameAs>
</rdf:Description>
```

In total, 124 links are generated to the DBPedia dataset, whereas 15,956 links are established with the GeoNames dataset (Table 8). By establishing those links, the already existing relations in the aforementioned datasets, enable the original data – e.g. in relation to the latitude and longitude of start and end points, as well as the various intermediate stops in different administrative areas – to be connected with resources about the census population (DBPedia) and instances describing nearby POIs (DBPedia and GeoNames). The resulting ROUTE linked dataset currently complies with all the Linked Data criteria and, therefore, constitutes a 5-star dataset (see Sect. 3.2.4). However, to enable public access and usage by different stakeholders it remains to be published as Linked Open Data to the LOD cloud.

TABLE 8 Links with other datasets.

Instances of ROUTE class	Type of link	Instances of external datasets	Links
schema:Administra-tiveArea	owl:sameAs	DBPedia	120
gtfs:Agency	owl:sameAs	DBPedia	3
dbpedia-owl:City	owl:sameAs	DBPedia	1
wgs84_pos:SpatialThing	owl:sameAs	GeoNames	15,956

§ 3.3.5 Publishing to the LOD Cloud

In order to enable the exploitation of the generated linked dataset by third-party stakeholders (e.g. policy makers, urban planners etc.), so that it could be used in urban modeling or other related applications, the final step is to publish it – along with its ontology and its external links – on the Web and more specifically to the LOD cloud. As mentioned in Sect. 3.3.1, the various source data that comprise the integrated dataset are provided under an open license, namely Creative Commons Attribution 3.0, which allows further adaptation, transformation, reuse, republication, and sharing. As a consequence, the generated Linked Data are also compliant with the Linked Open Data principles and can, therefore, be published to the LOD cloud. To this end, aside from the integrated

dataset, the domain ontology, and the external links, all documentations have to be made available online, prior to being featured in the LOD cloud (Radulovic et al., 2015).

§ 3.3.5.1 Ontology and RDF Dataset Publication on the Web

To make the integrated dataset publicly available and to allow different stakeholders take full advantage of its potential, the generated linked dataset and the corresponding ontology are published under a Creative Commons Attribution 4.0 license[21]. The online publication of the ontology[22] can facilitate the understanding of the concepts to which the source data are mapped and the relationships that enable their combination. Moreover, it could be reused partially or fully for mapping relevant data from different sources or domains, in the same way that the demonstrated dataset reuses concepts from external knowledge models.

In Sect. 3.2.3, it has been described that the real potential of integrated RDF datasets lies not so much in the online availability of the files per se, but rather in the possibility to query those data and extract useful (combined) knowledge. To this end, the generated triples, together with the external links, are first stored in an online RDF repository – namely, the OpenLink Virtuoso server[23] – to allow public access and retrieval. Following this, a dedicated SPARQL Endpoint[24] is set up for querying the integrated data.

Since the ROUTE dataset comprises a fusion of the nine different data sets, it enables the performance of more complex queries, as well as the retrieval of information that traverses the individual source data. For instance, besides simple queries such as "How to get from an origin (*route:StartPoint*) to a destination (*route:EndPoint*) and by which means of transport (*gtfs:RouteType*) within a specific period of time (*time:DateTimeInterval*)?", one can extract information about "How many stops (*gtfs:isStop*) exist within a specific area/bounding box (*schema:AdministrativeArea / gtfs:Shape*)?", or "Which stops (*gtfs:isStop*) nearby specific POIs (link to *dbpedia-owl:locationOf*) in an area (*schema:AdministrativeArea*) are accessible to disabled people (*gtfs:WheelchairBoardingInformation*)?".

21 https://creativecommons.org/licenses/by/4.0/

22 As mentioned in Sect. 3.3.2.3, the implemented ROUTE ontology is available online as a file at: https://route-owl.github.io/ontology.owl

23 http://virtuoso.openlinksw.com (Accessed on February 23, 2016).

24 https://route-owl.github.io/sparql (Accessed on February 23, 2016).

§ 3.3.5.2 Documentation Accessibility

Next to the online availability of the ontology and the linked dataset, ensuring public access to their respective documentations is particularly crucial, when it comes to increasing the reuse potential in modeling, planning, or decision-support applications. Although there are certain similarities to the logic governing API documentations for various software platforms, the main difference in the case of Linked Open Data is that the documentation accompanying ontologies and RDF datasets has to be readable not only by humans but also by machines. Following Semantic Web and LOD principles, it needs to be possible for both people and computational systems to understand the rationale behind semantically integrated data and the models governing the relationships between concepts.

In this respect, the human-readable documentation of the ROUTE ontology and dataset is semi-automatically generated, using the Wizard for Documenting Ontologies (Widoco)[25]. The outcome of this process is an HTML document that contains the description of the ontology, drawing on its machine-readable hierarchy and axioms. This semi-automatic process further enables on-the-fly completion of the documentation, when future modifications occur (e.g. by introducing new components to the ontology). In addition to this, the machine-readable documentation is generated by means of describing the dataset and the ontology in both DCAT[26] (Data Catalog vocabulary) and VoID[27] (Vocabulary of Interlinked Datasets) vocabularies (see Appendix B). Both of these descriptions comprise W3C standards. The former refers to an RDF vocabulary for describing integrated data and for enabling them to be discovered and processed by machines. The latter is also an RDF vocabulary for describing metadata of integrated datasets, as well as the links with external resources. VoID is also intended for increasing accessibility to the generated linked dataset. All aforementioned documentations are publicly-available online[28].

25 https://github.com/dgarijo/Widoco (Accessed on February 23, 2016).

26 https://www.w3.org/TR/vocab-dcat/ (Accessed on February 23, 2016).

27 https://www.w3.org/TR/void/ (Accessed on February 23, 2016).

28 The human-readable documentation of the ROUTE ontology and dataset is available at: http://osmosys.hyperbody.nl/files/ROUTE_doc. The machine-readable documentations in DCAT and VoID vocabularies are respectively available at: http://osmosys.hyperbody.nl/files/dcat and http://osmosys.hyperbody.nl/files/void, and are also available in Appendix B.

§ 3.3.5.3 Registration into an Urban Linked Data Catalog and Publication to the LOD Cloud

After publishing the human- and machine-readable documentations of both the ontology and the RDF dataset, the generated linked dataset can be published to the LOD cloud, so that other datasets can create new links with it. Although a 5-star linked dataset (the level reached in Sect. 3.3.4) already contains the combined knowledge from the various source data and the links with other semantically enriched datasets, its exploitation by a wider community of stakeholders is ensured only through its publication as LOD. In achieving this goal, it needs to secure (a) the way in which it will be discovered on the Web, and (b) the compliance with the LOD cloud requirements.

In general, the predominant mechanism for discovering documents on the Web is by means of a certain search engine. This discovery process is further facilitated by sitemaps, which describe the content and structure of web pages. In the context of the Semantic Web, specialized sitemaps can be used to inform both generic and data-oriented search engines about an online RDF dataset.

In the case of the example presented here, the sitemap4rdf[29] tool is used for automatically creating a sitemap of the published dataset. The entire set of URIs contained in the RDF data is extracted from the dedicated SPARQL Endpoint (see Sect. 3.3.5.1) and used for producing the XML sitemap document, which is then uploaded to both generic and data-oriented search engines. In addition to this, and in order for the dataset to be more accessible to city stakeholders, the linked dataset is registered into a specialized catalog, dedicated to urban linked data, namely the READY4SmartCities[30] catalog. This particular choice is driven by the domain-oriented nature of this catalog for Linked Data, which differentiates it from other more popular, yet rather generic, data catalogs.

Lastly, to further raise awareness about the newly generated dataset, a validation against the fulfillment of the LOD cloud publication requirements is necessary to be performed. To this end, the dedicated LOD cloud Validator[31] is used to assess the completeness of the generated dataset, so that it appears in the next version of the LOD cloud diagram[32].

29 http://lab.linkeddata.deri.ie/2010/sitemap4rdf/ (Accessed on February 24, 2016).

30 http://smartcity.linkeddata.es/datasets/ (Accessed on February 24, 2016).

31 http://validator.lod-cloud.net/ (Accessed on February 24, 2016).

32 http://lod-cloud.net (Accessed on February 24, 2016). It should be noted that the ROUTE dataset does not still appear on the LOD cloud diagram, as the last update of the latter has occurred on August 30, 2014, whereas ROUTE has been published on June 15, 2015.

The inclusion in the LOD cloud diagram increases the visibility potential of the linked dataset and, therefore, the possibility of other semantically enriched datasets to establish links with it.

§ 3.4 Summary and Conclusions

The growing availability of urban data from various sources opens up new perspectives in understanding aspects of city systems that have hitherto been difficult to study. However, the integration of heterogeneous data remains a significant challenge.

In addition to presenting various approaches to interoperability, this chapter addressed the above-mentioned challenge by designing a methodology for the transformation of heterogeneous urban data into multidimensional linked urban data. The methodology follows an ontology-based data integration approach and accommodates a variety of semantic (web) and linked data technologies. Each of the steps it comprises, was demonstrated through a use case, using real-world datasets from multiple sources. The use case illustrated how various collected sets of large-scale spatiotemporal data from three different sources, covering the entire public transport network of the city of Athens, Greece, can be fused together into a semantically rich dataset. It further presented how to establish links with external geo-data and how to eventually publish the resulting linked dataset to the LOD cloud.

The methodology comprises three distinct, yet interrelated, processes: (a) semantic integration of heterogeneous source data, (b) interlinkage with external datasets from different or relevant domains, and (c) publication as LOD to allow exploitation by third parties. Each of these stages resembles different levels of data openness, reusability, reproducibility, connectivity, and retrieval. Semantic integration may refer to the fusion of local data that can be either open or proprietary and that stem from different sources, though mostly adhering to a certain domain. The outcome of this process, which is an integrated dataset, can be further linked to other integrated datasets from different domains. Yet again, the resulting linked data can be either proprietary (and only be exploited within a group of stakeholders) or publicly distributed. Contrariwise, the publication of integrated datasets as LOD requires that the former can be freely retrieved, reused, republished, transformed, connected to other datasets, and exploited in various applications. As a result, the latter stage ensures the highest degree of openness, reproducibility, and reusability by a wider community of stakeholders.

This methodology can be replicated with relatively low effort and be applied (with minor adjustments) to different types of urban data, irrespective of the chosen sources.

Moreover, the fact that it is based on ontologies enables the semi-automatic iteration of the data mapping for any future updates of the source data, provided that the latter maintain their initial schemas. This can be beneficial for contemporary social urban data, which are characterized by very frequent update rates. Besides minimizing data redundancy and ensuring semantic interoperability, ontologies can also be used as a basis for querying and retrieving the resulting linked datasets, e.g. from SPARQL endpoints.

One of the major limitations encountered, especially with regard to the interlinkage process, is that the LOD cloud presently incorporates more generic datasets than domain-specific ones. Even though linked geospatial data are becoming increasingly available – such as GeoNames and Linked Geo Data (Stadler, Lehmann, Höffner, & Auer, 2012) – more specific integrated datasets pertinent to human mobility behavior, social activity in cities, or flows between urban systems are hardly available. This has hampered the creation of meaningful links with external datasets, thus limiting the number of potential interconnections.

Another important difficulty concerns the legal terms accompanying source data. Finding appropriate data for the applied example was tremendously difficult and the available options were very limited. Despite the fact that a growing number of organizations worldwide are providing their internal data as open data, there are only a few cases where the licenses clearly specify whether or not the data can be processed, republished, and reused in different applications. This is crucial when aiming to publish urban data as LOD, as it could hinder their applicability to other domains.

Nevertheless, if an increasing number of urban data referring to various facets of the city (e.g. social, economical, spatial etc.) are linked with one another, contemporary analytics, simulation, and decision-support systems can be highly benefited. As the analysis of urban dynamics requires the simultaneous consideration of various aspects of the urban environment, data integration and interlinkage becomes paramount. This would allow stakeholders to perform complex queries and extract knowledge that exceeds the context of the source data. The methodology that has been presented in this chapter could be helpful in this regard.

4 Designing and Implementing Tools for the Visual Exploration of Multidimensional Linked Urban Data[33]

§ 4.1 Introduction

The nature of the analysis of urban dynamics is inherently multidimensional, in the sense that it requires the simultaneous consideration of spatial, social, and temporal parameters. Recent developments in complexity theory have reinforced the notion of cities as complex systems (Bettencourt, 2013; Portugali, 2011; Schlapfer et al., 2014). Cities in fact incorporate a multiplicity of interrelated networks, operating at varied spatial and temporal scales. Examples include the physical structure of the urban fabric, land uses, transportation, organizations, infrastructure, social networks of people, social interactions and activities, among many others. In studying and analyzing these networks, the key challenge is to understand the various relationships between the elements that comprise them, not only in terms of how these relationships are structured, but also in terms of how they evolve over time. To achieve this, data from different sources need to be fused together. The previous chapter addressed the challenges pertaining to data integration, and introduced a comprehensive methodology for interlinking datasets from different domains, in order for multidimensional linked urban data to be generated that are more appropriate for the analysis of urban dynamics.

Nevertheless, the generation of integrated and linked urban data still remains a non-trivial task among urban planners, researchers, and policy makers. This is evident from the present scarcity of domain-specific urban datasets on the LOD cloud, as well as from the limited amount of ontologies related to cities (Benslimane, Leclercq,

33 This Chapter is largely based on the following publications:
Psyllidis, A. (2015). *Ontology-Based Data Integration from Heterogeneous Urban Systems: A Knowledge Representation Framework for Smart Cities*. In Proc.: 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015), Cambridge, MA, USA: MIT, pp. 240:1–21
Psyllidis, A. (2015). *OSMoSys: A Web Interface for Graph-Based RDF Data Visualization and Ontology Browsing*. In P. Cimiano, F. Frasincar, G.-J. Houben, & D. Schwabe (Eds.) (Vol. 9114). Switzerland: Springer International Publishing Switzerland, pp. 679–682

Savonnet, Terrasse, & Yétongnon, 2000; Falquet et al., 2011). This can be partly explained by the stakeholders' lack of familiarity with data integration and interlinkage technologies. As a consequence, the majority of urban data remain stored in disparate silos and any fusion of information across sectors is carried out in a labor-intensive, manual fashion.

As the available sources of data for cities increase and the necessity for interoperable systems becomes crucial, the adoption of ontology models and other semantic (web) technologies for data integration and interlinkage becomes pivotal. It is argued that city stakeholders need to be encouraged to engage in the design and development of ontologies, as well as to increasingly consume multidimensional linked data for analysis, planning, and decision-making purposes. At present, ontology engineering experts are, to a great extent, engaged in performing these tasks. However, they lack the domain knowledge required for creating essential correlations between systems.

Although the methodology that was developed in the previous chapter provides a means to integrate data from multiple sources, it requires one to be familiar with the formalisms of ontologies and other semantic technologies. Therefore, it is essential to provide tools that would enable and encourage city stakeholders to leverage the potential of integrated and linked data, and would make the communication between them and other experts less cumbersome.

To address this challenge, this chapter presents the design and implementation of a set of web-based tools for the visual representation of ontologies and multidimensional linked urban data. The tools provide graphical user interfaces for the visual representation, browsing, and interactive exploration of both ontologies and linked data. The use of different visualizations – in the form of interactive web documents and force-directed graphs – aim to support the adoption and consumption of linked urban data, without requiring extensive knowledge of the technology stack that underpins them. Therefore, the tools provide easy-to-use interfaces, accessible to a wide range of users, either experienced or amateur ones.

To further support the production of multidimensional linked urban data, an upper-level ontology is developed that formally describes and represents the relationships between the various elements of urban networks, pertinent to both the social and spatial sphere of urban systems. Individual datasets with heterogeneous attributes can be mapped to the aforementioned ontology and fused into a single dataset that combines the different attributes together.

The chapter first reviews existing domain ontologies pertinent to cities and planning, as well as related approaches to ontology visualization. The limitations of existing work set the requirements for the design of the proposed tools. Next, the chapter describes their architecture and the design of the upper-level ontology of urban networks. The latter, in

combination with the outputs of the previous chapter (i.e. the ROUTE ontology and the linked dataset), are used as benchmark tests for the tools. The chapter concludes with reflections on the potential and limitations of the presented set of tools.

§ 4.2 Related Work

§ 4.2.1 Modeling Urban Systems through Ontologies

The role of ontologies in semantically describing real-world entities and the relationships between them has been emphasized in the previous chapter. In the last decade, mainly stimulated by the diversity of available data sources, there has been a growing use of ontologies in several domains of knowledge, as a standard for the integration of various domain data. Unlike other scientific fields, there is a lack of domain ontologies that specifically cater to urban analysis and planning. The limited existing examples comprise ontologies that represent concepts pertinent to a specific facet of the urban environment (e.g. transportation networks, land uses, spatial geometry of the urban fabric etc.). Examples of ontologies combining more than two facets are rather scarce, to date.

In providing a more generic framework for urban information management, one of the first and few examples existing hitherto, is the one proposed by Benslimane et al (Benslimane et al., 2000). The developed framework does not establish a single domain ontology merging information from various urban networks together. Instead, it sets the foundations for a multi-layered modeling system, in which each layer corresponds to a specific domain ontology, configured by external experts (e.g. urban planners). These ontologies are then mapped to a top-level ontology that defines the relationships between the layers. However, in order for the framework to operate properly, there is a need for tools that would allow the various parties involved to collaborate in the generation and evaluation of each ontology representing a specific urban network, which are not developed in this particular work. In a more recent example, Bellini et al. (Bellini, Benigni, Billero, Nesi, & Rauch, 2014) introduce an ontological model for smart city services, which fuses concepts related to public administration, street features, POIs, and transport networks together with data from various sensor systems and time-related entities.

Montenegro et al. (Montenegro, Gomes, Urbano, & Duarte, 2012) introduce an ontology for land use planning. Having as a starting point the Land Base Classification Standards (LBCS) model, established by the American Planning Association, the ontology provides semantically annotated land use descriptions of spatial data, to allow enhanced integration and reuse possibilities across GIS systems. In semantically enriching 3D models with integrated information about the physical geometry of the urban structure, Métral et al. (Métral et al., 2009) present an ontology of the CityGML model, which is an OGC standard for storage and exchange of virtual 3D city models (Gröger, Kolbe, Czerwinski, & Nagel, 2008). Based on this, the ontology incorporates both geometrical and topological concepts of urban objects (e.g. vegetation, water elements, geometrical features of buildings, topological relations between spatial geometry objects etc.). In addition to this, a first instance of the Ontology of Urban Planning Processes (OUPP) is also demonstrated, yet only focusing on issues related to soft mobility.

A further collection of ontologies, specifically created for modeling systems related to the domains of urban planning and development, is presented in (Falquet et al., 2011). The demonstrated cases cover a wide range of domains, from urban mobility to urban morphology, with varying levels of completeness. In a more recent case, Poveda-Villalón et al. (Poveda-Villalón, García-Castro, & Gómez-Pérez, 2015) introduce a comprehensive and structured catalog that accumulates existing ontologies in domains related directly or indirectly to cities, ranging from energy to building geometry and air quality. Such collections can be particularly useful for discovering and reusing existing models in the ontology development process.

§ 4.2.2 Approaches to Ontology Visualization

The growing number of ontologies in various scientific fields has generated an equally increasing demand for visualization methods and tools. This necessity becomes even more significant as users with varying levels of expertise are progressively involved with the process of ontology development and evaluation. Visual representations of semantic models could also facilitate the work of ontology engineering experts, who are faced with the growing volume and complexity of the various components that comprise ontologies (i.e. classes, properties, instances, restrictions, axioms etc.). A variety of methods and tools for the visual representation of ontologies have been developed to date (Katifori, Halatsis, Lepouras, Vassilakis, & Giannopoulou, 2007). The visualization and interaction techniques used in each case, largely depend on the application area and the target user groups. Without intending to provide a comprehensive overview, this section focuses mainly on the presentation of work that closely relates to the tools described later in this chapter.

In general, the most frequently used methods in ontology visualization include 2D or 3D graphs of various layouts, tree diagrams, nested sets, UML diagrams, and knowledge graphs, to name a few. Graph-based visualizations are particularly interesting to the focus of this chapter. This type of visual representation is also consistent with the nature of OWL ontologies, which are in fact extensions of RDF graphs and, hence, RDF triples. As mentioned in the previous Chapter (Sect. 3.2.3), graph nodes can represent the various subjects and objects (classes and instances), whereas connecting lines can illustrate the predicates (object and data properties) of the triples.

A widely used tool for graph-based ontology visualization is OntoGraf (S. Falconer, 2010). The tool mainly owes its popularity to the Protégé ontology editor (Stanford University), as it comprises a plugin for the platform. Although the tool offers many possibilities for visualizing several ontology components (e.g. classes, sub-classes, individuals, domain/range object properties etc.), the fact that it is dependent to a sophisticated ontology editor constitutes a major obstacle for non-expert users. Similar limitations apply to the example of OWLPropViz (Wachsmann, 2008). The two aforementioned examples refer to 2D graph visualizations. Conversely, the Onto3DViz (Guo & Chan, 2010) is an attempt to represent ontologies in a three-dimensional graph. Despite the advantage of being a standalone application, rather than a plugin for a platform, it incorporates a very limited amount of ontology components and does not give the possibility to interactively search specific entities. Besides desktop-based applications and plugins, an example of a web-based service for ontology visualization is FlexViz, presented in (S. M. Falconer, Callendar, & Storey, 2010). In addition to allowing online access, it also provides various layout, navigation, search, and export functions, which make it an interesting case of graph-based ontology viewer.

Alternative graph-based approaches that further focus on the visualization of RDF data have also been developed recently. An early example is RelFinder (Heim, Hellmann, Lehmann, Lohmann, & Stegemann, 2009) for interactively searching and browsing instances of ontology classes and their relationships. In a similar way, LODWheel (Stuhr, Roman, & Norheim, 2011) focuses mainly on RDF data visualization, yet it is only standardized for datasets stemming from DBPedia. A more recent and advanced example in this regard is LodLive (Camarda, Mazzini, & Antonuccio, 2012), which allows data that are retrieved from SPARQL endpoints to be visualized in a dynamic graph-based fashion, while also supporting various types of ontology components. Lastly, one of the latest examples hereof, which has been developed later than the set of tools presented in this Chapter, is the Visual Notation for OWL Ontologies (VOWL) (Lohmann, Negru, Haag, & Ertl, 2016). The latter offers visual representations of ontologies, based on a force-directed graph layout covering several ontology components, and is implemented as both a plugin for the Protégé ontology editor and a web-based interface. An overview of the aforementioned tools for ontology and RDF data visualization is presented in Table 9.

TABLE 9 Tools for ontology (OWL) and structured data (RDF) visualization.

Tool	Dependency	Web/Desk-top-based	Visualization	OWL/RDF	Filtering/ Editing	Interactive Navigation	Source
OntoGraf	Protégé	Desk-top-based	2D graphs	OWL	Filtering & Editing	Yes	S.Falconer (2010)
OWLPropViz	Protégé	Desk-top-based	2D graphs	OWL	Filtering & Editing	Yes	Wachsmann (2008)
Onto3DViz	Standalone	Desk-top-based	3D graphs	OWL	None	No	Guo & Chan (2010)
FlexViz	Standalone	Web-based	2D graphs	OWL	Filtering & Editing	Yes	Falconer et al. (2010)
RelFinder	Standalone	Web-based	2D graphs	OWL & RDF	Filtering & Editing	Yes	Heim et al. (2009)
LODWheel	Standalone	Web-based	2D graphs	RDF	Filtering & Editing	Yes	Stuhr et al. (2011)
LodLive	Standalone	Web-based	2D graphs	OWL & RDF	Filtering & Editing	Yes	Camarda et al. (2012)
VOWL	Standalone & Protégé plugin	Web & Desk-top-based	2D graphs	OWL	Filtering & Editing	Yes	Lohmann et al. (2016)

§ 4.3 A Framework of Web-Based Tools for the Visual Exploration of Ontologies and Multidimensional Linked Urban Data

The nature of the analysis of urban dynamics is inherently multidimensional, in the sense that it requires the simultaneous consideration of spatial, social, and temporal parameters. To achieve this, data from different sources need to be fused together. The previous chapter addressed the challenges pertaining to data integration, and introduced a comprehensive methodology for interlinking datasets from different domains, in order for multidimensional linked urban data to be generated that are more appropriate for the analysis of urban dynamics. Here, the focus is on providing mechanisms to facilitate the understanding and consumption of multidimensional linked urban data for analysis, planning, and decision-making purposes.

To this end, the *OSMoSys* framework[34] is presented, comprising a set of web-based tools for the interactive visualization and exploration of ontologies and multidimensional linked urban data. The goal of the proposed framework is to potentially facilitate the consumption and employment of linked urban data in city analytics. In achieving this, *OSMoSys* consists of an interface for the visualization of RDF data and OWL ontologies, using a force-directed graph layout, as well as of an ontology browser for interactive navigation through the hierarchy of classes, properties, and individuals, using a multi-pane user interface (UI). In addition, *OSMoSys* is supplemented by an upper-level ontology that describes the different networks in cities and relationships between their elements, based on relevant established standards and roadmaps. The ontology can in turn be used as a reference framework in domain-specific ontologies that model resources pertinent to a particular facet of the urban environment. Moreover, *OSMoSys* supports uploading of custom ontologies and integrated or linked datasets, so that non-experienced users understand their structure (through visual exploration), evaluate their completeness, and potentially exploit them in other applications (Figure 8).

Unlike existing tools (see Sect. 4.2.2) that depend on specialized software or require installation, the proposed tools are fully accessible through the Web and rely on open-source technology. The following paragraphs describe, first, the technology stack used in *OSMoSys* and, then, its various components.

34 The OSMoSys framework is available online at: http://osmosys.hyperbody.nl

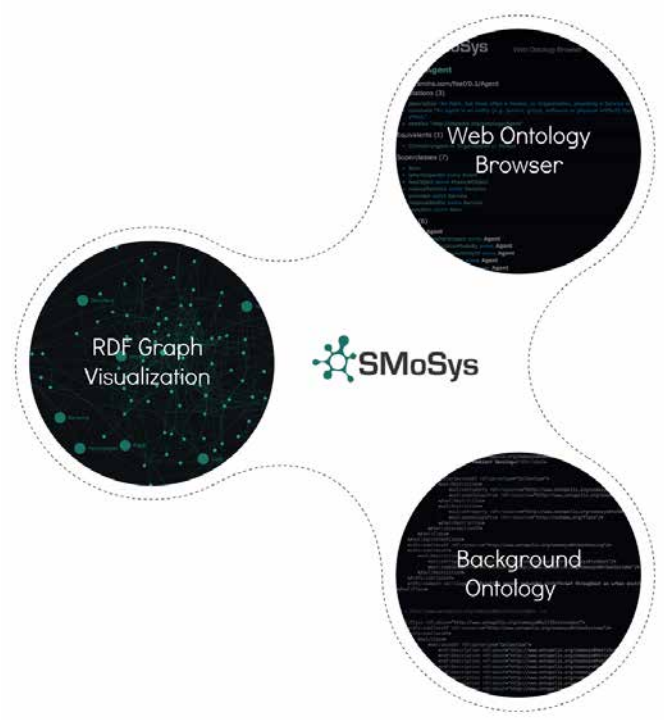


FIGURE 8 Components of the OSMoSys framework.

§ 4.3.1 Technology Stack

A key requirement in the development of the proposed framework is that *OSMoSys* is accessible to everyone. Moreover, it should provide an easy-to-use and intuitive interface, so that it stimulates mainly non-experienced users to benefit from its features, while also making it possible for other users to further modify and extend its functionality. In line with these requirements, *OSMoSys* makes use solely of open software and programming languages. Additionally, it is implemented as a web-based platform, in an attempt to overcome the limitations that characterize plugins, especially in terms of accessibility.

The overarching framework of *OSMoSys* is implemented using JavaScript, HTML5, CSS3, jQuery[35], and to a lesser extent PHP[36]. JavaScript is specifically chosen for its compatibility with all contemporary web browsers, without the need for additional plugins. The graph-based RDF/OWL visualization interface uses the Sigma.js JavaScript library[37], which is dedicated particularly to graph drawing. In generating the force-directed graph layout, the Yifan Hu multilevel algorithm is employed (Hu, 2005). The rationale of this algorithm lies in the dynamic distribution of graph nodes, according to the conceptual proximity of the concepts they represent. The Web Ontology Browser is implemented using JavaScript, OWLDoc[38], HTML5, and CSS3. The upper-level ontology has been created using the Protégé ontology editor. *OSMoSys* receives as input RDF data – either retrieved from SPARQL endpoints or uploaded directly to the platform from local files – and OWL ontologies, which are converted into JSON format prior to being visualized, so as to be easily integrated into the JavaScript library (Table 10). The converted datasets and ontologies are then stored in a SQL Server database. Both graph visualization and web ontology browser are created from the JSON files at runtime.

TABLE 10 OSMoSys – Technology stack.

Component	Programming language / Software
Overall *OSMoSys* framework	JavaScript, HTML5, CSS3, jQuery, PHP
RDF/OWL graph visualization	JavaScript (Sigma.js library), HTML5, CSS3, jQuery
Web Ontology Browser (WOB)	JavaScript, OWLDoc, HTML5, CSS3
OSMoSys upper-level ontology	Protégé

§ 4.3.2 Interactive Graph-Based Visualization of RDF Data and OWL Ontologies

OSMoSys visualizes integrated and linked datasets, as well as their underlying ontologies, as networks of nodes and links (or edges), using a force-directed graph layout. In the case of ontologies, classes (either super- or sub-classes) and their instances are illustrated as nodes, whereas properties (object, data, and annotation

35 https://jquery.com. Accessed March 10, 2016.

36 http://php.net. Accessed March 10, 2016.

37 http://sigmajs.org. Accessed March 10, 2016.

38 https://github.com/co-ode-owl-plugins/owldoc. Accessed March 10, 2016.

ones), which in fact correspond to the relationships between the classes, are depicted as edges connecting the nodes together. Similarly, in the case of integrated and linked datasets, subjects and objects of an RDF statement (see Sect. 3.2.3) are displayed as nodes, whereas predicates (relationships) as edges (Figure 9). This is also in agreement with the RDF graph notation (Schreiber & Raimond, 2014). In this way, a more intuitive and straightforward illustration of the relationships between various local data, or between real-world objects of the urban fabric, is provided, as opposed to the intricate, machine-oriented RDF serializations.

To further increase readability, the graph incorporates varied node sizes. These variations may indicate either the position of the class in the ontology hierarchy (top-classes appear larger than sub-classes) or the amount of instances that belong to it (larger node size indicates a larger number of class instances). In the case of RDF datasets, the size of nodes is proportional to the amount of established links with other nodes. Therefore, a data instance with multiple links to other instances will be illustrated as a node of a larger size. In this way, concepts, objects, or data elements of greater significance (or, at least, centrality) are instantly recognized. Besides node variation, the force-directed layout algorithm clusters those nodes that either have a large number of links to other nodes or contain several instances, and places them more centrally, as opposed to smaller nodes, which are placed at the outskirts of the graph.

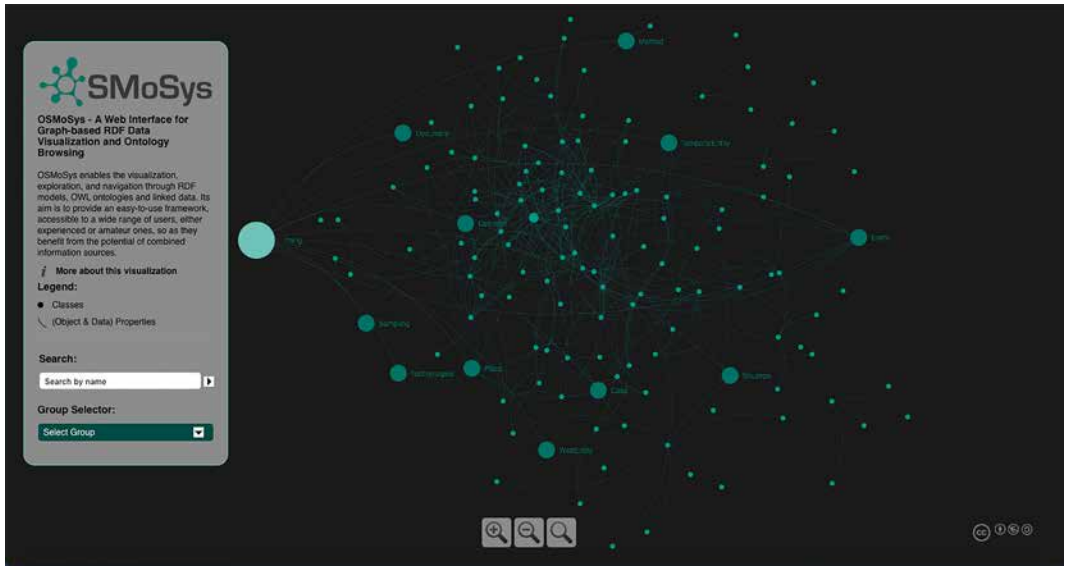


FIGURE 9 General overview of the web-based interface for ontology and RDF data visualization.

TOC

In addition to the above, the *OSMoSys* graph supports zooming. In the initial state of the graph, users have an overview of the entire network. As they zoom in, the graph displays different levels of detail, gradually displaying the names of classes or instances, or sometimes depending on the dataset, their complete URIs. This process is generally called *semantic zooming*. This allows users to understand the structure of the entire network, as well as to focus on specific objects and relationships. In addition to zooming, the interactive graph also supports panning, so that users easily navigate the graph by using their mouse or trackpad (Figure 10).

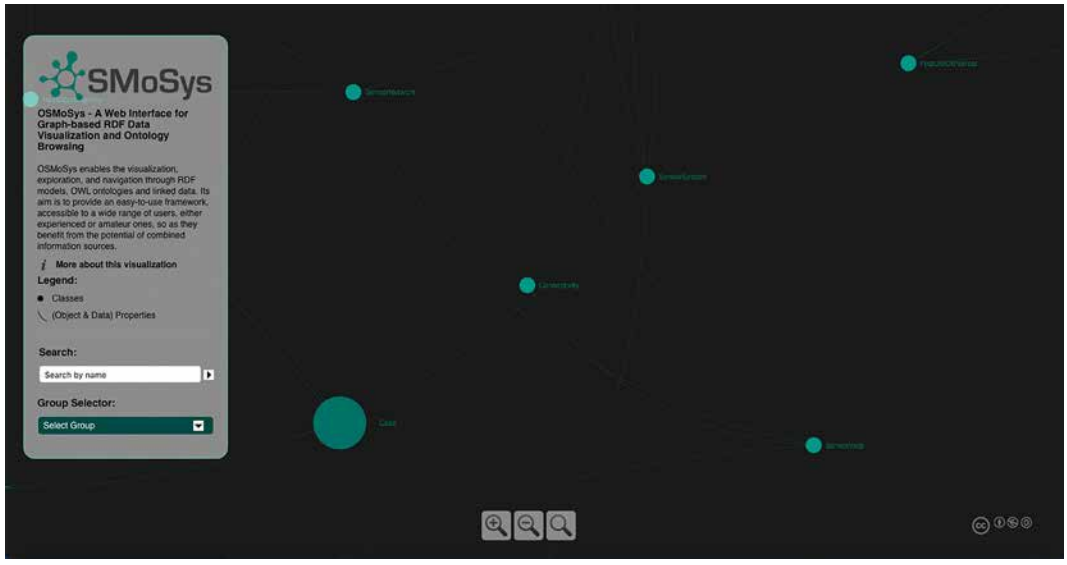


FIGURE 10 Semantic zooming function.

However, a common limitation of network visualizations is that they can become quite messy as the number of nodes increases. This may essentially hinder the readability and, hence, the visual discovery of correlation patterns between the nodes (Michael Batty, 2013b). To overcome this obstacle, the *OSMoSys* graph implements a set of additional features. In particular, when a user hovers over a node, its corresponding label as well as the nodes that are connected with it are highlighted, irrespective of the zoom level (Figure 11). Moreover, when a particular node is selected, the graph shows an isolated view of the chosen object and its links. Thus, the global network minimizes into a local network, consisting only of a certain chosen entity and the elements that are directly linked with it (Figure 12). When one of the other linked nodes is selected, its corresponding local network of relations appears. Users can switch between the global and the local structure of the network at any time, by using the corresponding options provided by the UI. In addition to the above, the selection of a highlighted

node activates a pop-up sidebar, containing detailed information about its attributes (e.g. the name of a class or instance, the description of an object, URIs etc.). Moreover, it further shows a list of all nodes/objects that are linked with the selected node. Therefore, aside from directly selecting nodes on the graph, a user is also enabled to navigate the different nodes by using the list included in the sidebar (Figure 13).

To further increase the node discovery potential and the readability of complex graphs, *OSMoSys* implements search and grouping functions. Thereby, a user is given the possibility to perform keyword search, without having to know necessarily the full name of a class or instance. The search component allows users to enter in the corresponding field at least three letters of a class or instance in focus, and subsequently all relevant results are listed right below the search field. In this way, a user may select any of the provided search results (in the case that there are multiple options) and, thereby, focus on the certain node and its immediate links. In addition to the search component, users are also enabled to group similar nodes of a certain type together. For instance, one is given the possibility to group together the entire set of super-classes included in an ontology hierarchy or those instances that contain owl:sameAs links to external data resources. Since both search and group functions are used for the discovery and selection of certain nodes and their relationships, they also activate the pop-up sidebar, mentioned in the previous paragraph (Figure 13). In the case of ontologies, where nodes represent classes or instances, the sidebar further contains links of the selected class or instance to the web ontology browser, which is described in the following section.

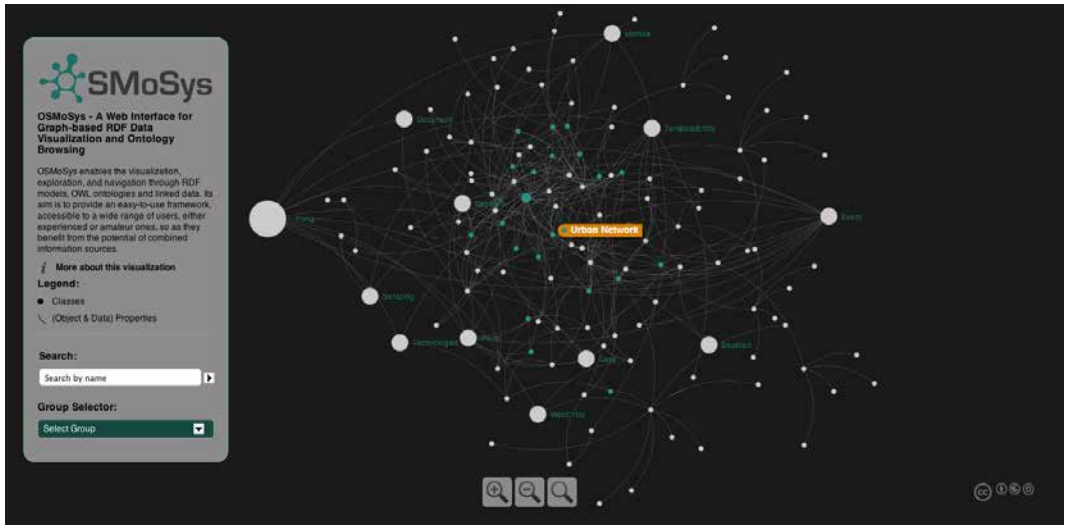


FIGURE 11 Highlighted node label on mouse over.

TOC

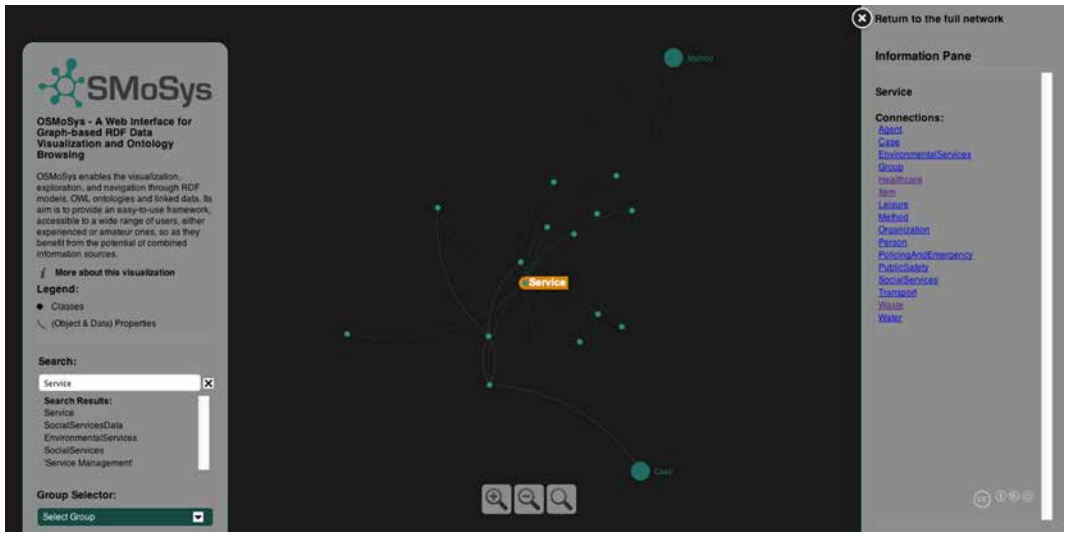


FIGURE 12 Isolated view of a selected node (i.e. class or data record) and its immediate links.

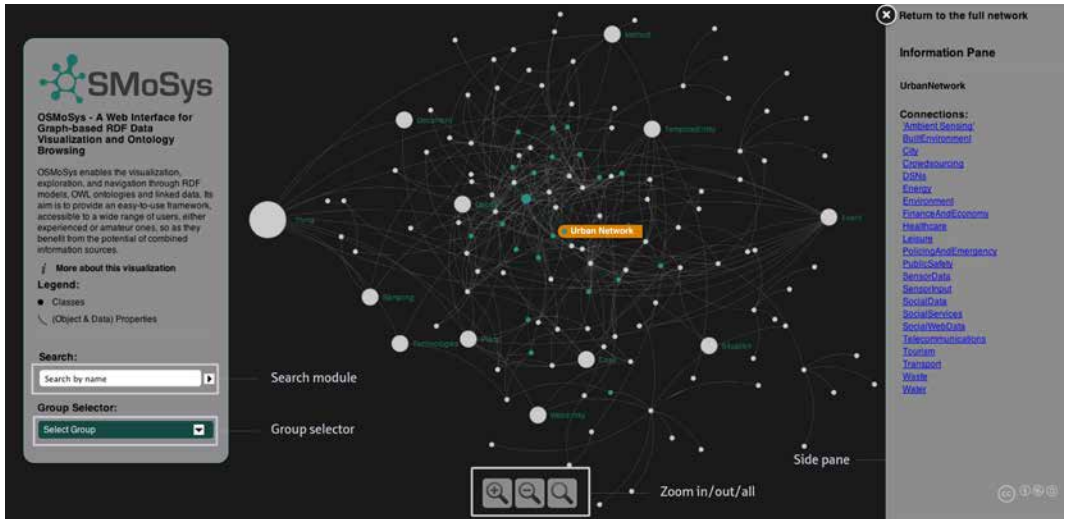


FIGURE 13 Side pane, zoom controls, "search" and "group" features of the visualization interface.

§ 4.3.3 Web Ontology Browser

A key component of the urban data integration and interlinkage procedure is the development of domain ontologies. The latter are used for modeling the concepts and objects of a particular domain (e.g. social networks, public transport networks etc.) and for mapping data from heterogeneous sources to these concepts, as described in the previous Chapter. The process of designing, developing, implementing, and evaluating ontologies, requires the involvement of experts in the domain that is modeled, to reach consensus on the way the various concepts and their relationships are described and represented. Having an understanding of the knowledge model is crucial in establishing links between sources within and across domains, as well as in making sense of how different objects or systems relate to one another. However, ontology navigation and evaluation by domain experts are non-trivial tasks. As the majority of ontologies are presently designed and implemented by ontology engineers, domain experts are faced with the task of evaluating the developed models for their completeness, conciseness, and intelligibility. To date, this process demands familiarity with ontology editing platforms, thereby hampering the involvement of and collaboration between non-experienced users.

In overcoming this obstacle, the *OSMoSys* framework incorporates – in addition to the interactive visualization interface – a web ontology browser (WOB) for navigating the components of an ontology and their entire set of metadata (i.e. annotations, object and data properties, individuals, descriptions, URIs, namespaces etc.), using a multi-pane layout (Figure 14). Besides the layout, the main difference between the WOB and the interactive visualization interface, is that the former provides a complete overview of all the components, properties and annotations comprising an ontology that is uploaded to the system. Unlike the graph-based visualization component, which also supports the representation of integrated and interlinked datasets, the WOB solely caters to ontologies. However, for users who aim to visualize an integrated RDF dataset, it is also possible to upload the domain ontology, to which the source data are mapped. Thus, the WOB is complementary to the interactive visualization component of *OSMoSys*. Having the entire set of ontology modules documented in the WOB, it prevents the network visualization from becoming cluttered. At present, the WOB does not support editing, as it is mainly intended for ontology evaluation and exploration by domain experts, who may have limited, if any, knowledge of ontology engineering. As is the case with the entire *OSMoSys* framework, the WOB does not require any plugin installation and is accessible by virtually all modern browsers.

The layout of the WOB is organized into three panes, each one providing different navigation possibilities and views of the ontology. More specifically, the upper-left pane lists the different ontology entities into groups of classes, object, data, and annotation properties, individuals, and data types. When any of the the aforementioned groups is

selected, a list containing all the entities corresponding to the selected group appears in the lower-left pane of the layout. The various entities are listed in alphabetical order. Users are given the possibility to return back to the general overview of the ontology at any time, by selecting the corresponding option in the upper-left pane. The main pane, covering the major part of the layout, accommodates the entire set of semantics, descriptions, and annotations of a selected entity (e.g. a class), in addition to its relationships with other ontology modules. Users can interactively browse through the various entities (i.e. classes, object and data properties etc.) and explore relationships between concepts, either through the side-pane indexes or by directly clicking on any term included in the main pane. Therefore, users with different levels of expertise can explore in detail and evaluate a given ontology and the domain it models. Moreover, in the case where some of the concepts are directly imported from external ontologies or are aligned with terms of external structured vocabularies, the WOB provides hyperlinks to the URIs of these resources or the documents describing them. Thereby, it could assist in discovering new interrelations, supporting the fundamental idea behind the Web of Data, as well as the generation process of multidimensional linked urban data.

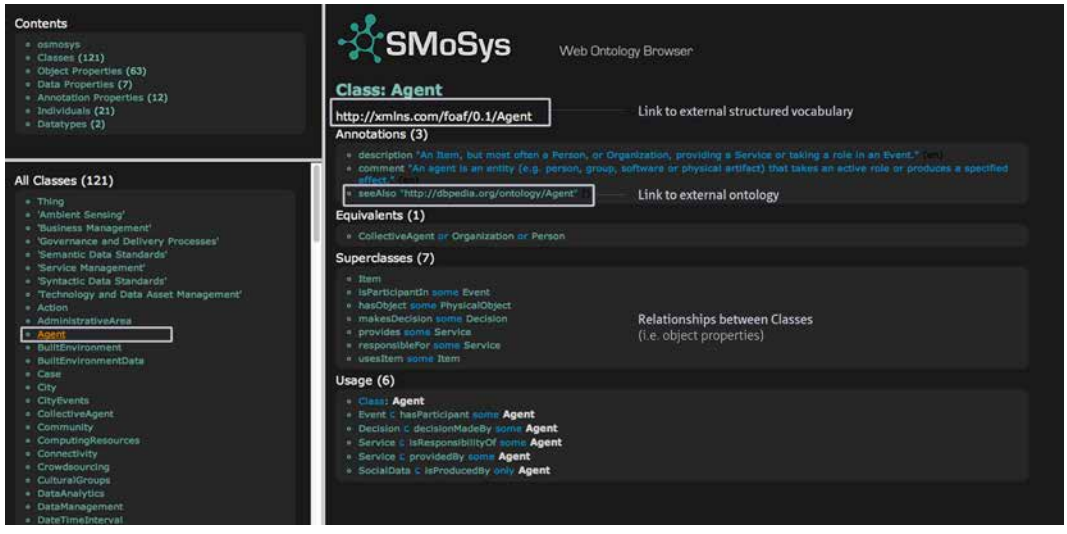


FIGURE 14 Interface and features of the Web Ontology Browser (WOB).

§ 4.3.4 Developing an Ontology of Urban Networks

§ 4.3.4.1 Requirements and scope definition

Besides the diversity of data sources about cities, it was previously described (see Sect. 4.2.1) that a variety of ontologies are being increasingly implemented to model and conceptualize different facets of urban systems. The result is a collection of several, mainly task-oriented, knowledge models that capture part of the relationships between the elements that comprise cities. It would therefore be useful to establish an overarching knowledge model, incorporating generic concepts about networks in cities and the ways their various facets could relate to one another.

Given this assumption, the *OSMoSys* framework implements an upper-level ontology of urban networks. Here, urban networks primarily refer to various types of networks between the elements comprising a city (also referred to as intra-urban networks), rather than networks between urban systems (also referred to as inter-urban networks). However, the knowledge model can easily be extended to also include networks at regional, national, or international scales. The ontology defines and axiomatizes general concepts about urban networks (e.g. social, spatial, economic, information networks etc.), the respective data sources and technology enablers, in addition to establishing a set of potential relationships between them. These are in fact indicators of relationships and are by no means comprehensive. In turn, domain-specific ontologies can reuse the generic concepts offered by the proposed knowledge model, to establish richer interrelations at different scales. Therefore, the goal of the proposed ontology is to serve as a generic, shared conceptualization for the coupling of urban networks to facilitate the generation of multidimensional linked urban data, and as a common foundation for relevant domain-specific ontologies.

§ 4.3.4.2 Ontology conceptualization

In conceptualizing the proposed ontology, an initial simplified model containing the main concepts and – part of – the relationships between them is first defined. To this end, relevant terms are extracted from recent European and American standards about cities, planning, urban governance and management, and integrated into the knowledge model. The aforementioned standards provide generally agreed definitions of concepts pertinent to these domains. The standards, vocabularies, and roadmaps from which the majority of terms are extracted are namely: the European Spatial Planning Observation Network – ESPON final report (ESPON, 2007), the British

Publicly Available Specifications (PAS) on the Smart Cities Vocabulary (Institution, 2014a), and on the Smart City Concept Model (Institution, 2014b), the Operational Implementation Plans of the European Innovation Partnership on Smart Cities and Communities (Communities, 2014), and the Smart Cities Readiness Guide (Council, 2014). The transfer of concepts included in standards to a knowledge model largely benefits from its reproducibility and extensibility, as well as from its modular structure.

Although the backbone of the proposed ontology is based on concepts pertinent to urban networks, its term hierarchy is not solely limited to them. Drawing on the fact that urban networks are embedded in an urban system, the ontology incorporates broad concepts (e.g. in the form of super-classes) that aim to provide context. Examples of these broad concepts are terms such as *Agent*, *Event*, *Item*, *Method*, *Process*, *Place*, *City*, *Information object*, *Physical object*, *Technology enabler*, *Temporal entity*, among others. These notions serve as the foundations, upon which the relationships between the concepts about urban networks are built. In other words, they assist in shaping the structure of the ontology hierarchy, described in the implementation section. The majority of these broad concepts are already described in external upper-level ontologies and are, therefore, reused in the proposed knowledge model. The ontology also takes into account the development of new types of (physical and digital) networks, such as sensor networks, geo-enabled social media, and LBSNs that are the major sources of social urban data, as discussed in Chapter 2. Thereby, it incorporates relevant concepts to model them semantically, either by reusing terms from external ontologies or by introducing new ones.

§ 4.3.4.3 Reuse of ontology modules

The reuse of modules that are already defined in implemented ontologies is key to the development of the proposed knowledge model. These ontologies and structured vocabularies are retrieved from Linked Open Vocabularies[39] and relevant ontology catalogs, such as the one described in (Poveda-Villalón et al., 2015). The selected strategy is to reuse specific ontology modules, instead of directly importing entire ontologies to the proposed knowledge model (with the exception of the *Time* ontology). The reused terms and statements are extended and enriched with additional attributes.

39 http://lov.okfn.org/dataset/lov/. Accessed March 14, 2016.

In particular, the *OSMoSys* ontology of urban networks reuses broad concepts, such as *Action*, *Method*, *CollectiveAgent*, and *InformationObject*, among others, from the *DUL* ontology[40] (Dolce + DnS Ultralite upper level ontology). To capture knowledge about sensor networks, modules from the *SSN* (Semantic Sensor Network) ontology are reused (Compton et al., 2012). Contributed concepts include the *ssn:Sensor*, *ssn:SensingDevice*, *ssn:Observation*, *ssn:Process*, *ssn:Stimulus*, *ssn:SensorInput*, and *ssn:SensorOutput*, which are appropriate for the representation of objects and features that are fundamental to sensor networks. However, broader concepts for describing the latter are introduced by the proposed ontology, as there is a lack of relevant terms in existing models. Concepts pertinent to transportation networks reuse ontology modules included in *OTN*[41] (Ontology of Transportation Networks). Conversely, the *CityGML*[42] ontology contributes concepts about the geometry and topology of spatial urban networks (e.g. *citygml:CityDistrict*, *citygml:Building* etc.). In the analysis of urban dynamics, besides understanding the spatial distribution of urban phenomena, it is equally important to understand how they evolve over time. Therefore, concepts pertinent to time intervals and temporal scales are reused from the *Time* ontology[43] (e.g. *time:TemporalEntity*, *time:Interval*, *time:Instant* etc.).

In addition to the aforementioned ontology modules, *OSMoSys* makes use of related terms from external structured vocabularies. More specifically, *dc* and *dct* (collections of metadata terms maintained by the Dublin Core Metadata Initiative) provide concepts such as *dct:Event* and *dc:PhysicalObject*. The *foaf* vocabulary is used for describing concepts pertinent to social networks, such as *foaf:Agent*, *foaf:Group*, *foaf:Organization*, *foaf:Person*, among others. Terms such as *Place* and *Administrative Area* are derived from the *schema* vocabulary. A selection of geographical features is derived from the *gml* vocabulary, an XML-based grammar and encoding standard of OGC, and *dbpedia-owl*. Additionally, a small number of data types and annotation properties are respectively retrieved from the *skos* and *vann* controlled vocabularies. Lastly, the proposed ontology complies with the following data modeling formalities: *owl*; *owl2xml*; *rdf*; *rdfs*; and *xsd* (Table 11).

40 http://www.loa-cnr.it/ontologies/DUL.owl. Accessed March 14, 2016.

41 http://www.pms.ifi.lmu.de/rewerse-wga1/otn/OTN.owl. Accessed March 14, 2016.

42 http://www.opengis.net/citygml/2.0/. Accessed March 14, 2016.

43 http://www.w3.org/2006/time. Accessed March 14, 2016.

TABLE 11 OSMoSys – Reuse of ontologies, structured vocabularies, and terms from standards.

	Ontology | Vocabulary	Prefix	URI (Namespace) / Source	Import
ONTOLOGIES	DUL (Dolce + DnS Ultralite top-level ontology)	DUL	http://www.loa-cnr.it/ontologies/DUL.owl#	partial
	SSN (Semantic Sensor Network)	ssn	http://purl.oclc.org/NET/ssnx/ssn#	partial
	CityGML	citygml	http://www.opengis.net/citygml/2.0/	partial
	Ontology of Transport Networks	otn	http://www.pms.ifi.lmu.de/rewerse-wga1/otn/OTN.owl	partial
	Time ontology	owl-time	http://www.w3.org/2006/time#	direct
STRUCTURED VOCABULARIES	Dbpedia-owl	dbpedia-owl	http://dbpedia.org/ontology#	partial
	Dublin Core	dc	http://purl.org/dc/elements/1.1/	partial
	Dublin Core Terms	dct	http://purl.org/dc/terms/#	partial
	Friend Of A Friend	foaf	http://xmlns.com/foaf/0.1	partial
	GML (Geography Mark-up Language vocab.)	gml	http://www.opengis.net/gml	partial
	Web Ontology Language vocabulary	owl	http://www.w3.org/2002/07/owl#	direct
	Web Ontology Language 2 vocabulary	owl2xml	http://www.w3.org/2006/12/owl2-xml#	direct
	Resource Description Framework	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	direct
	Resource Description Framework schema	rdfs	http://www.w3.org/2000/01/rdf-schema#	direct
	Schema vocabulary	schema	http://schema.org/#	partial
	SKOS (Simple Knowledge Organization System)	skos	http://www.w3.org/2004/02/skos/core#	partial
	VANN vocabulary	vann	http://purl.org/vocab/vann/	partial
	Extensible Markup Language vocabulary	xml	http://www.w3.org/XML/1998/namespace	direct
	Extensible Markup Language schema	xsd	http://www.w3.org/2001/XMLSchema#	direct

>>>

TABLE 11 OSMoSys – Reuse of ontologies, structured vocabularies, and terms from standards.

	Ontology | Vocabulary	Prefix	URI (Namespace) / Source	Import
STANDARDS	ESPON		ESPON (2007)	partial
	British Publicly Available Specifications (PAS)		British Standards Institution (2014a, b)	partial
	EIP on Smart Cities & Communities		EIP (2014)	partial
	Smart Cities Readiness Guide		Smart Cities Council (2014)	partial

§ 4.3.4.4 Ontology implementation

Following the integration of reused terms and modules, new conceptual structures (i.e. concepts and axioms) are introduced, to address the scope and requirements of the proposed ontology. Overall, the entire *OSMoSys* ontology conceptualization comprises 226 entities, classified into 121 classes, 82 object, data, and annotation properties, 23 individuals and datatypes, implemented with 736 axioms. The ontology is developed using the Protégé ontology editor and the OWL2-*EL* coding formalism (Table 12). The chosen resource naming strategy is that of hash URIs (see also Sect. 3.3.2.2), inasmuch as the ontology contains a quite small and rather stable set of resources. The URI domain used is: *http://osmosys.hyperbody.nl*, while the base URI for the ontology is: *http://osmosys.hyperbody.nl/files/Ontology*[44]. In accordance to the hash URI resource naming strategy, the generic URI paths for class names and properties respectively follow the pattern:*http://osmosys.hyperbody.nl/files/Ontology#<ClassName>* and *http://osmosys.hyperbody.nl/files/Ontology#<propertyName>*.

44 The *OSMoSys* ontology is available at this link, in both OWL and RDF formats.

TABLE 12 OSMoSys – Reuse of ontologies, structured vocabularies, and terms from standards.

	Metrics, types of correspondence & annotations	Counts / Annotations	Examples
ONTOLOGY METRICS	Total number of ontology entities	226	Classes, properties etc.
	Classes	121	*UrbanNetwork, City, Sensor, SocialMedia* etc.
	Object properties	63	*isPointOfInterest, isLocatedIn* etc.
	Data properties	7	*day, timezone* etc.
	Annotation properties	12	*language, prefix, creator* etc.
	Individuals	21	*"Twitter", "Instagram", "Foursquare"* etc.
	Datatypes	2	lat, long
AXIOMS	Axioms	736	Logical statements
	Subsumption correspondences (*subClassOf* axioms)	318	Municipality is a *subClassOf* City
	Mereology correspondences (*partOf* axioms)	21	End point is *partOf* route
	Assertion correspondences (*isA* axioms)	358	PointOfInterest *isA* Place
	Equivalence correspondences (*equivalentTo* axioms)	12	Day is *equivalentTo* temporal unit
	Disjointness correspondences (*disjointWith* axioms)	4	Instant is *disjointWith* proper interval
	Domain axioms	21	The domain of *hasParticipant* is *Event*
	Range axioms	2	The range of *isParticipantIn* property is *Event*
ANNOTATIONS	Namespace prefix	osmosys	
	URI	http://osmosys.hyperbody.nl/files/Ontology#	
	Languages	EN	
	Coding formalism	OWL2-*EL*	

The key component in a knowledge model of urban networks is the systems in which the networks are embedded. These systems are in fact the cities comprising the various types of networks, represented in the ontology by the *dbpedia-owl:City*[45] class.

45 Throughout the thesis, the various ontology components are preceded by a prefix denoting the ontology name or structured vocabulary, followed by the name of the component. Class names are capitalized and follow the CamelCase naming convention, while (object, data, annotation) properties start with a lowercase letter. Therefore, ontology components are identified as follows: [ontology prefix]:<ClassName> and [ontology prefix]:<propertyName>.

Not having found appropriate modules in existing ontologies to describe urban networks, a new *osmosys:UrbanNetwork* class is introduced. A subsumption correspondence is subsequently established, by means of an *rdfs:subClassOf* property, between the *osmosys:UrbanNetwork* and the *dbpedia-owl:City* (axiom: *osmosys:UrbanNetwork* is *rdfs:subClassOf* a *dbpedia-owl:City*). The generated axiom explicitly defines that cities are at their essence systems of networks, which are embedded in the physical structure of the city through various types of infrastructural components.

To better refine this statement, additional modules are introduced to the knowledge model. The – social and spatial – structure of urban systems generally comprises several types of elements with different attributes and behaviors that enable them to be related in some way to one another through networks. As these elements may refer to concepts as diverse as infrastructural components, organizations, groups of people, or individuals, the generic class *osmosys:Item* is introduced. The latter is thus a sub-class of *dbpedia-owl:City* and it encompasses all types of urban elements, by establishing subsumption correspondences (i.e. *rdfs:subClassOf* relationships) with them. In particular, the *osmosys:Item* class contains the sub-classes *foaf:Agent*, *DUL:InformationObject*, *dct:PhysicalObject*, and *osmosys:Service*. These classes address the variety of interrelated components in urban systems and, thereby, establish various forms of relationships with the *osmosys:UrbanNetwork* class.

To capture the variety of actors in social networks, the class *foaf:Agent* further incorporates modules about individuals, groups of people, and organizations (*foaf:Person*, *DUL:CollectiveAgent*, *DUL:Community*, *foaf:Group*, *foaf:Organization*) and introduces several object properties (i.e. relationships) between them. In turn, the *dct:PhysicalObject* class conceptualizes the various infrastructural components of cities and also encompasses modules to semantically represent sensor networks (e.g. *osmosys:SensorNetwork*, *ssn:Sensor*, *ssn:SensingDevice*, *osmosys:SensorSystem* etc.). The latter mainly stem from the SSN Ontology but are enriched with additional axioms. Yet, as the reused modules that are derived from the SSN Ontology solely refer to types of sensing devices, the knowledge model further introduces the *osmosys:HumanSensor* class, to represent the contemporary notion that people can also operate as "sensors" (e.g. by generating content on social media, by providing VGI, by being the main actors of crowdsourcing etc.). To prevent semantic discordance, it establishes an equivalence correspondence (*owl:equivalentClass*) with the *foaf:Person* class. Conversely, the *DUL:InformationObject* class is used for modeling the various immaterial objects (e.g. *ssn:SensorInput*, *osmosys:SensorOutput*, *osmosys:SocialMediaFeed* etc.) that comprise information flows and networks. The *osmosys:Service*, which completes the hierarchy of sub-classes that make up the generic *osmosys:Item* class, semantically represents the various functions of particular elements in an urban system.

The *osmosys:UrbanNetwork* class that was mentioned above, further specifies the different kinds of networks between the components that are semantically represented by the *osmosys:Item* class. It, therefore, incorporates classes about the various networks of streets and buildings, the transport, social, economic, telecommunications, energy, waste, and water networks, among other that are found in cities. Through the subsumption correspondence, the instances of the *osmosys:UrbanNetwork* class and all its sub-classes directly inherit the entire set of properties, attached to *dbpedia-owl:City*. In turn, a city – and hence all its sub-modules – is engulfed by the overarching concept of *schema:Place*. Further relationships are established with modules representing events in time (*dct:Event*), activity types (*DUL:Action*), POIs (*osmosys:PointOfInterest*), processes (*ssn:Process*), among others.

Besides the spatial attributes of networks, time is an intrinsic parameter in understanding interactions between urban elements that take place on networks. In capturing such temporal dynamics, the knowledge model incorporates several components about time-related entities. To this end, it imports the entire set of modules included in the Time Ontology. Examples of these components are, among others, the *time:Instant*, *time:Interval*, *time:DateTimeInterval*, under the overarching *time:TemporalEntity* class. The aforementioned classes are capable of capturing any type of temporal unit.

The above concepts allow relationships to be established between attributes of different dimensions (i.e. spatial, social, temporal) contained in the various source data that would be mapped to the ontology. For example, a pair of social contacts (e.g. derived from a social media platform or inferred from mobile phone data), with each individual being an instance of the *foaf:Agent* class, could be linked with a specific POI location (i.e. instance of the *osmosys:PointOfInterest* class) in a given city (i.e. instance of the *dbpedia-owl:City* class), at a given point in time (i.e. instance of the *time:DateTimeInterval* class). The semantic network of the OSMoSys ontology hierarchy is (partially) shown in Figure 15.

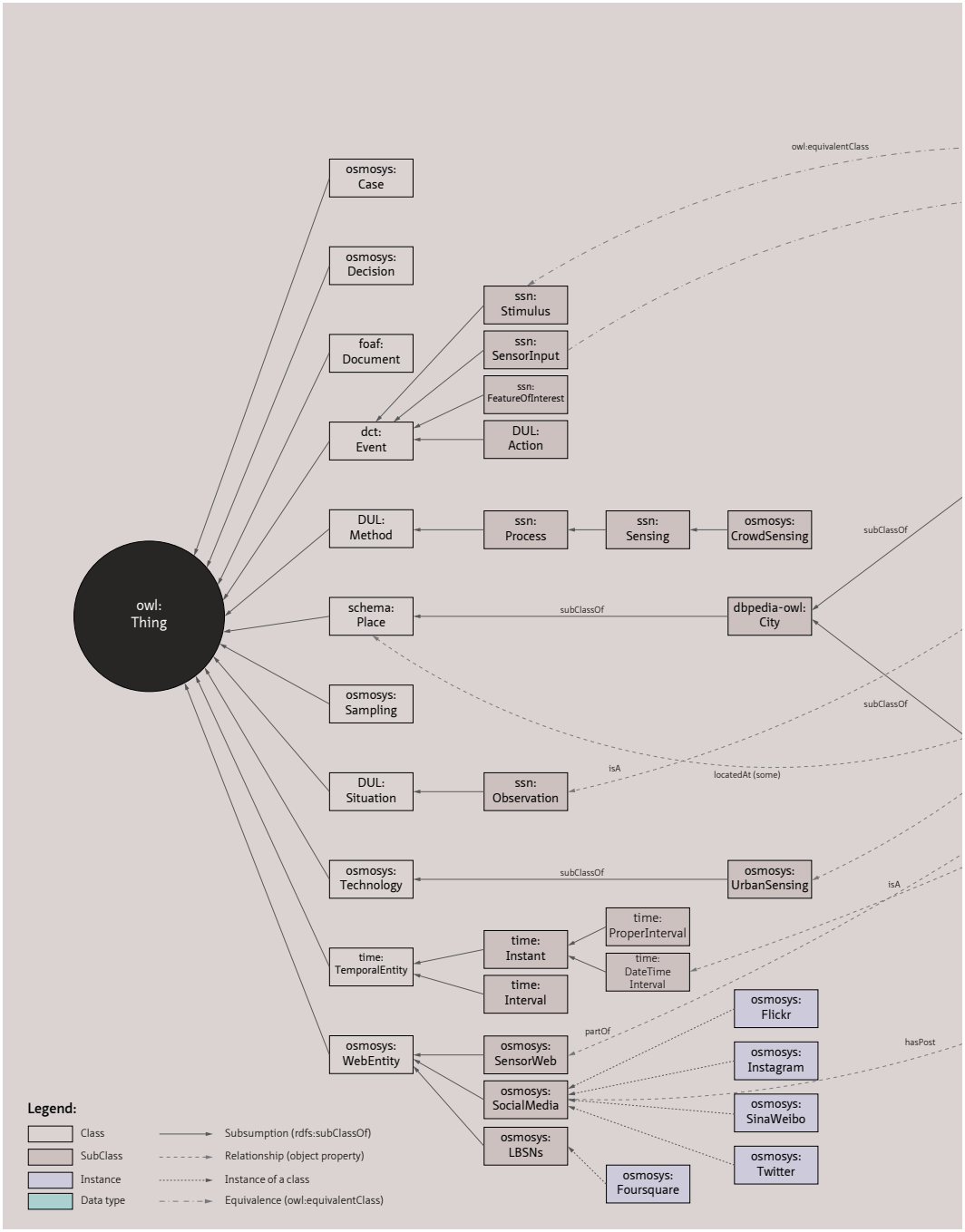
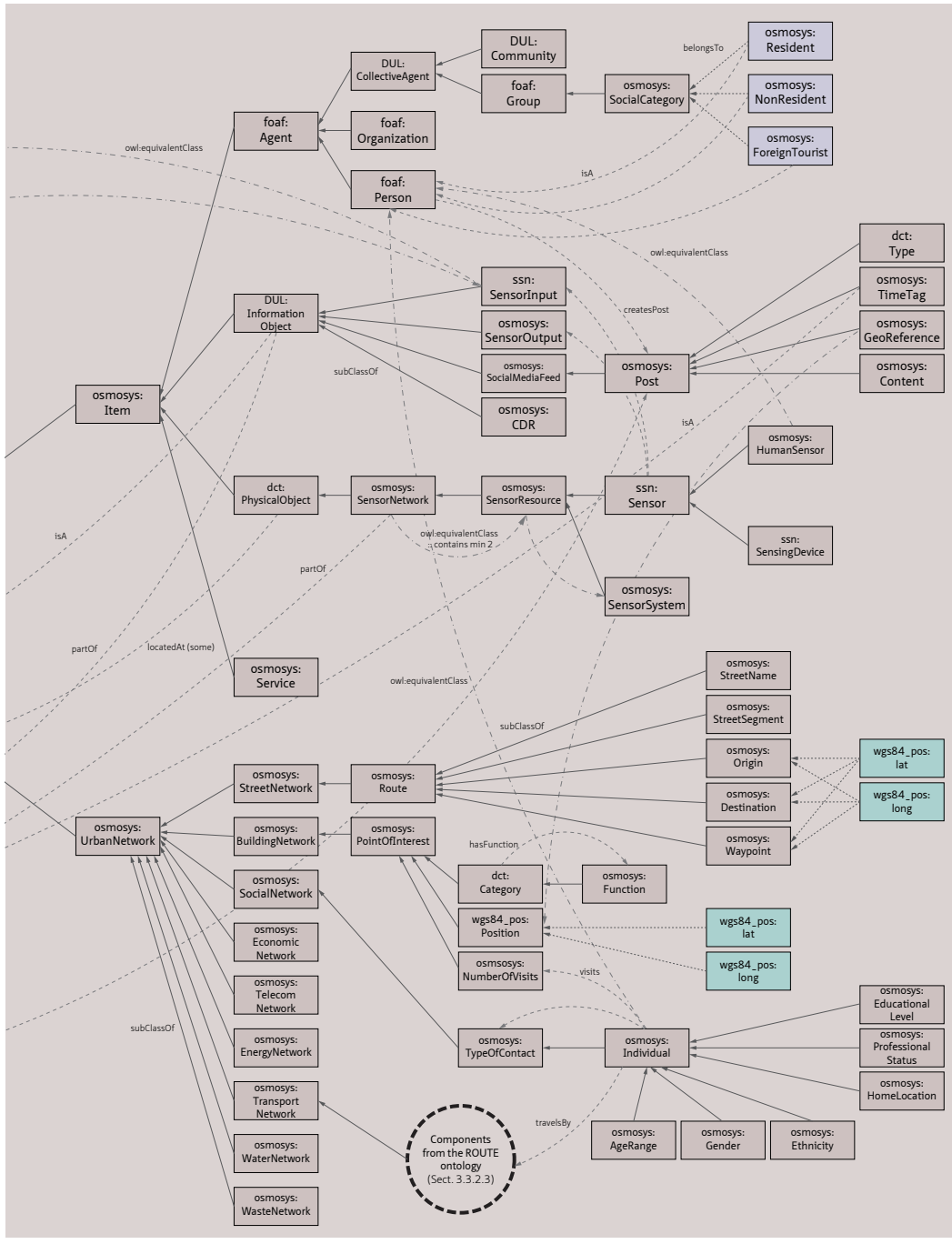


FIGURE 15 OSMoSys ontology of urban networks – Semantic network representation of class hierarchy and indicative relationships.

TOC


DUL: Community
DUL: CollectiveAgent
foaf: Group
foaf: Agent
foaf: Organization
foaf: Person
osmosys: SocialCategory
osmosys: Resident
osmosys: NonResident
osmosys: ForeignTourist
belongsTo
owl:equivalentClass
isA
dct: Type
osmosys: TimeTag
osmosys: GeoReference
osmosys: Content
ssn: SensorInput
DUL: Information Object
osmosys: SensorOutput
osmosys: SocialMediaFeed
osmosys: CDR
osmosys: Post
createsPost
subClassOf
osmosys: Item
osmosys: HumanSensor
dct: PhysicalObject
osmosys: SensorNetwork
osmosys: SensorResource
ssn: Sensor
ssn: SensingDevice
owl:equivalentClass
contains min 2
partOf
osmosys: SensorSystem
locatedAt (some)
osmosys: Service
osmosys: StreetName
osmosys: StreetSegment
osmosys: Origin
osmosys: Destination
osmosys: Waypoint
wgs84_pos: lat
wgs84_pos: long
osmosys: StreetNetwork
osmosys: Route
osmosys: UrbanNetwork
osmosys: BuildingNetwork
osmosys: PointOfInterest
hasFunction
dct: Category
osmosys: Function
wgs84_pos: Position
osmsosys: NumberOfVisits
visits
osmosys: SocialNetwork
osmosys: Economic Network
osmosys: Telecom Network
osmosys: EnergyNetwork
osmosys: Transport Network
osmosys: WaterNetwork
osmosys: WasteNetwork
osmosys: TypeOfContact
osmosys: Individual
osmosys: Educational Level
osmosys: Professional Status
osmosys: HomeLocation
travelsBy
osmosys: AgeRange
osmosys: Gender
osmosys: Ethnicity
Components from the ROUTE ontology (Sect. 3.3.2.3)

§ 4.3.4.5 Evaluation

From the various criteria and types of ontology evaluation (see also Sect. 3.3.2.3D), the implemented knowledge model is evaluated for logical consistency, as well as for conciseness. Given that the proposed ontology is in fact an upper-level ontology that can be further extended by third parties, it is not evaluated in terms of completeness. The chosen method for checking potential inconsistencies is carried out through several tests, using different types of reasoners, namely FaCT++, HermiT, and Pellet. The ontology proved to be fully consistent, in terms of the description logic of its axioms. In evaluating conciseness, the open-source OOPS! pitfall scanner[46] is used (see also Sect. 3.3.2.3). Multiple tests were carried out throughout the ontology development procedure, in order to minimize issues of critical importance. Detected missing annotations, mainly in the reused modules, are completed with relevant descriptions and comments. There are yet some cases, in which missing domains and/or ranges in properties are detected. However, this was a deliberate action, since the comprehensive definition of exact domains and ranges of properties can lead to undesired axiomatic classifications.

§ 4.4 Visually Exploring Ontologies and Multidimensional Linked Urban Data: A Benchmark Test

§ 4.4.1 Visualizing Ontologies

In examining the capacities of the *OSMoSys* framework, as regards the representation and visual exploration of ontologies and linked data, the developed tools are tested against reference data and knowledge models. In particular, the previously described *OSMoSys* ontology, as well as the *ROUTE* ontology, presented in Chapter 3, are used as benchmarks of both the interactive graph-based visualization and the WOB. In addition, an instance of the integrated ROUTE RDF dataset, focusing on the bus network of Athens is used as a benchmark of the developed interactive graph. With regard to the two aforementioned ontologies, emphasis is put on the identification of important classes in the graph (usually super-classes in the ontology hierarchy

46 http://oops.linkeddata.es

or classes with multiple connections, which in turn represent important real-world objects in the network) and the visual clarity of relationships established with other ontology modules. Conversely, in the case of the RDF dataset the focus is on the performance of the graph visualization, especially in terms of handling large-scale data, represented by thousands of nodes and edges.

Using the interactive graph tool to visually represent both *OSMoSys* and *ROUTE* ontologies, an immediately distinguishable element is the the node corresponding to the *owl:Thing* class. According to the OWL specification, *owl:Thing* is the default super-class containing all classes of an ontology (Bechhofer et al., 2004). Therefore, it has by definition the largest amount of connections and is, subsequently, illustrated as the largest node in the graph. In the case of *ROUTE*, *owl:Thing* is centrally placed (Figure 16), whereas in the *OSMoSys* graph it is located in the outskirts of the network visualization (Figure 9). This has to do with the size of each network, which in turn affects the layout of the force-directed graph and the placement of nodes in it. *ROUTE*, containing a smaller amount of classes (nodes) results in a radial layout, whereas *OSMoSys* has a more complex structure that is described later. Although *owl:Thing* does not have a functional role in the network, it serves as a starting point for the exploration of the represented concepts and their relationships.

In the complete overview, primary classes (i.e. the ones that are higher in the ontology hierarchy and contain a larger number of connections) are easily distinguishable through their node sizes. The larger the size of the node, the larger the number of connections it has with other components. These in turn refer to important real-world objects of the domain being modeled. In the case of *ROUTE* important entities of the transport network (e.g. route, route type, duration, transfer rule, stop etc.) surround the *owl:Thing* node (Figure 17), whereas in the case of *OSMoSys* the primary classes are located at the periphery of the graph, yet at a distance from *owl:Thing*. Entities with less established connections – hence represented by smaller nodes – are clustered mainly in the center of the graph, instead of being placed at the outskirts of it (as is the case with *ROUTE*), to keep its size more compact. The labels of the components representing the aforementioned entities (primary classes) are the only ones that are visible in the complete overview. This increases the readability of the complex graphs and it prevents them from becoming cluttered with entity labels. In revealing the labels of smaller nodes, a user may take advantage of the semantic zooming, hover, and pan functions (see Sect. 4.3.2). Property names (types of relationship) are also entirely missing from graph, again for readability reasons. Instead, the complete set of relationships is included in the WOB.

At present, edge thickness remains constant throughout the graph. This is because edges in the OSMoSys framework are used as indicators of interaction between two components, rather than as measures of the intensity of this interaction. In the case of ontologies, interactions between classes (nodes) also indicate correlations between

the members that belong to each one of the interconnected classes. In exploring the relationships between elements, the search and group functions play a crucial role, especially in cases of complex networks. Instead of looking for a particular class in the graph, one can type in the relevant term in the provided search box. For instance, a user looking for the class representing urban networks, can type in either "urban" or "network", and a list of relevant components appears right below. By choosing the term of interest – here, "UrbanNetwork" – the graph depicts an isolated view, containing only the components that have some kind of interaction with the osmosys:UrbanNetwork class. Accordingly, clicking on any of the connected nodes results in a new graph, comprising the objects that are directly connected with the chosen node (Figure 12). This can assist domain experts and ontology engineers to identify missing concepts from the knowledge model.

The platform currently assigns a single color to all nodes and edges of a visualized knowledge model. Some of the existing ontology visualizations distinguish sub-classes from super-classes, as well as imported concepts from newly introduced ones, by assigning different colors to nodes (Lohmann et al., 2016). Although this could be helpful for experienced ontology engineers, in terms of understanding the structure and role of ontology modules, it might seem quite redundant to users with little or no experience in ontology modeling. The latter, being the main target group of the *OSMoSys* framework, might find color and type variations confusing. Therefore, the only variation is in terms of node size, as explained in the previous paragraphs.

Users may switch between the interactive graph and the WOB, by using the information pane on the right side of the display, or the corresponding option in the WOB (Figure 14). The WOB displays the entire set of properties, annotations, individuals, and data types included in an ontology. Although the WOB does not incorporate any search function, the various components of an ontology are listed in alphabetical order, which could also be helpful when looking for a particular entity (e.g. class, object property, annotation etc.). By clicking on any term on the left, its full description appears in the main pane of the display. This in turn contains the URI of the selected entity and a hyperlink to it, the accompanying annotations, the list of its super-classes, as well as an overview of its usage patterns in the knowledge model. The aforementioned attributes are clustered into groups to facilitate readability. By selecting any of the entities included in these groups, the WOB displays the definitions, annotations, and relationships corresponding to this entity. Moreover, all types of correspondences (e.g. subsumption, equivalence, mereology, cardinality constraints, disjointness etc.) pertaining to an ontology component, are displayed in the WOB and are further distinguished by color coding.

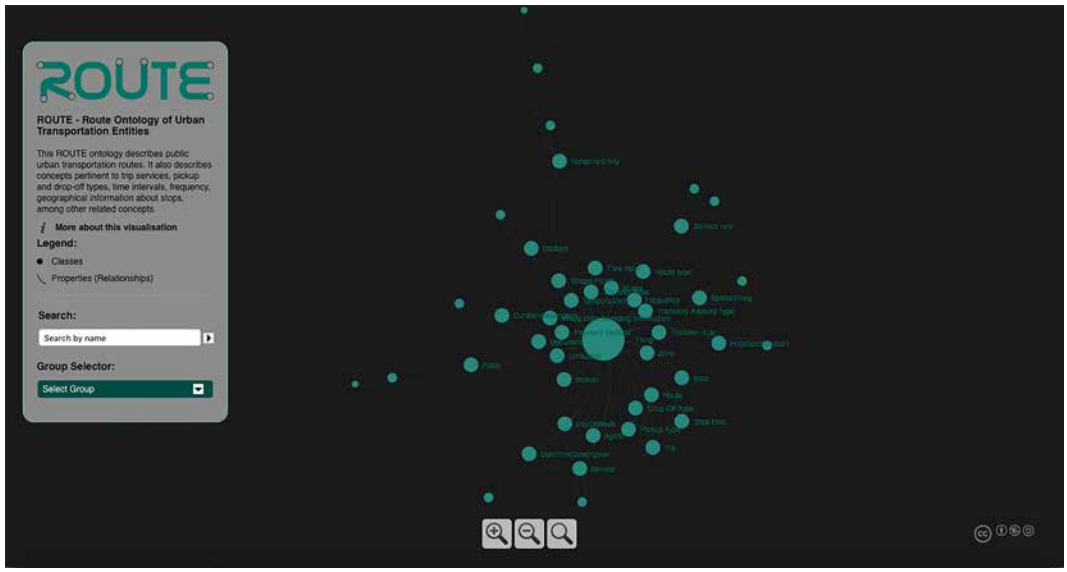


FIGURE 16 Visualization of the ROUTE ontology using the web-based interface of *OSMoSys* (Full network).

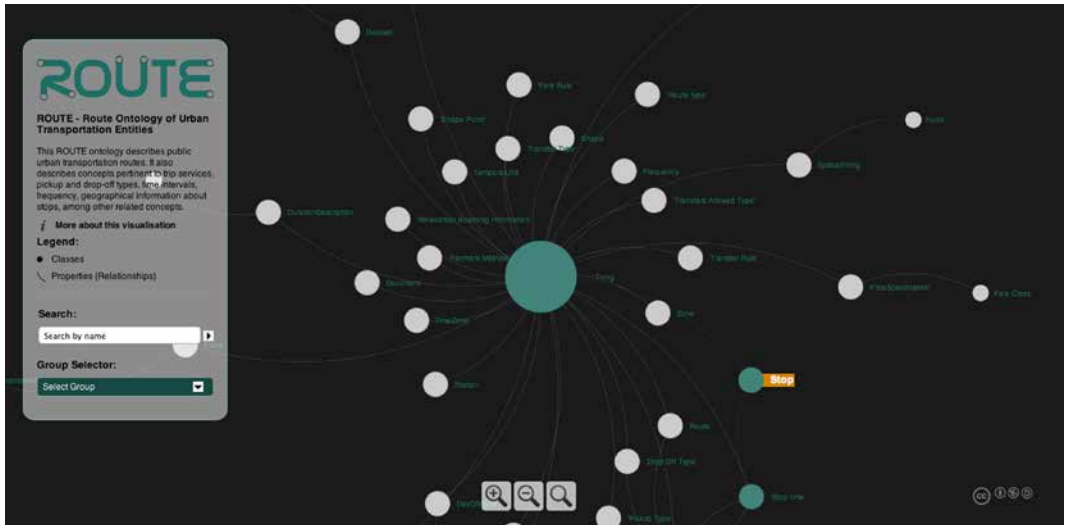


FIGURE 17 Zoomed view of the ROUTE ontology graph, highlighting the *gtfs:Stop* node (i.e. Class).

§ 4.4.2 Visually Exploring Multidimensional Linked Urban Data

Besides the two ontologies, the interactive graph is further tested for its ability to represent large-scale linked datasets. To this end, an instance of the wider ROUTE RDF dataset is used, concerning specifically the bus network of the city of Athens. The RDF triples are retrieved from the dedicated SPARQL endpoint of the ROUTE dataset, described in the previous Chapter (Sect. 3.3.4). The visualization of the RDF triples results in a directed graph, comprising 37,249 nodes (i.e. individuals referring to the start and endpoints of all bus lines, the entire set of intermediate stops, and their geo-coordinates), connected together with 47,900 edges (i.e. relationships between the nodes) (Figure 18). The force-directed graph has a radial layout, with less connected objects being placed at the periphery of the graph. The majority of the network components are instances of the *gtfs:Stop* class (part of the *ROUTE* ontology), which subsequently constitutes the largest node in the network and is centrally placed in the graph configuration. Due to the large size of the network, navigating the graph directly is cumbersome. Although pan, zoom, and hover functions perform well, without processing delay, the muddled visualization in the general overview makes it difficult to identify nodes of interest. In cases like this, entity filtering by means of search and group functions could prove to be beneficial.

To illustrate this, an example of entity exploration is carried out. Assuming that a user is interested in discovering the available bus stops around the area of Syntagma square, which is the largest square in Athens city center, the first step is to identify the nodes in the graph that correspond to relevant entities. To this end, the fastest method is to use the search field. By typing in a generic term, such as "Syntagma" (or "Σύνταγμα" in Greek, since the source data are provided in the Greek language), the system returns a number of related nodes. These are namely "Syntagma square", which refers solely to the square and its adjacent streets; "Syntagma", which also covers the area surrounding the square; "On Syntagma Square", which only refers to bus stops that are located on the square; and "Syntagma station", which denotes the closest metro station to the square.

Since the user is interested not only in the square itself, but also in the area surrounding it, the option that best fits the criteria is that of "*Syntagma*". By choosing this particular node, the graph display shows only the nodes that have a direct connection with the chosen entity (Figure 19). These in turn correspond to all the available bus stops in and around the square. By hovering over each node, the name of the bus stop along with its ID number are highlighted. The selection of any of these bus stops gives back additional information, by displaying the entities that it relates to. For instance, by choosing the node labelled "*On Vassilissis Sofias Avenue (Επί ΛΕΩΦ. ΒΑΣΙΛΙΣΣΗΣ ΣΟΦΙΑΣ)*", which refers to the bus stop at the intersection of the aforementioned avenue with the square, a set of additional nodes is revealed. These

denote bus lines traversing, yet not stopping at this particular point, nearby bus stops serviced by the previous bus lines, latitude and longitude coordinates of the selected bus stop, as well as the "*Syntagma*" node, which is the source entity. In a similar manner, the selection of any of the related nodes would reveal further connections. Moreover, the information pane on the right side of the interface contains links to the URI of each entity.

In cases where links have been established with external datasets, such as the DBPedia linked data, the sidebar further incorporates hyperlinks to those resources. In this way, the user can take advantage of the additional information, described in the external dataset, about the entity at hand. One example could refer to the nearby POIs, which are included in the DBPedia dataset. This would in turn enable users to understand the interconnection between components of the bus network and the network of POIs.

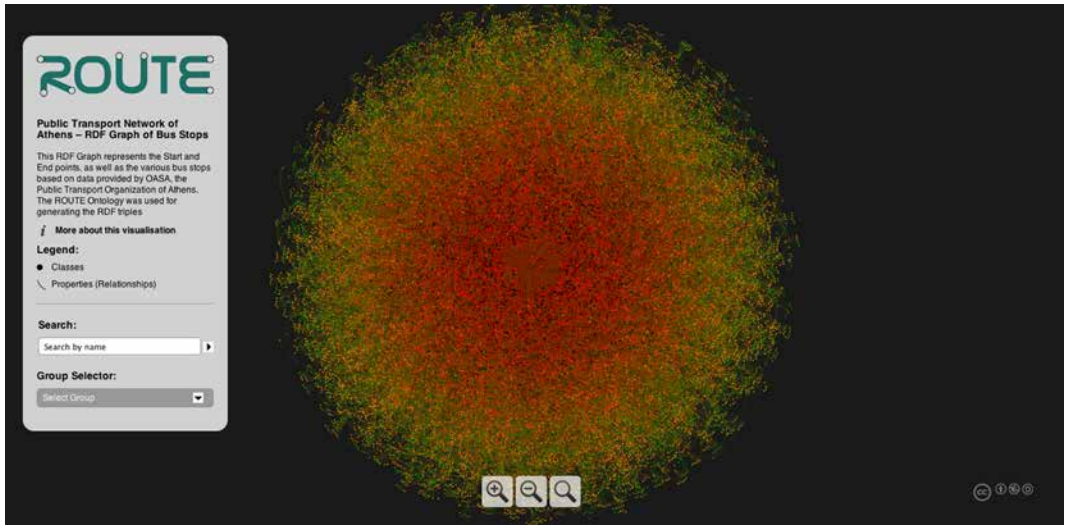


FIGURE 18 Graph visualization of an instance of the ROUTE RDF dataset. The large amount of triples results in a muddled visualization in the full network view.

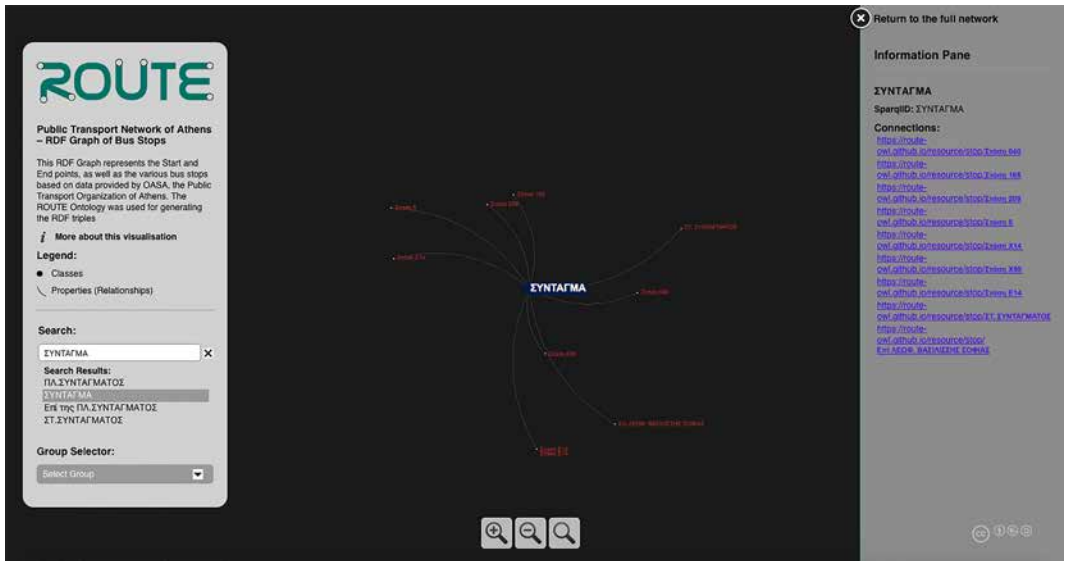


FIGURE 19 Using the search field to choose a specific node on the RDF graph, returns an isolated view containing only the nodes that are directly linked to the chosen one.

§ 4.5 Summary and Conclusions

The key challenge that is addressed by this Chapter refers to the difficulties in making sense of knowledge models and employing linked urban data in urban analytics, planning, and decision-making. As pointed out in previous Chapters, the increasing necessity to understand different aspects of urban systems in conjunction with one another, encourages the combination of data from heterogeneous sources. This in turn requires the establishment of approaches to data fusion. One such approach can be based on the employment of ontologies, as discusses in the previous Chapter. However, the majority of city stakeholders are not familiar with the formalisms that the data integration methods pertain to. This could have a strong influence on the extent to which these datasets can actually be consumed and exploited. It could also hamper the understanding the relations between the different facets of cities and their dynamics. Providing visual representations of ontologies and linked data can be useful in mitigating this problem. Yet again, as ontologies and integrated datasets grow in size and complexity, the resulting visualizations can easily become muddled.

In addressing these challenges, a set of tools is designed and implemented to support the interactive visual exploration of ontologies and multidimensional linked data for cities. In addition to the toolset, an upper-level ontology of urban networks

is developed, which can be used as a reference framework in domain-specific ontologies that model resources pertinent to a particular facet of urban systems. The contributions are thus threefold:

- A set of web-based tools for the visual representation and exploration of ontologies and multidimensional liked urban data;
- A set of navigation and filtering functions to increase the readability and exploration potential of complex graph visualizations of ontologies and linked data;
- A conceptualization that models the networks in cities, the elements that comprise them, and (an indication of) their relationships, which can be used as a shared vocabulary among city stakeholders.

The tools and the knowledge model use solely open software and standards, are provided under open licenses, and can be accessed through commonly-used web browsers. One of the aims of the aforementioned tools is to assist in bridging, to some extent, the gap between linked data consumers and ontology engineers. Moreover, the tools can be used by domain experts as a basis to evaluate ontologies under development (see also Sect. 3.3.2.3).

A limitation of the developed graph-based visualizations of both ontologies and, especially, linked data is that the illustrated resources are not spatially referenced. Although some of the data elements represented might be, in reality, referenced by a spatial location, their placement on the graph is independent of this location. Hence, the generated graph visualizations are essentially a-spatial. This means that the visual exploration is focused on understanding the topological, rather than the spatial, relations between the elements and, in the case of ontologies, the semantic relationships between the modeled concepts. Moreover, the tools do not provide for any sort of measurement or network analysis statistics (e.g. degree distribution, centrality, betweenness etc.), as this is not in their scope. The tools are targeted to city stakeholders (e.g. urban planners, policy makers etc.), as a means to facilitate their understanding of interrelations between heterogeneous urban data. Also, they can be valuable in the process of ontology development, as a shared platform between domain experts and ontology engineers.

In reference to the developed ontology of urban networks, the implementation presented in this Chapter refers to the first version. Presumably, additional axioms, annotations, properties, and classes will be necessary in future versions, to further enhance the coverage of the domain and knowledge it models. To this end, further contributions and evaluations by relevant stakeholders are deemed necessary in the future. However, given that it is an upper-level ontology, it deliberately contains generic concepts, which in turn can encompass modules that represent more specific elements of urban systems (e.g. to serve as super-classes of concepts included in domain specific ontologies, so that the latter also inherit the properties of the former). Taking into

consideration the lack of analogous ontologies to date, it aims to set the foundation for a sharable and extendable knowledge model that could facilitate the fusion of heterogeneous urban data.

Another relevant issue, which has also been broached in the previous Chapter, is the current lack of comprehensive linked data for cities. This would allow the performance of additional tests, to examine the capacities of the developed tools, in making sense of the relations between resources. With regard to public transport networks, the generated linked dataset, described in Chapter 3, is one such contribution, relative to a specific urban system (i.e. the Athens metropolitan area). If available, the generation of links with integrated data from social media, for instance, would assist in exploring relations between the social activities of individuals and the transportation system of the city. Despite these limitations, the tools presented in this Chapter provide a way to mitigate the problems related to the formalisms of ontologies and linked data, which generally prevent non-expert users from engaging in them. They could potentially facilitate the understanding of ontologies, the consumption and employment of linked urban data in urban analytics, and hopefully instigate the generation of new linked datasets.

5 Deriving Human Activity Attributes from Social Urban Data

§ 5.1 Introduction

The analysis of urban dynamics, especially in regards to human activity and movement, requires that spatial, social, and temporal properties of cities and people are taken into consideration. Urban systems have been generally analyzed, to date, through two distinct approaches. The first one emphasizes the physical structure of cities, in particular the geometrical and morphological attributes of the urban fabric. This approach borrows concepts from location theory and urban morphology. The primary proxies that are used to understand spatial phenomena pertain to area size, population density, physical proximity, employment rates, and land-use locations, to name but a few. In this perspective, social activity is assumed to be influenced directly by the changes occurring in the built environment. The second approach focuses predominantly on the social networks of cities, studying the relationships and interactions between individuals or groups of people. This approach imports best practices and techniques from disciplines such as sociology and urban geography. Unlike the first approach, the proxies used to understand social phenomena primarily pertain to social connectivity, betweenness, network centrality, embeddedness, topological (dis)similarity, and other relational measurements (M. Burger & Meijers, 2011; Sevtsuk & Mekonnen, 2012; Wang et al., 2015). However, the networks of social contacts and interactions are often considered independently of the physical structure of cities.

Although the two aforementioned approaches are mutually related, spatial and social urban phenomena have hitherto been analyzed independently from one another (Andris, 2011). As a consequence, cities have been understood either as clusters of locations or as – largely a-spatial – networks of (social) interactions, yet rarely as combinations of the two (Michael Batty, 2013b; Hristova, Williams, Musolesi, Panzarasa, & Mascolo, 2016). To a great extent, the limitations of traditional urban data (e.g. spatially aggregate attributes, infrequent updates etc.) contributed to this separation.

Nowadays, however, the availability and distinguishing characteristics of social urban data, as defined in Chapter 2, together with the possibilities given by linked data

(described in Chapter 3 and partially in Chapter 4), open new avenues for richer and more detailed descriptions of urban systems than it has been possible hitherto. This chapter presents a set of methods and techniques to derive attributes from social urban data, pertinent to people, places, and the interactions between them, at the disaggregate level (e.g. activity, movement, demographic, and sentiment attributes of individuals; functions of single POIs etc.). The focus is primarily on data generated from geo-enabled social media and LBSNs. The aim is to support the analysis of urban dynamics with detailed descriptions of people and their activity at different places over time. Further, the chapter presents how the extracted attributes can enrich existing metrics of the built environment. The methods and techniques to extract disaggregate attributes from social urban data set the foundation for the design of a system that performs analyses on these attributes and provides insight into the dynamics of human activity in cities (presented in Chapter 6).

The remainder of the chapter is structured as follows. First, in Sect. 5.2, different approaches to measuring, modeling, and characterizing urban space are discussed, by reviewing existing literature. The focus is on the attributes – derived from both traditional and emerging sources of data – that have been used hitherto to measure and model urban systems and their dynamics. Next, Sect. 5.3 describes a set of attributes pertinent to human activity, in terms of both people and places, which can be derived from geo-enabled social media and LBSNs, and presents methods and techniques for extracting these attributes. It also presents how the derived attributes help measure the functional density and diversity of urban areas, as well as the geographical extents of activity spaces over different periods of time. Finally, Sect. 5.4 summarizes the conclusions.

§ 5.2 Approaches to Measuring, Modeling, and Characterizing Urban Space

§ 5.2.1 Measuring the Geometry and Morphology of the Physical Urban Structure

Traditionally, the predominant approach to studying spatial phenomena and the activities of people in cities is pivoted around the geometry and morphological configuration of the physical urban fabric. The main assumption hereof is that the structural characteristics of the built environment have a direct influence on the social and economic aspects of cities and vice versa. A plethora of existing literature adopts methods and tools from the field of urban morphology to analyze urban space, with

a focus on the study of geometrical patterns of the built environment. These patterns refer to arrays of elements such as buildings, streets, open spaces (e.g. lots, public squares, parks etc.), and urban blocks.

The prevailing methods used by the majority of empirical studies to measure the aforementioned patterns are namely: (a) the comparison of different urban systems to evaluate the diversity of distinct physical city characteristics, and (b) the analysis of urban growth and development over different periods of time (Cervero & Kockelman, 1997; Crane, 2000; Forrester, 1969; Song & Knaap, 2004). Both methods use similar proxies, which can be classified into the following four categories: (1) geometrical attributes, such as the area and distance (e.g. Euclidean, cost distance etc.) to measure the density, proximity, and continuity of the urban fabric; (2) land use composition and distribution to measure the concentration, clustering, segregation, and centrality of different regions in a city; (3) topological attributes of spatial networks (e.g. streets) to evaluate accessibility; and (4) socio-economic attributes, such as population size and employment rates, to infer (economic) activities.

The understanding of aspects of urban dynamics in this approach is rather rudimentary. In fact, it is limited to the investigation of urban growth patterns and how these evolve over long periods of time. The proxies mentioned in the previous paragraph are used to inform urban models, which are usually built upon the foundation of classic spatial theories, such as the location theory (Weber, 1909) and the central place theory (Christaller, 1933; Lösch, 1944). The former determines the spatial distribution of land uses and other economic activities, driven by the optimization of costs and benefits. Conversely, the central place theory focuses on the relationship between urban (cities) and rural (towns) settlements, with regard to their mutual trade interactions, but it is mostly concerned with their spatial distribution and size, rather than the actual dynamic flows between them (Berry & Garrison, 1958; Berry & Parr, 1988; Berry & Pred, 1965; M. Burger & Meijers, 2011; Parr, 1987). Both of these theories are based on the assumption that urban space is rather invariable and human behaviors are homogeneous (Fotheringham et al., 2000; Sayer, 1992).

Drawing on the above theories, a wide range of empirical studies has approached properties of the urban environment, such as spatial density, proximity, diversity, segregation, and centrality solely from the standpoint of physical urban structure and its attributes. Jacobs (Jacobs, 1961) and Gehl (Gehl, 1996) have made important qualitative observations about the effects of density and diversity on the social prosperity of cities, yet without any empirical grounding on the basis of quantifiable measurements and observations. In contrast, there exist a number of scholars who employed various land-use data, housing price records, population demographics, and employment rates as proxies to quantify urban density and its connection with the gradation of land values (Alonso, 1964) and the amount of inhabitants (Clark, 1951). These measurements were pivoted around the geometrical (Euclidean or

cost) distance from the central business district (CBD) of a city. As such, they largely associated with von Thünen's model (Thünen, 1966) for the study of optimal agricultural land uses, based on travel costs to the central marketplace (Michael Batty, 2013b). The density of urban space has also been used as a proxy to measure the centrality of a place (inter alia, (Hall & Pain, 2006; Kloosterman & Musterd, 2001; Riguelle, Thomas, & Verhetsel, 2007) or the existence of multiple sub-centers in a city (a spatial phenomenon often referred to as "polycentricity"). These approaches were primarily based on the identification of local employment maximums (McDonald, 1987; McDonald & Prather, 1994) and the clustering of merchandising activity in sub-centers (Thurstain-Goodwin & Unwin, 2000).

Accessibility and connectivity have been traditionally studied through the analysis of spatial networks, and specifically through the space syntax method (Hillier, 1996). Based on the spatial configuration of public spaces, space syntax focuses on the street network of cities, which is approximated by a two-dimensional network of linear elements (axial sightlines). Connectivity and accessibility are respectively measured in relation to the amount of connections of street segments to other adjacent streets, and the number of direction changes from a particular street. Thereby, the space syntax method does not use points as proxies to represent specific locations, but it is rather based on a linear representation of the urban spatial structure. Socioeconomic activities are therefore considered simply as the aftereffect of the way these elements interconnect with one another.

Although space syntax can be useful in analyzing the effects of future urban interventions on the wider city fabric, it is also characterized by several limitations that may lead to significant misinterpretations, especially with regard to the social activities in cities. Ratti (Ratti, 2004) has specifically focused his criticism on the lack of metrical properties and building height information in space syntax, while Batty et al (Michael Batty, Jiang, & Thurstain-Goodwin, 1998) have pointed out the absence of elements pertinent to land uses, and the corresponding limitations in representing human mobility and interactions between locations. In addition, space syntax is entirely independent of any cultural or socio-economic factor that may influence significantly the way people interact with the urban environment.

§ 5.2.2 Modeling Spatial Flows and Interactions

The approaches described in the previous section, tend to portray the structure of cities in a more or less static fashion. Dynamic relations among different locations are hardly ever taken into consideration. Essential features, such as urban density and diversity, are approached solely from the perspective of population size, physical proximity and

the array of different land use types, but scarcely from the viewpoint of the volume and type of activities. In measuring, simulating, and predicting the movement of people and goods between different location, various spatial interaction theories have been developed over the years (Michael Batty, 2009).

The early approaches to spatial flows adopted theories primarily from the domain of physics. As a matter of fact, the first spatial interaction models for the measurement and prediction of migration flows were created as analogies to gravity models (Michael Batty, 2013b; Fotheringham et al., 2000). In these models, the movement of people from an origin *i* to a destination *j* is described as a function of their respective population size, divided by the distance d_{ij} between the two cities and defined as:

$$M_{ij} = K\frac{P_i P_j}{d_{ij}} \tag{5.1}$$

Where represents the amount of movements between the two locations *i* and *j*, P_i and P_j represent their respective population size (used as a proxy for measuring their size), and *K* is a scaling constant, which could be measured, for instance, in kilometers/person, provided that the distance is measured in kilometers and the movement by the amount of people traveling from the origin to the destination. In analogy to the Newtonian gravitation, the movement model of the equation (5.1) implies that the flow of movement between two locations increases when their respective populations increase in size. Conversely, their mutual attraction will be weaker as the distance between them becomes larger.

However, it was later acknowledged (Fotheringham et al., 2000; Fotheringham & O'Kelly, 1989) that the assumptions of the above-described model were quite abstract and generic and, therefore, were not appropriate for representing the complexities of actual movement. For instance, human mobility, traffic flow, or the flow of goods usually differ from one another, and this diversity is not addressed by the model in (5.1). Therefore, in order to allow for these differences to be represented, the model was extended to accommodate additional parameters for calibration:

$$M_{ij} = K\frac{P_i^{\alpha} P_j^{\beta}}{d_{ij}^{\gamma}} \tag{5.2}$$

TOC

where α is an exponent related to the ability of an origin location to generate movement, β is an attractiveness indicator determined by the types of activity at the origin, and γ is a factor for describing the friction of distance, usually determined by the type of transport system connecting the two cities. The value of these indicators is not fixed. Changes in the type of the activities in question or the way in which they are performed, in addition to advancements in the transportation systems connecting the two locations can essentially affect the value of the exponents and, subsequently, the value of the overall flow. In general, historical data on flows (e.g. migration) are used as proxies for estimating the aforementioned indicators. These types of data – sourced primarily from travel surveys – are usually not up-to-date (travel surveys are updated approximately every 5 – 10 years). Emerging social urban data could be useful, in this regard.

A later improvement to these models, resulted in the development of a series of spatial interaction models that were built as analogies to the principle of maximum entropy. These particular models are also known as the "family of spatial interaction models", and were developed by Wilson (Wilson, 1967, 1975). The "family" comprises four types of models, namely (a) the unconstrained, (b) the production-constrained, (c) the attraction-constrained, and (d) the production-attraction-constrained models (Fotheringham et al., 2000; Wilson, 1970, 1975).

A recent alternative to both the gravity model and Wilson's spatial interaction models is the Radiation Model for human flows and interactions, introduced by Simini et al (Simini, Gonzalez, Maritan, & Barabasi, 2012) and described as follows:

$$M_{ij} = M_i \frac{P_i P_j}{(P_i + S_{ij})(P_i + P_j + S_{ij})} \tag{5.3}$$

where P_i and P_j represent again the populations at origin i and destination j respectively, while S_{ij} signifies the total population in the circle with radius d_{ij} centered at i, excluding the populations at both origin and destination. M_i denotes the total outgoing flux, originating from i. Unlike the models that have been described thus far, which take into account only the population size of the origin and destination locations, the radiation model also accommodates the population density of regions surrounding the origin. In comparison with the gravity and entropy-based models, the main advantage of the radiation model lies in its parameter-free nature, in the sense that it excludes any exponents/indicators that are characteristic of the previous models.

Spatial interaction models have been widely applied to the analysis of the movement dynamics between locations. In recent literature, Thiemann et al (Thiemann, Theis, Grady, Brune, & Brockmann, 2010) explore a network of interregional human flows to operationalize and assess whether existing political and administrative borders reflect the contemporary mobility and connectivity patterns of people. Banknote transaction data are used as proxies for the measurement of human mobility. Similarly, van Oort et al (van Oort, Burger, & Raspe, 2010) employ a dataset about the interrelations of firms within the most important conurbation in the Netherlands – known as the "Randstad" – and use it as a proxy for evaluating whether this urban system functions as an integral (economic) entity. (De Goei, Burger, Van Oort, & Kitson, 2010) and (M. Burger & Meijers, 2011), employ commuting data and incorporate them into gravity models to study the spatial interaction patterns of people across cities in the UK and the Netherlands respectively.

§ 5.2.3 Integrating Social Networks into the Physical Structure of Cities

The approaches presented thus far focus on the interactions between physical locations (e.g. between two cities or two places in a city). In contrast, research on the interactions between individuals has followed a different path, usually ignoring the spatial parameters of the built environment and focusing primarily on topological attributes of the social networks (Freeman, 1979; Hanneman & Riddle, 2005; Jackson, 2010; Prell, 2012). As a consequence, social connectivity is usually studied outside of the physical space. Research on multilayered networks (Boccaletti et al., 2014) is only recently gaining in popularity and, drawing on the emerging possibilities of social urban data, attempts are made to interconnect social relationships with geographic space (Andris, 2016; Boccaletti et al., 2014; Hristova et al., 2016; Wang et al., 2015). The integration of social connections into the physical urban space has potential to provide richer descriptions of urban areas and their dynamics.

The mutual exploration of both social and spatial aspects of urban dynamics has hitherto been hampered by the limited availability of multidimensional urban data at finer spatial and temporal resolution. Nowadays, however, the increasing availability of social urban data, has instigated a wealth of research that touches upon social and spatial aspects of urban systems. Though, it should be noted that the "social" aspects in related literature do not necessarily refer to social connections, but may instead pertain to social activity (e.g. mobility patterns, check-ins to specific venues) that derives from a – usually online – social network (e.g. social media platforms). In this context, the following paragraphs classify related literature that uses social urban data as proxies for understanding urban dynamics, according to the source they employ, in particular: (1) mobile phone records (CDRs); (2) GPS traces; (3) human-generated content from social media; and (4) combinations of the previous sources.

Call detail records from mobile phones are increasingly gaining in significance, as regards the study of human activity patterns in cities. Departing from the more common study of aggregate flows of people, Calabrese et al (Calabrese, Smoreda, et al., 2011) analyze a large-scale dataset of CDRs, derived from a single mobile phone operator in Portugal, in order to explore the person-to-person activity over space and time. The analysis showed that about 93% of users calling each other, irrespective of the distance between their homes, have also been physically co-located in the same place at the same time, as inferred from the cell tower area within which the calls have been made. Wang et al (Wang et al., 2015) enhance the approach of the previous work by incorporating geometrical data about the physical urban fabric to create links between the social networks of users and the spatial locations where their activity occurs. In an attempt to alleviate the impact of the several limitations that characterize CDRs, (Diao et al., 2015) combine mobile phone records with data from traditional travel surveys, to infer the spatial and temporal distribution of everyday activities of individuals. In contrast, (Grauwin et al., 2015) employ CDRs at an aggregate level, in order to perform a comparative analysis on the dynamics of human activities across three major cities – namely, New York, London, and Hong Kong – and juxtapose them with land use data to explore potential correlations and influences. By superimposing the aggregate activity dynamics in all three cities, the authors discover a large degree of pattern similarity, which could be explained as the repercussion of the globalized economy on the shaping of contemporary cities, while several diversities exist at the individual level. Correspondingly, (Amini et al., 2014) conduct a comparative analysis based on large-scale mobile phone records, yet at the national level and between a developing and an industrialized country. A particularly interesting feature of this work lies in testing the performance and suitability of several spatial interaction models, described in Sect. 5.2.2, in the challenging context of developing nations.

Several studies on human mobility exploit the potential of GPS traces, as these become increasingly available through sensing devices and sensor networks embedded in urban infrastructure. Yuan et al (Yuan, Zheng, & Xie, 2012) use trajectory data from GPS-enabled taxicabs, in combination with sets of POIs, to infer the real usage and function of different areas within Beijing and their evolution over time. Gao et al (S. Gao, Wang, Gao, & Liu, 2013) also employ taxi trajectory data to approximate the distribution of traffic flows. In addition, these data are integrated into spatial network models, focusing specifically on the betweenness centrality, in order to investigate its capacity as an indicator for simulating and predicting traffic flow patterns. However, due to the low degree of semantic expressiveness characterizing GPS and other sensor-generated data (see Chapter 2), it is generally difficult to infer individual behaviors of people, which have a strong influence on the configuration of mobility patterns.

Contrariwise, data generated by people through various geo-enabled social media have a great capacity to create connections between the social ties among users and the spatial location of their activities over time. Scellato et al (Scellato, Lambiotte, Noulas,

& Mascolo, 2011) use datasets from three different social media platforms, namely Brightkite, Foursquare, and Gowalla, as proxies for approximating several socio-spatial attributes of (groups of) people and are, then, incorporated into urban gravity models. This work is further extended in (Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012), which focuses on a comparative analysis of human mobility patterns among 34 cities across four continents, by exploiting large-scale check-in data from Foursquare. Unlike the several limitations of traditional urban data sources (see also Chapter 2), the global scale of LBSNs allowed the authors to simultaneously investigate the diversities and similarities of human flows and interactions across disparate urban environments. (Cranshaw et al., 2012) also utilize Foursquare data to study the dynamics of activity patterns in Pittsburg, Pennsylvania. These are subsequently incorporated into models for clustering areas in the city, based on the activity types, taking into account both spatial and social proximity between venues and users respectively. What is particularly interesting in this work is the validation of the social media analysis results, by comparing them with data from several interviews with residents. In a similar way, Silva et al (Silva, Melo, Almeida, Salles, & Loureiro, 2013) explore urban dynamics through a mutual comparison of large-scale datasets from two social media platforms, namely Foursquare and Instagram, yet without performing a cross-validation against survey data. Likewise, (Del Bimbo, Ferracani, Pezzatini, D'Amato, & Sereni, 2014) employ geo-referenced data from Facebook and Foursquare to classify different venues in the city, based on users' interest profiling. (Shelton et al., 2015) use Twitter data from Louisville, Kentucky to investigate the intensity of segregation between two neighborhoods in the city. A significant characteristic of this study is the coupling of the aforementioned web data with local knowledge about cultural, historical, and political factors that largely supplement the interpretation of the online datasets.

The studies described in the previous paragraph employ a single source of data. There exist a few examples, in which multiple sources of social urban data are used. (Sagl, Resch, Hawelka, & Beinat, 2012) combine CDRs with data from Flickr to analyze aggregate spatiotemporal human mobility patterns. Earlier to this study, (Vaccari, Calabrese, Liu, & Ratti, 2009) introduced an urban information system that allows datasets from diverse sources to be collected, stored, and integrated. More recently, (Stanislav Sobolevsky et al., 2015) investigated the operationalization of city attractiveness factors for foreign visitors, by using heterogeneous data from bank card transactions and social media platforms (Twitter and Flickr), to approximate economical, social, and behavioral aspects of human activities.

§ 5.3 Deriving Disaggregate Attributes of Human Activity

The studies discussed thus far provide an indication of the new possibilities given by social urban data. Besides having an understanding of the distinguishing characteristics of social urban data (see Chapter 2), it is also important to investigate what attributes can be derived from these emerging sources, which have potential to provide detailed descriptions of the spatiotemporal dynamics of human activity.

This section describes a set of attributes pertinent to human activity, in terms of both people and places, which can be derived from social urban data, and presents methods and techniques for extracting these attributes. The focus here is primarily on data from geo-enabled social media and LBSNs. The derived attributes refer to characteristics of both the people who perform a certain (social) activity (e.g. socio-demographic characteristics, home location, individual trajectory, activity space, sentiments etc.) and the places where activities occur (e.g. land use, type of activity). The attributes are classified into four categories according to the nature of the feature they describe, namely: (1) *socio-demographic attributes*, (2) *functional attributes of places*, (3) *individual spatial movement patterns*, and (4) *topical attributes* (Table 13).

Further, the section presents how the derived attributes can be used to enrich existing metrics of the urban built environment (e.g. functional density and diversity), which could support the characterization of urban areas according to the activities performed over time.

TABLE 13 Attributes of human activity and methods for deriving them from geo-enabled social media and LBSN data.

Category	Attribute	Method / Technique
Socio-demographic attributes	– Home location – Age range, gender, ethnicity	– Recursive grid search / Geohashes – Directly from profile information / Using specialized software
Functional attributes of places	– Land use	– POI categories
Individual spatial movement patterns	– Individual trajectory – Activity space	– Sequence of posts – Online visits (i.e. check-ins)
Topical attributes	– Semantics – Sentiments	– Natural language processing / Keyword filtering – Sentiment analysis

§ 5.3.1 Estimating Socio-demographic Attributes

Social urban data and, in particular, geo-enabled social media and LBSNs are tagged to space and time and can provide estimates of a person's socio-demographic attributes (e.g. home location, age range, gender etc.).

The following paragraphs present methods and techniques to address the demographic diversity of the individuals contributing content to geo-enabled social media and LBSNs.

§ 5.3.1.1 Home location approximation

Residential areas play a significant role in the analysis of human mobility in cities, as the majority of movements are generated from each person's home location and it is this same place where most human movements end up. Therefore, having knowledge of home locations at the disaggregate level is particularly important when it comes to studying human dynamics and to understanding the mobility choices of people with regard to commuting or leisure.

In approximating the home location of an individual from social media data, recursive grid search could be used (Cheng, Caverlee, Lee, & Sui, 2011). By taking into account the complete history of geo-referenced posts of an individual (i.e. social media user), the home location can be approximated as the one from which this person appears to post most frequently. Instead of considering the average location of the entire set of geo-referenced posts as evidence of a person's home location, the recursive grid search method could reach more accurate approximations. To achieve this, a set of consecutive steps need to be followed. First, the geo-referenced posts are clustered into a grid, which comprises square cells of a certain (larger) size that can be freely defined. Second, the cell that contains the largest amount of posts, along with the eight cells that are adjacent to it, are divided into a smaller grid that comprises cells of a much smaller size than the initial ones (about one tenth of the size of the original squares). Next, the aforementioned procedure is repeated as many times as needed, until the grid cells are about a thousand times smaller than the original ones. Eventually, the centroid of the cell that contains the largest amount of posts can be used as a proxy for the home location of an individual.

In enriching this method, geohashes could be used in place of custom grid cells (Bolivar, 2014; Fox, Eichelberger, Hughes, & Lyon, 2013). Geohashes consist of a latitude/longitude geocoding system and spatial data structure, which represents coordinates in a grid-like fashion. Posts are clustered into geohashes of increasing

length – the larger the length of a geohash, the larger the level of detail and the lesser the amount of error – following the exact same procedure as described in the previous paragraph (Figures 20 – 21).

An additional improvement with regard to accuracy, involves the consideration of the posts that are generated between 6pm and 8am, assuming that these would originate from the actual location of a person's home. This approach has been followed by (Calabrese et al., 2013) in order to approximate the home location of a mobile phone user, derived from CDRs. By adapting it to the recursive grid method, the centroid of the grid cell or geohash containing the largest amount of posts that were generated in the aforementioned time interval can be used as a proxy for the home location of an individual. Validation of the extracted results can be done by cross-checking the collected geo-referenced social media data with disaggregate census data, where available.

Drawing on the approximation of the home location, it is possible to classify users of social media into social categories and, in particular, residents, non-residents, and foreign tourists. In particular, if the estimated place of residence of an individual is located in the same city as a city in question, then it could be assumed that this person represents a resident. Conversely, in the case where a user's approximated home location is placed outside of a city under consideration, but both are located in the same country, it could be assumed that this person represents a non-resident or a commuter. Finally, if a user's approximated home location is placed outside of a city in focus, and is also located in a different country, then it could be assumed that this person represents a foreign tourist (Psyllidis et al., 2015a). The identification of different groups of people, instead of treating all social media users as uniform, could provide new insights into the spatiotemporal dynamics of human activity in cities.

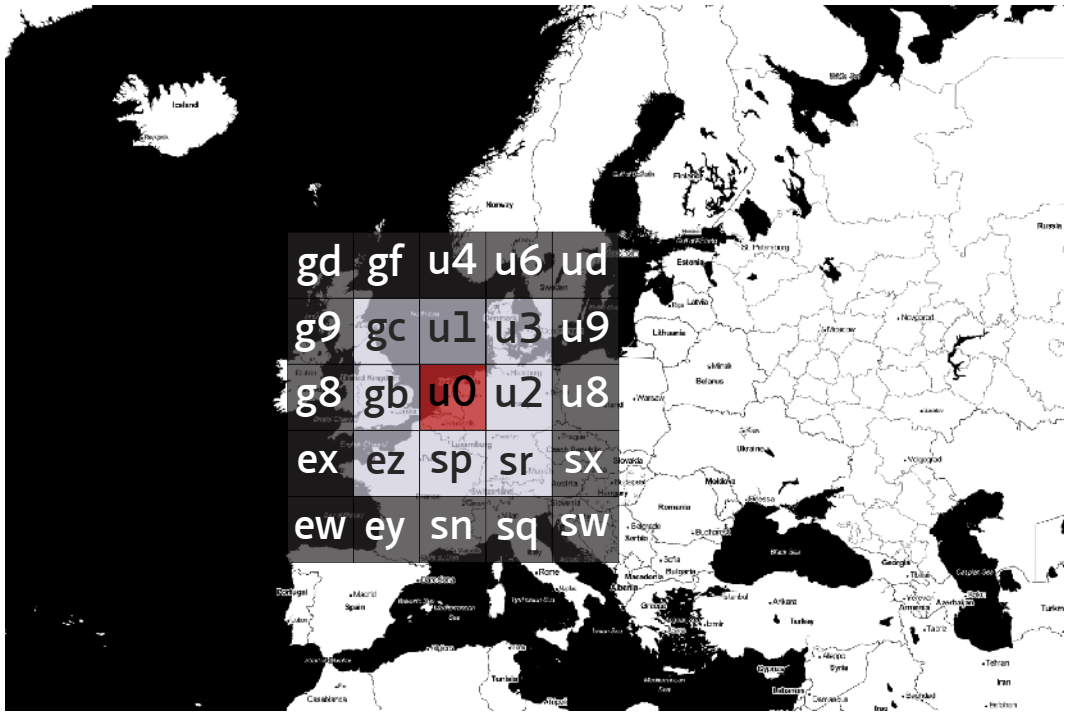


FIGURE 20 Recursive grid search with geohashes. The geohash containing the largest amount of posts (here *u0*) and the eight cells adjacent to it are further divided into smaller geohashes.

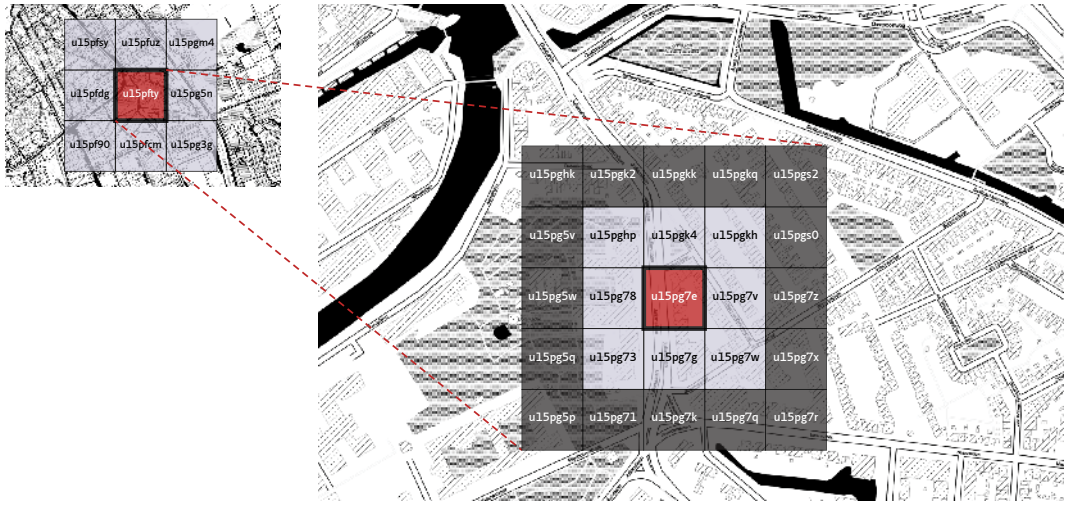


FIGURE 21 Iterative division of geohashes. The centroid of the cell that contains the largest amount of posts is used as proxy for the home location.

TOC

§ 5.3.1.2 Socio-demographic attributes of individuals

Among the most common socio-demographic attributes that can be derived from social media are the *age*, the *gender*, and the *ethnicity* of an individual. The ways to extract these attributes vary significantly from platform to platform, as each of them uses a different model to describe the profile of a user. For instance, Sina Weibo asks for an explicit indication of gender and age, whereas platforms such as Twitter and Instagram allow users to optionally specify this type of information. Therefore, it is difficult to derive demographic attributes by using a single technique that is applicable to all types of social media.

With regard to the *age* attribute, in cases where it is not explicitly provided by the users, the content of posts can be used as a proxy to derive an estimate of a person's age range, using natural language processing techniques (Rao, Yarowsky, Shreevats, & Gupta, 2010). Yet, this method allows only for a very broad classification of age (e.g. below or above the age of 30). In achieving more accurate estimations, the analysis of profile pictures using facial recognition algorithms is increasingly gaining in popularity (Bolivar, 2014). However, not all users provide a real or recent picture of themselves, which subsequently decreases the estimation accuracy. Even in the cases where recent and real profile pictures are provided, the facial recognition algorithms will return a certain age range, rather than a precise age value.

Similarly, the *gender* of social media users can be inferred from the characteristics of the profile picture. Burger et al (J. D. Burger, Henderson, Kim, & Zarrella, 2011) tested a variety of attributes that are frequently included in a social media profile model (e.g. full name, screen name, profile description etc.), and by additionally performing textual analysis on sample posts, it is concluded that the most informative gender estimation variable is the full name of a person. However, a significant limitation of this variable is that only a limited number of social media platforms require the real full name of a user.

An additional demographic characteristic that can be inferred from social media is the *ethnicity* of an individual. According to (Pennacchiotti & Popescu, 2011) the most reliable proxy for inferring a person's ethnicity is the linguistic content of her/his posts, whereas profile-based features (e.g. profile pictures) have limited potential in this regard. Therefore, a combination of linguistic content analysis along with analysis on the full name of a user can be employed, in order to potentially increase the accuracy of the obtained results. This combination of techniques could also be useful for the extraction of age and gender-related attributes.

§ 5.3.2 Inferring Functional Attributes of Places

An advantage of geo-enabled social media and LBSN data in comparison with other sources of social urban data, such as mobile phones and sensors, is that geo-referenced posts can be mapped to specific places (i.e. POIs), which are in turn connection with a given function (e.g. museum, restaurant, theater etc.). Moreover, they are usually enriched with semantics about the type of activity carried out in a given place, which may deviate from its function. Topical attributes will be touched upon in Sect. 5.3.4.

The following paragraphs focus on the extraction of land use attributes from social media data. These attributes are then used to enrich metrics of the functional density and diversity of the built environment.

§ 5.3.2.1 Land use approximation

Each spatially and temporally tagged post of an individual on a social media platform reflects an instance of a certain activity, linked with a physical location. Some of these locations appear to attract further attention from people, either because they are interesting or because they are characterized by a useful function that allows several types of social interactions to occur. These particular places are referred to as *points of interest* (POIs) and constitute important attractors or generators of human flows. The majority of the functions characterizing online POIs correspond to the actual land uses of buildings or public spaces. Therefore, it is assumed that they can be used as reliable proxies for estimating real-world land uses of the built environment at the disaggregate level. The extent to which these POIs reflect the actual underlying physical locations and functions is subject to several biases of social media that need to be considered in the extraction and analysis process.

Each social media platform follows a different approach to POI description and assignment to a post. Geo-enabled social media platforms such as Instagram, Twitter, and Sina Weibo, provide a set of geo-coordinates (latitude and longitude) from which POIs in the vicinity can be inferred. Conversely, LBSNs such as Foursquare, enable the extraction of specific types of function, venue categories, exact place names, links to websites, and the level of popularity. In addition to this, there exist a variety of taxonomies, which are used to classify POIs according to their category. For instance, the entire Foursquare taxonomy consists of more than 400 POI types, which are classified into 10 general categories (e.g. arts and entertainment, college and university, event, residence, professional places etc.). In contrast, Sina Weibo, a popular Chinese social media platform, uses a hierarchy of more than 300 POI types, which are grouped into 9 categories (e.g. travel and accommodation, office and

organization, education, shopping etc.). Table 14 presents an alignment of the general POI categories between Foursquare and Sina Weibo, as listed in their respective APIs.

TABLE 14 Alignment of POI categories between Foursquare and Sina Weibo (based on API documentation).

POI category (Foursquare)	POI category (Sina Weibo)	Indicative function
Arts & Entertainment	Life & Entertainment	Movie theater, Museum, Theater etc.
College & University	Education	College academic building, University faculty etc.
Event	Other places	Conference, Convention, Festival etc.
Food	Food & Beverage	Restaurant, Coffee shop, Diner etc.
Nightlife spot	Life & Entertainment	Bar, Pub, Nightclub etc.
Outdoors & Recreation	Park & Plaza	Sports field, Fitness center, Sports club etc.
Professional & Other places	Enterprise Office building & Organization	Office building, Government building, Non-profit organization etc.
Residence	—	Private home, Residential building
Shop & Service	Shopping	Various types of shops, Bank, Super Market etc.
Travel & Transport	Travel & Accommodation	Airport, Bus station, Hotel etc.

The mapping process of a certain post to a POI is fully dependent on the social media platform in question and the possibilities provided by its respective API. The completeness and coverage of available POIs that are included in the corresponding taxonomies vary according to the technology penetration levels in different regions. In general, affluent regions where the use of social media is popular are characterized by ample coverage of online POI venues. The scarcity of place-related information strongly affects the characterization of urban areas (Sengstock & Gertz, 2012). Moreover, there exist several geo-tagged posts that are not linked with a specific POI, but rather with a larger spatial unit (e.g. a neighborhood, a city, a county etc.), which further hampers the approximation of land uses at the disaggregate level.

§ 5.3.2.2 Measuring density and diversity

The approximation of land use types from social media data enables the reformulation of measurements pertinent to the density of activity patterns and the diversity of functions. As it has been described earlier in this Chapter (see Sect. 5.2), the measurement of these features of the built environment has been a long-standing issue of urban analysis and planning. However, to date, they have been approached

more qualitatively than quantitatively, due to lack of sufficient data. To be able to measure the density and diversity of an urban area, the key variables to be considered are the array of different functions, the number of people performing activities, the amount of visits, and the temporal variation of activity patterns. In the previous sections, it has been explained how each of these parameters can be derived from geo-enabled social media and LBSNs. This section describes how these attributes can enrich measurement of urban density and diversity.

According to (Cervero & Kockelman, 1997), the *density* of a place refers to the intensity and relative compactness of activities in a given area, taking into account the population and employment rates in this area, as well as the proximity of residential units to other land uses. As an alternative to this, the number of individuals visiting an array of POIs in a certain area, the corresponding amount of visits, and the temporal distribution of activity patterns can be used as proxies to measure the density of an urban area. Therefore, the density of a spatial unit (*u, v*) (e.g. a neighborhood), within a wider area *m* x *n*, in a given time period *t*, can be formally expressed as:

$$D(u,v,t) = \frac{[N(u,v)]t}{\sum_{i=1}^{m}\sum_{j=1}^{n}[N(i,j)]t} \tag{5.4}$$

Where *N(u, v)* resembles the number of people attracted to the spatial unit *(u, v)* in a given time period *t*, and *N(i, j)* is the total amount of visits in the wider area in question (e.g. an entire city). In defining the variables, the number of individual social media users visiting a predefined area *(u, v)* is used as proxy to measure the *N(u, v)* parameter, whereas the number of individual posts (i.e. excluding multiple posts from a single user in a short time interval) that are mapped to a specific POI within the wider region *m* x *n* serves as a proxy for calculating the *N(i, j)*.

On the other hand, the functional *diversity* describes the degree of land-use heterogeneity in a given area, denoting the array of activities performed by people in this district (Cervero & Kockelman, 1997; Hess, Moudon, & Logsdon, 2001). Thereby, it can be expressed by an entropy index, as follows:

$$H(u,v) = -K\sum_{j=1}^{J} P_j(u,v)\ln\left[P_j(u,v)\right] \tag{5.5}$$

TOC

Where K is a positive constant (equivalent to Boltzmann's constant in thermodynamic entropy) defined as $K = 1/ln(J)$, in which J resembles here the total amount of activity types in the area (u, v), and $P_j(u, v)$ indicates the proportion of people who perform an activity type j within the area (u, v) and for a time period t (Zhong et al., 2015). Respectively, the total number of activity types J can be approximated by the array of different POI categories within a predefined area (u, v), and the amount of people carrying out a certain activity type j in this district can be measured by the number of individual users visiting a specific POI category.

§ 5.3.3 Deriving Individual Spatial Movement Patterns

Traditionally, information on the spatial movement of people is extracted from travel surveys, at the individual or household level. Although these sources are more trustworthy and accurate than social media data, they take several years to be updated and to subsequently become available for use.

The following paragraphs describe attributes of spatial movement that can be derived from geo-enabled social media and LBSNs and can also be used to measure the spatial distribution of human mobility over time.

§ 5.3.3.1 Individual trajectory

A significant part of the dynamics of human activity in cities refers to the movement patterns of individuals. Origin and destination locations are pivotal to measuring and modeling flows (see Sect. 5.2.2). However, knowledge of the trajectories followed by individuals that are mapped to the physical street network can provide better insights into the laws governing human movement than origin-destination matrices. However, accurate trajectory data are scarcely available. Typically, the main source used to derive mobility patterns are travel surveys, at household or individual level, but information is limited to a set of origin and destination locations at the time the survey is conducted.

Recently, mobile phone data and GPS records have been used as proxies for inferring the actual trajectories followed by individuals (inter alia (Alhasoun et al., 2014; Bayir, Demirbas, & Eagle, 2009; Calabrese, Colonna, Lovisolo, Parata, & Ratti, 2011; Diao et al., 2015). Unlike travel surveys, these data sources are updated frequently and individual trajectories can be extracted in the form of a spatiotemporal sequence of activities (i.e. a set of geo-referenced phone calls or a GPS track).

In a similar way, data from geo-enabled social media can also be used as proxies for extracting human trajectories at the disaggregate level. However, compared to CDRs or GPS records, the volume of social media posts generated by an individual is usually much smaller. While a person can transmit several call signals throughout the day, the posting activity of the same person on online social networks may be much more sporadic. Therefore, the most comprehensive sources of social urban data, when it comes to human movement patterns, are mobile phones and GPS systems.

Nevertheless, individual trajectories from geo-enabled social media can also be inferred by means of extracting a spatiotemporal sequence of posts. In particular, the geo-coordinates of a post – where available – are collected, in combination with the time tag, together forming a triple $\{x_i, y_i, t_i\}$, where x_i and y_i denote the coordinates of a post in a location i at a given point in time t_i. Although this sequence of posts indicates a set of activities in chronological order, it is not by itself sufficient to approximate the actual trajectory followed by an individual. In fact, it only indicates how a person moves from one place to another, representing a set of origin and destination locations, without evidence of the path that this person chose to follow. In order to approximate the trajectory followed by an individual between two consecutive posts, and to align it with the actual street network of a city, intermediate waypoints between each origin and destination have to be determined. One way to determine the intermediate waypoints is to collect the total daily social activity (i.e. the total amount of geo-referenced posts generated in a single day) of a person in question and, subsequently, calculate the distance between each two consecutive posts, in addition to the time interval between these two posts. These data can then be fed into a route calculation algorithm (such as the one provided by the Google Directions API[47]) to roughly approximate the intermediate waypoints of the trajectory (Figure 22). An implementation of this trajectory extraction method is further explained in Chapter 6.

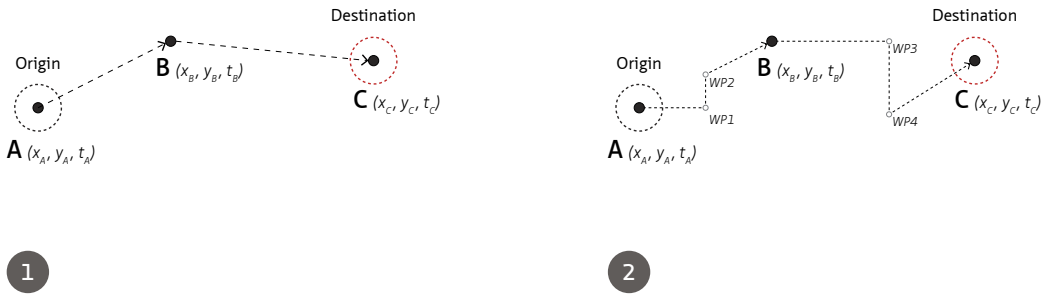


FIGURE 22 Individual trajectory inferred from social media posts (1) as a simple spatiotemporal sequence, and (2) as a sequence with intermediate waypoints.

47 https://developers.google.com/maps/documentation/directions/ (Accessed on December 30, 2015).

TOC

§ 5.3.3.2 Activity spaces

The set of extracted POIs (see Sect. 5.3.2.1) can be used as proxies to infer the activity territory of an individual over different periods of time. In general, the set of locations frequently visited by a person on a typical day or a longer period of time comprises her/his *activity space* (Axhausen, 2007). Based on this definition, the activity space of an individual may include her/his place of residence, workplace(s), and a group of other locations pertinent to leisure and recreational activities (Figure 23). Places of residence can be inferred by means of the methods described in Sect. 5.3.1.1. Conversely, workplaces and locations of leisure activities can be inferred from the extracted POIs, as described Sect. 5.3.2.1.

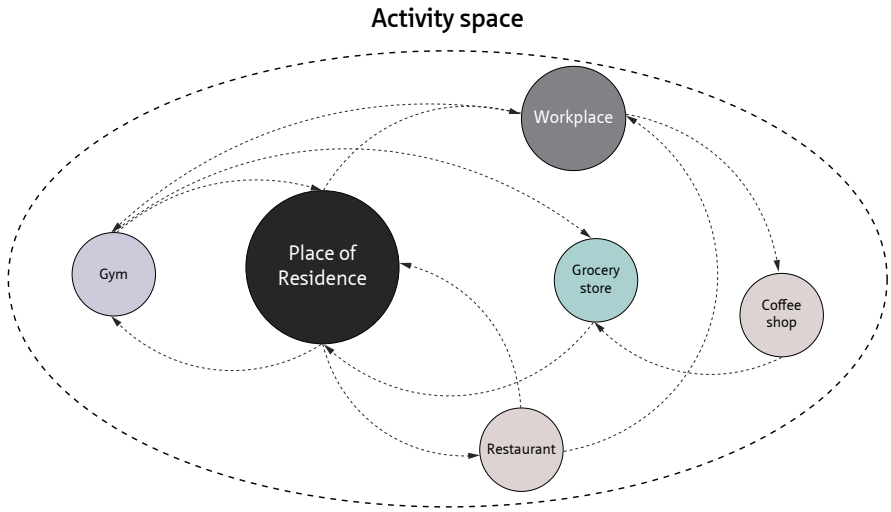


FIGURE 23 Activity space of an individual consisting of the place of residence (1st place), workplace (2nd place), and a set of locations pertinent to other activities (3rd places).

The sum of the POIs visited by a person in a given period of time, in addition to her/his estimated home location and her/his approximated workplace, indicate her/his activity space, derived from social media data. Formally, this could be described as follows:

$$(AS)_u = \sum_{i=1}^{n_{POI}^u} POI_i(t_i) + (HL)_u + (WP)_u \tag{5.6}$$

TOC

Where $(AS)_u$, is the activity space of a user u, n^u_{POI} resembles the total number of extracted POIs visited by u, $POI_i(t_i)$ indicates the location of a POI at a specific point in time, $(HL)_u$ is the estimated home location of the user, $(WP)_u$ and is the approximated workplace (if applicable).

Unlike traditional urban data which usually indicate where people live – that is, in most cases, where they spend the night – the extraction of activity spaces from social urban data enables the characterization of urban areas according to the places where they spend time during the day. Moreover, the extraction of activity spaces of two or more social contacts (e.g. online "friends" or "followers"), derived from geo-enabled social media and LBSNs, could give an indication of geographical places that are shared by individuals belonging to a common network of social contacts (Wang et al., 2015).

However, one major limitation of geo-enabled social media and LBSNs, is that they do not give an indication of the amount of time a person has spent at a specific location. Also, the time tag accompanying social media posts does not necessarily resemble the actual time a person visits a certain place. Further, the majority of social media platforms do not give the possibility to users to explicitly indicate their departure time from a location. This can only be estimated, to a certain extent, through textual analysis of the post content (Cramer, Rost, & Holmquist, 2011; Frith, 2014; McKenzie et al., 2015). Moreover, a certain amount of posts is not generated or tagged to the actual moment an activity took place. In some cases, the online social activity (i.e. creating a post) may refer to an actual activity (e.g. a visit to a specific place) that was carried out in the past or is about to happen in the future. These temporal biases need to be considered when using social media data as proxies for the analysis of human activity over space and time.

§ 5.3.3.3 Radius of gyration

The extraction of individual trajectories from social media data provides further insight into the geographical extent of human activity (i.e. how far a person travels) and the intensity of the movement flow (i.e. how often a person travels between places). This information can be used to enrich metrics of human mobility and, more specifically, the *radius of gyration*. The radius of gyration determines the extent of the activity space of an individual (see also Sect. 5.3.3.2), in addition to the intensity of her/his flow patterns (Figure 24). The time tags accompanying social urban data enable the calculation of the radius of gyration for a given time interval (e.g. hourly, daily etc.).

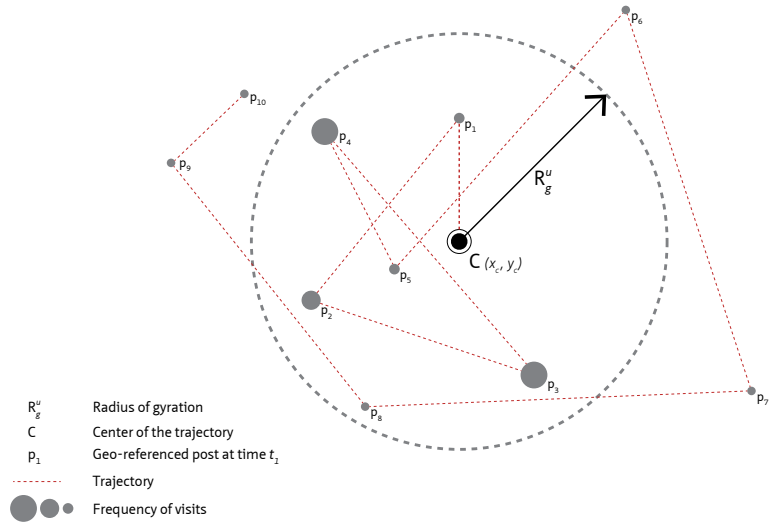


FIGURE 24 Radius of gyration, based on a person's trajectory as inferred from the sequence of social media posts. Rarely visited places have low impact on the radius of gyration.

The radius of gyration is computed as the root mean square distance of an individual's location from the average location of all her/his transmissions within a given time interval. (Gonzalez et al., 2008) used mobile phone data as proxies to calculate the radius of gyration in the analysis of individual human mobility patterns.

Accordingly, in the case of geo-enabled social media data, the radius of gyration can be described as the distance covered by an individual between locations (based on the post history), as well as the frequency of this trip. By adding the temporal dimension, the radius of gyration can be calculated as follows:

$$R_g^u(t) = \sqrt{\frac{1}{n_p^u(t)} \sum_{i=1}^{n_p^u(t)} (x_i - x_c)^2 + (y_i - y_c)^2} \tag{5.7}$$

Where $n_p^u(t)$, represents the total number of posts generated by an individual within a period of time t, x_i and y_i are the coordinates of a post in location i, x_c and y_c and represent the center of the trajectory that is created by joining the single posts together, and are defined as follows:

TOC

$$x_c = \frac{1}{n_p^u(t)} \sum_{i=1}^{n_p^u(t)} x_i \tag{5.8}$$

$$y_c = \frac{1}{n_p^u(t)} \sum_{i=1}^{n_p^u(t)} y_i \tag{5.9}$$

Therefore, the variance between (x_i, y_i) and (x_C, y_C) represents the distance between a specific post of the user u and the average location of the total number of posts over a specific period of time. In other words, it indicates how far a user travels to perform certain activities.

From the equation (5.4) it can be inferred that a person who only carries out activities within a short distance from the average location of posts will result in a low radius of gyration. Conversely, a high radius of gyration characterizes a person who travels long distances (in relation to the average location) to carry out certain activities. Also, a person who only performs a few long-distance activities will produce a higher radius of gyration, compared to a person who performs a large number of activities, yet only within a short distance. As a consequence, the radius of gyration can be an important metric in the analysis of the spatiotemporal dynamics of human mobility.

§ 5.3.4 Extracting Topical Attributes

Unlike other contemporary sources of social urban data, such as sensors or mobile phones, a distinguishing characteristic of geo-enabled social media LBSNs pertains to their semantic richness. The majority of social media data consist of human-generated content represented in a textual or other (e.g. image, video etc.) format, from which topical attributes of human activity can be derived.

This section outlines methods to extract topical attributes of human activity by analyzing the semantics and sentiments of social media content.

§ 5.3.4.1 Semantics and sentiments

The analysis of the content of social media data has potential to derive topical information on the type of activity one performs in a given place, which could deviate from the place's original function (e.g. studying in a coffee shop). *Semantic analysis*

can be used to extract topical attributes from the content of social media data. Further, it could be used to infer a certain type of activity, whenever a post is not directly linked with a specific POI category.

The techniques that can be used to extract topical attributes from social media depend on the nature of the platform in question. Microblogging platforms, such as Twitter and Sina Weibo, primarily contain textual posts, expressed in natural language. Thereby, natural language modeling and processing techniques can be used in this regard. Conversely, platforms such as Instagram and Flickr, are mainly used for photo-sharing purposes, but also give users the possibility to accompany photo posts with textual content and hashtags. In this particular case, semantic analysis focuses predominantly on the the textual elements of the post. Keyword and hashtag-based filtering comprise the most frequently used techniques in this regard (Becker, Naaman, & Gravano, 2011). The mapping of keywords to ontologies, following the methodology presented in Chapter 3, enables the semantic annotation of social media posts in a machine-processable format, as well as the discovery of semantically similar topical attributes between different posts.

Besides semantics, it also possible to extract *sentiments* from the textual content of social media data. Sentiment analysis can be valuable for the estimation of a person's views and feelings about a particular activity and, therefore, provide new perspectives on aspects of human behavior in cities.

To explore and extract sentiments from human-generated textual content, established taxonomies are required, to enable the classification of concepts into general categories of emotions (Q. Gao, 2013). An exemplary taxonomy thereof is the one developed by Ekman (Ekman, 1972), which identifies six general types of emotions. These are namely: anger, disgust, fear, happiness, sadness, and surprise. This taxonomy can be used to automatically detect emotions in microblogging streams (e.g. Twitter) (Purver & Battersby, 2012).

In accurately inferring the sentiments of people, the main challenges are the limited length of textual content, in combination with the extensive use of informal language and abbreviations. Substantial filtering is, therefore, required in order for very short terms – often referred to as "stop words" – to be excluded. Common methods for the extraction and analysis of sentiments include keyword spotting, lexical affinity, statistical methods, and concept-based approaches (Cambria, Schuller, Xia, & Havasi, 2013), the detailed explanation of which is beyond the scope of this thesis. The approximation of sentiments from social media streams has potential to provided insights into the behavioral aspects of human activity.

§ 5.4 Summary and Conclusions

Contemporary sources of social urban data offer new possibilities for the analysis, measurement, modeling, and characterization of urban space, and can be used as proxies for gaining knowledge about the spatiotemporal dynamics of human activity.

This chapter focused on different types of attributes that can be derived from social urban data and, in particular, geo-enabled social media and LBSNs. It described methods and techniques that enable the extraction of socio-demographic attributes of individuals, functional attributes of places, individual spatial movement patterns, and topical attributes of human activity from social media content. The incorporation of these attributes into urban analytics helps deviate from traditional approaches, in which people and places are usually perceived as aggregate (i.e. average, mean, or summed values) parameters within spatial subdivisions (e.g. census tracts). Further, the chapter presented how the derived attributes help measure the functional density and diversity of urban areas, as well as the geographical extents of activity spaces over different periods of time. These measurements, along with the various methods and techniques for attribute extraction, have potential to improve urban modeling, simulation, as well as planning and decision support systems.

However, the inherent diversity and biases of social urban data and, in particular, geo-enabled social media and LBSNs, pose challenges of accuracy with regard to the approximated attributes. Accuracy levels may vary according to the platform in focus and the policy it follows in terms of data sharing, modeling, and geo-referencing. Validation by means of cross-checking with traditional urban data – where available – is therefore recommended.

The set of attributes described here demonstrate that, besides space and time tags, social urban data contain multiple information in relation to different aspects of human movement, human activity and social behavior, and urban space that, if accommodated in urban analytics, can provide richer descriptions of urban dynamics. The methods and techniques described in this chapter set the foundation for the design and are implemented in a system for the visualization, exploration, and analysis of the spatiotemporal dynamics of human activity that is presented in the following chapter.

TOC

6 Designing and Implementing a System for the Visualization and Exploration of the Spatiotemporal Dynamics of Human Activity in Cities[48]

§ 6.1 Introduction

Throughout this thesis, it has been stressed that the study of complex dynamic processes in cities at various scales, requires the integration of data from several different sources. A key issue of such processes is to understand how people interact with the city – in fact, with the various components that comprise a city – and with each other. Human mobility, activity patterns, and socio-spatial interactions play a pivotal role in the establishment of planning strategies and policies related to land use, transport, and infrastructure configuration. In deciphering the laws governing these processes, the emergence of new data sources (e.g. sensor networks, mobile phones, social media etc.) provide additional viewpoints to those extracted from traditional urban datasets (Lazer et al., 2009).

The increasing availability of emerging data sources has recently instigated numerous research studies, covering a variety of aspects pertinent to human mobility and interactions. Large-scale trajectory data from sensors and GPS devices (Bazzani et al., 2010; Giannotti et al., 2011) or taxi trips (Sagarra, Szell, Santi, Diaz-Guilera, & Ratti, 2015) have been used to unveil aspects of human mobility in cities at disaggregate levels. Similarly, public transportation records from RFID cards have also proven

48 Sect. 6.3 – 6.4 of this chapter are largely based on the following publications:
Psyllidis, A., Bozzon, A., Bocconi, S., & Titos Bolivar, C. (2015). *Harnessing Heterogeneous Social Data to Explore, Monitor, and Visualize Urban Dynamics*. In Proc.: 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015), Cambridge, MA, USA: MIT, pp. 239:1–22. (Main author, 95% contribution)
Psyllidis, A., Bozzon, A., Bocconi, S., & Titos Bolivar, C. (2015). A Platform for Urban Analytics and Semantic Data Integration in City Planning. In G. Celani, D. M. Sperling, & J. M. S. Franco (Eds.) *Computer-Aided Architectural Design Futures – New Technologies and the Future of the Built Environment* (CCIS, Vol. 527). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 21–36. (Main author, 95% contribution)

beneficial, in this regard. Analyses on such datasets have revealed strong diversities as regards the volume of intra-urban flows (Roth et al., 2011) across temporal scales (Zhong et al., 2016), similarities in individual interactions across different contexts (Cattuto et al., 2010), or spatiotemporal patterns that can be used in models for the prediction of individual visits to certain locations (Hasan, Schneider, Ukkusuri, & González, 2012). Empirical data from the geographic circulation of banknotes have also been used as proxies for mathematically describing human travelling behavior (Brockmann, Hufnagel, & Geisel, 2006), and for assessing the extent to which current territorial subdivisions and administrative boundaries reflect the organizational structure of present-day human connectivity (Thiemann et al., 2010).

In inferring the features that characterize human interactions and mobility patterns, the majority of recent studies have employed primarily mobile phone call detail records (CDRs), from the palette of emerging social urban data. Some of the valuable insights relate to the discovery of spatiotemporal regularities in individual human trajectory patterns (Gonzalez et al., 2008) and mobility behavior (Bayir et al., 2009), the appearance of temporal co-locations in physical space between people who belong to the same social network, irrespective of distance (Calabrese, Smoreda, et al., 2011), the influence of urban morphology on the distribution of intra-urban travels (Kang et al., 2012), and to the extraction of attributes that can be used in transportation models (Alhasoun et al., 2014; Calabrese et al., 2013; Diao et al., 2015). Relevant analyses have inferred the spatial cohesiveness of regions across several countries, in relation to their geo-political boundaries (S. Sobolevsky et al., 2013), the influence of social segregation on human mobility patterns in both developed and developing countries (Amini et al., 2014), commuting patterns across countries and global cities (Grauwin et al., 2015; Kung, Greco, Sobolevsky, & Ratti, 2014), and to the impact of city size on human interactions (Schlapfer et al., 2014).

Increasingly, more attention has turned to data from various LBSNs, especially because of their semantic richness and ability to provide information about both spatial and social properties of human activities (Noulas et al., 2011; Scellato et al., 2011). Social media data can, in effect, be used as proxies for investigating the spatial distribution of social activity and its evolution over time. From a network perspective, the analysis of LBSN data can shed light on a significant type of urban network, in particular the *activity network* of cities, which extends the notion of mobility networks. Activity networks comprise interactions between places of social activity, which in turn accommodate the networks of social interactions between people. It is generally difficult to explore such complex socio-spatial networks using the data sources that were mentioned in the previous paragraphs, since they lack semantic annotations, such as place-based geo-tags and comments enriched with topics. The growing pervasiveness of geo-enabled social media and LBSNs, in combination with the relatively accessible APIs, has allowed the performance of studies about activity networks across several urban systems (Noulas et al., 2012), and nation-

wide explorations of inter-urban travel patterns (Liu, Sui, Kang, & Gao, 2014). The semantic richness characterizing the content that accompanies LBSN records allows the extraction of information about POIs, which can be fed into LUTI models (Jiang et al., 2015), the approximation of social activities in urban space at high spatiotemporal resolution (Steiger et al., 2015), the measurement of city attractiveness for particular groups of people (Stanislav Sobolevsky et al., 2015), and the analysis of intra-neighborhood deprivation and social segregation (Quercia & Sáez-Trumper, 2014; Shelton et al., 2015).

Despite the multiplicity of insights into several aspects of human mobility and activity patterns, the results are generally prone to limitations and biases pertinent to each of the aforementioned data sources. For instance, GPS track records, sensor and RFID card data usually stem from a single provider and subsequently cover a certain type of transport modality. Mobile phone CDRs comprise proprietary sources that are also acquired from a single operator, and their representativeness is dependent on the operator's user penetration rates. Conversely, data from LBSNs can be rather "noisy", have several limitations in terms of data lifespan and changing API policies, and may insufficiently represent certain societal groups. And traditional urban data sources (e.g. censuses, surveys, socio-demographic records etc.) have infrequent update rates, varying spatial resolutions, and, in many cases, are not publicly available. These limitations have already been discussed at length in the previous Chapters (especially in Chapter 2). One way to mitigate the limitations and biases is the consideration of more than one data source. However, this requires data integration, which – as discussed in Chapter 3 – is not a straightforward procedure.

The aforementioned challenges motivate the need for tools that enable the simultaneous combination of urban data from various sources, to allow for enriched urban analytics. To this end, this Chapter presents the design and implementation of a web-based system – coined *SocialGlass*[49] – that facilitates the integration of heterogeneous social urban data, and enables the exploration and visualization of human activity patterns in cities, at various spatial and temporal scales. In its current implementation, the system combines data from several geo-enabled social media (namely, Twitter, Instagram, and Sina Weibo) and LBSNs (i.e. Foursquare), publicly available socio-economic urban data (presently focusing on Dutch urban systems), data from web-enabled sensor networks, and further provides mechanisms for the incorporation of custom sources. Data integration processes are founded upon the methodology described in Chapters 3 and 4. In addition, the system incorporates modules that extract and analyze the set of attributes described in Chapter 5

49 *SocialGlass* has been designed and developed in collaboration with Dr. Alessandro Bozzon, Dr. Stefano Bocconi, and Christiaan Titos Bolivar of the Web Information System (WIS) group, EEMCS, Delft University of Technology (TU Delft).

(e.g. home location, socio-demographic attributes, individual trajectory, radius of gyration etc.). Besides its module-based backend structure, the system provides an interactive map-based UI with various visualization and filtering possibilities, alongside a dashboard for real-time monitoring of social activity and flows. The system relies entirely on open-source technologies and open data.

In the remainder of the Chapter, the proxies for attributes of social activity that are used by the system are first introduced (Sect. 6.2). Next, the data sources that are (or could be) integrated into *SocialGlass* are described (Sect. 6.3). Sect. 6.4 presents the various components and modules comprising the system architecture. In Sect. 6.5, an instance of the system is put to use in a real-world case study, to assess the capacities and limitations of *SocialGlass* in exploring the dynamics of urban social activity. An additional spatial analysis on the obtained findings is also presented. The Chapter concludes by reflecting on the flaws and benefits of the system.

§ 6.2 Proxies for Attributes of Social Activity in Urban Space

The spatial patterns of social activity entail a multiplicity of interactions between people, places, and people with places that are built on different types of physical or social networks. In turn, each of the interacting components is further characterized by several attributes that distinguish it from others and also determine the type of activity and the type of relation between them (Michael Batty, 2013b; Herrera-Yague et al., 2015). Different sources of data represent differently the attributes that characterize the interacting components (i.e. people and places). Authoritative records, such as population censuses and socio-economic data still remain the most accurate and trustworthy source of information with regard to these real-world attributes. Conversely, data from online social media indicate a partial image of reality, using different notions to characterize real-world objects. To a great extent, this has to do with the nature of social media; in particular, the fact that they were designed to serve purposes, different than that of spatial analysis. Yet, inasmuch as the system that is presented in this chapter incorporates and integrates both of these sources, there is a need for a mapping between the real-world attributes of social activity and their approximations in social media data.

The introduced proxies revolve around two basic factors of urban social activity; people and places. Respectively, the level of disaggregation is by individuals and activity locations. The latter are characterized by certain positions in space, as well as by a type of activity, which in turn relates to how the different components (i.e. places and people) interact with one another. The most representative approximation of

an activity location within an online location-based network is a POI. In turn, POI categories operate as proxies for activity types. Individuals, on the other hand, are characterized by several socio-demographic attributes that are attached to them. In social media data, these can be approximated to a certain degree through the methods discussed in Chapter 5 and also described later in this chapter. Additional levels of disaggregation include the types of interaction that can be established between the two fundamental components. These specifically refer to interactions of individuals with activity locations, interactions between individuals, and interactions between activity locations. The taxonomy is, therefore, extended to include the proxies for the attributes that characterize the aforementioned interaction categories (Table 15). The entire set of introduced proxies is subsequently translated and organized into modules that comprise the system architecture of *SocialGlass*, as discussed in the following sections.

TABLE 15 Proxies for attributes of social activity in urban space.

Level of Disaggregation	Attribute (real-world)	Proxy (LBSN)
Activity location	Place name ("Third" place)	POI
	Activity type	POI category
Individual	Age	Profile picture (est. age range)
	Gender	Name & Profile picture (est.)
	Ethnicity	Name & Posts language (est.)
	Place of residence	Estimated Home Location (HL)
	Social category:	
	Resident	HL ∈ (City ∩ Country in focus)
	Non-resident	HL ∉ City | HL ∈ Country in focus
	Foreign visitor	HL ∉ (City ∩ Country in focus)
Interaction 1: Individual with Activity location		
	Visit	Check-in (Post)
	Unique visit	Post ID
	Time	Timestamp
	Location	Geo-tag
	Activity category	POI category
	Individual	LBSN ID
	Individual demographics	Age range (estimated)
		Gender (estimated)
		Ethnicity (estimated)
	Individual role	Resident (est. home location)
		Non-resident
		Foreign visitor
	Activity type	Topic (content semantics)
	Opinion	Sentiment (content semantics)

>>>

TABLE 15 Proxies for attributes of social activity in urban space.

Level of Disaggregation	Attribute (real-world)	Proxy (LBSN)
Interaction 2: Individual with Individual		
	Type of social tie	LBSN contact
	Individual demographics	Age range (estimated)
		Gender (estimated)
		Ethnicity (estimated)
	Individual role	Resident (est. home location)
		Non-resident
		Foreign visitor
Interaction 3: Activity location with Activity location		
	Link	Path segment
	Flow	Path segment weight (number of
		OD visits)
	Activity category	POI category

§ 6.3 Integrating Heterogeneous Data Sources

Human activity and its distribution over space and time is a complex phenomenon that merges together features of both the social and the spatial sphere of the urban system. It primarily refers to how people make use and experience urban space. This involves the daily trajectories of individuals around the city (i.e. human movement), which in turn determine the volume of connectivity between places, i.e. the spatial flows. But it also relates to – and is often affected by – the social connectivity (i.e. social interactions) between individuals who perform these activities over space and time (Grabowicz, Ramasco, Gonçalves, & Eguíluz, 2014; Toole et al., 2015; Wang et al., 2015). Moreover, as regards the aspect of experience, people's sentiments and opinions also play an important role. Activities are not necessarily in accordance with the function of the place. An example of this could be a café, where both leisure and study activities can be accommodated. To explore such a complex socio-spatial and dynamic phenomenon, it would be insufficient to use data from a single source. Thereby, the combination of multiple sources of social urban data is deemed necessary.

On the one hand, this poses great challenges to interoperability. As discussed in Chapter 2, social urban data are generally characterized by different levels of

TOC

completeness, representativeness, resolution, timeliness, semantic expressiveness, and trustworthiness, depending on the source that generates them. These in turn result in various syntactic, schematic, and semantic heterogeneities (see Sect. 3.2.1). On the other hand, the combination of heterogeneous data allows to mitigate one source's weaknesses, by using information included in another source. Drawing on this, the backbone of the *SocialGlass* system is data integration. In its current implementation, *SocialGlass* merges together three major types of sources, namely official open data repositories, sensor networks, and online location-based social media (i.e. geo-enabled social media and LBSNs).

First, open data from official repositories can be manually uploaded to the system. In the case where such repositories are supported by an API, custom requests can be made through the system to retrieve the data needed. However, the requests are carried out in the backend, which means that a user cannot simply call a specific API from the system's frontend. The Amsterdam instance of the CitySDK Linked Data API[50] is one such case, from which real-time data about parking capacity are retrieved and incorporated into the system. Socio-economic and demographic information is extracted from publicly available census data by the Dutch Central Bureau of Statistics (CBS)[51]. The currently implemented instance of *SocialGlass* contains data from the census of population in 2011. This specifically includes the population of residents, classified by age, gender, and income, along with their geographic distribution over census divisions. Information on crime rate is also included. Unlike CitySDK data, the CBS records cover the whole national region of the Netherlands and, hence, the entire Dutch urban system. However, this does not mean that the system can be used only in the context of Dutch cities. It can also accommodate relevant datasets from virtually any city worldwide, provided that they are first uploaded to the system. Nevertheless, census and other related data from open repositories do not comprise appropriate sources for retrieving information about human activities. Their main role is to provide aggregate information about the socio-economic and demographic characteristics of urban areas and their population and, hence, give context to social activity data.

Second, GPS traces extracted from sensor networks can also be valuable, when it comes to flow volumes between places, even in real time. Therefore, *SocialGlass* incorporates mechanisms to additionally support this type of source in its architecture (in particular, its real-time dashboard instance). However, it should be noted that sensor data are more appropriate for analyzing human movement than human activity. The lack of semantics or any other contextual information about the monitored elements, can only address aggregate observations on human movement and space occupation.

50 http://citysdk.waag.org. Accessed on April 3, 2016.

51 http://www.cbs.nl. Accessed on April 3, 2016.

Third, various geo-enabled social media and LBSNs constitute the most significant type of source in the *SocialGlass* system for deriving human activity data. In general, using online LBSN data one can extract information about the spatial distribution of social activity over time, and about the social connectivity between individuals (inferred by online "friendships"). It is also possible, through analysis, to derive information about topics, opinions, and sentiments in relation to the performed activity. The currently implemented instance of the system focuses primarily on the spatiotemporal dynamics of human activity, as well as the semantics and sentiments of human social activity. To this end, it presently contains mechanisms to integrate data from Twitter, Instagram, and Sina Weibo. Also, POI-related data are derived from Foursquare. Each one of these platforms is used for communicating different facets of social activity and, thereby, provides different opportunities to infer how people make use of the city. Twitter and Weibo are microblogging platforms whose posts mainly comprise short messages that can be accompanied by pictures or other media. Instagram is primarily an image-sharing platform, which also supports geo-location and textual descriptions. Foursquare is a local search and discovery service that is also used for place recommendation. User penetration varies from platform to platform, as well as from one geographical region to another. For instance, Sina Weibo is highly popular in China, but it is minimally adopted elsewhere. Therefore, each LBSN source has different levels of sample representativeness. Moreover, as the system takes into account only geo-located posts, the sample of available data dramatically decreases in volume. For this reason, *SocialGlass* integrates data from more than one social media platform, while giving further possibilities to incorporate new ones. In so doing, the greatest challenge is to simultaneously accommodate the different data crawling possibilities, based on each platform's API. Also, unlike the first two types of sources, the attributes derived from online social media are approximations of real-world features. One of the main requirements of the system is to provide mechanisms that support the extraction of the proxies presented in Table 6.1, and were described in the previous chapter.

§ 6.4 System Architecture

§ 6.4.1 Components

In implementing the system, a modular architecture[52] is employed to accommodate the various steps from data crawling to visualization (Figure 25). The basic idea behind it, is that different tasks of the process are assigned to different software modules. In turn, the functionality performed by each module determines its connection patterns, that is, the other modules that it directly communicates with. The inter-module communication is achieved by means of message queues (Silberschatz, Galvin, & Gagne, 2009). This means that each module receives information from related ones (e.g. modules that perform certain tasks pertinent to a particular process) in the form of messages. Then, after processing the received message, it may also send messages to other modules that expect this information to complete a certain procedure. In this way, the system can be quite easily extended with new modules, performing tasks that might have not been considered initially (e.g. to integrate a new source, such as an additional social media platform), without posing significant challenges to the already existing structure and system functionality.

The first step in the process is about data ingestion and retrieval. The former comprises mechanisms that support data upload (e.g. census data), whereas the latter relates to the process of data crawling (e.g. from social media APIs or from APIs of official repositories). Next, after filtering and storing the retrieved data in a database, semantic integration processes are carried out. This helps to enrich the extracted information and proxies, following an ontology-based integration approach (see also Chapter 3). The following step concerns the visualization of the extracted information, using multiple types of data visualization in the form of layers, on top of a map-based UI. Finally, in the case of monitoring activity patterns in real time, the various datasets are visualized in a dashboard format, containing several widgets. The creation, editing, and monitoring of running analyses can be carried out through an admin interface.

In accordance with the four aforementioned general steps, the various modules that comprise the system are organized into four main components. These respectively cater to (a) data ingestion and analysis, (b) semantic enrichment and integration, (c)

52 The system architecture documentation of the SocialGlass system is available on GitHub, at the following link: https://github.com/WISDelft/SocialGlass/wiki/Development-Guide-2.-Architecture. Accessed on April 5, 2016.

exploration and visualization, and (d) real-time monitoring. The first two components constitute the backend of the system, whereas the remaining two pertain to its frontend.

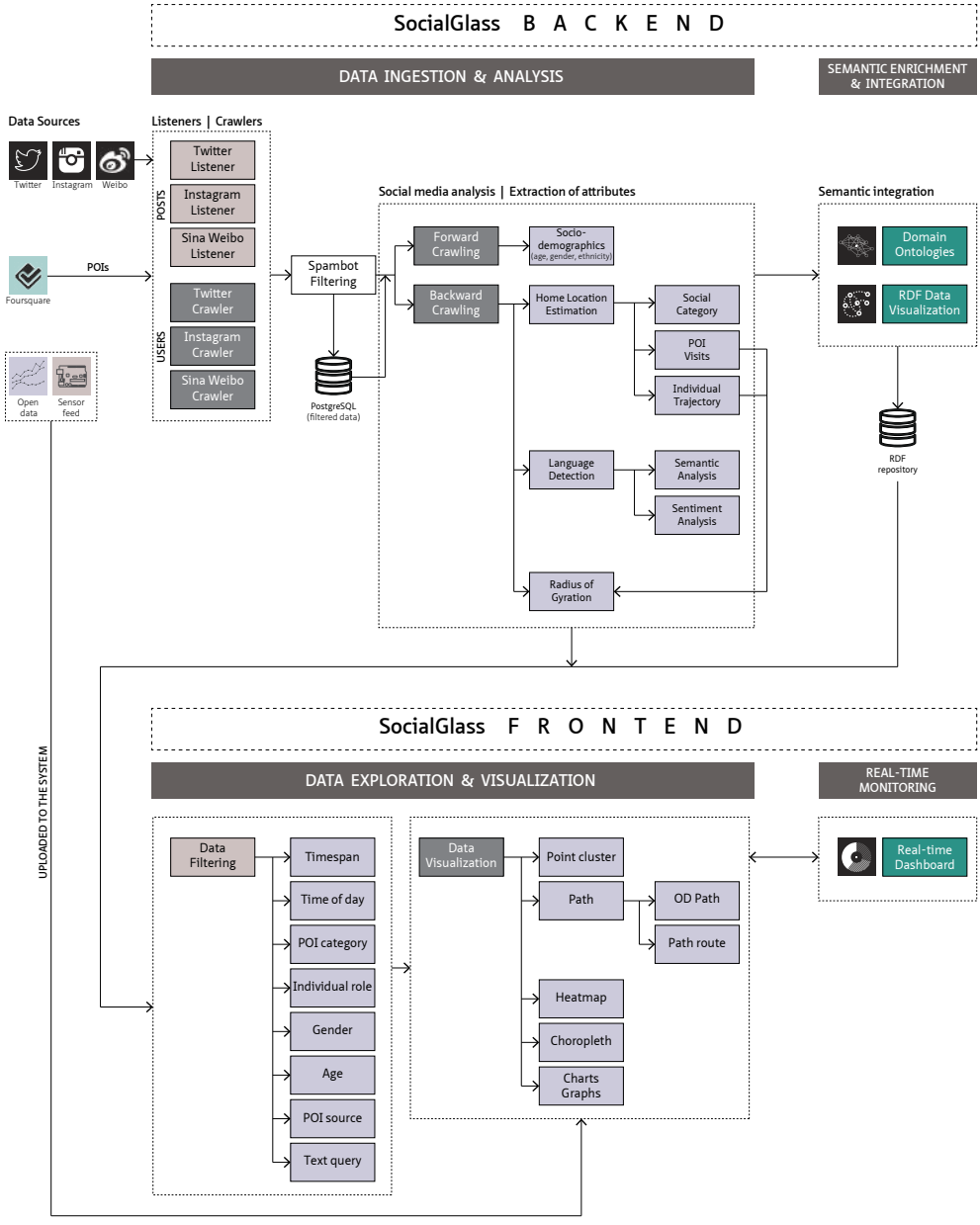


FIGURE 25 System architecture of the *SocialGlass* system (components and modules).

First, the *data ingestion and analysis component* accommodates modules that perform the tasks of data collection, filtering (cleansing), storage, user modeling, and mapping to specific POIs within a city in question. The collection of data from online social media is done by means of listener and crawler modules. At present, the system consists of three listeners that respectively retrieve data from the APIs of Twitter, Instagram, and Sina Weibo. The latter is mainly used in the context of Chinese urban systems. Since the focus is on the geographical distribution of human activities, the system takes into account only geo-tagged posts (either accompanied by an exact geo-location or mapped to a specific POI) from the aforementioned social media. In addition, data retrieval is limited to the posts that are generated within a predefined area of interest. This could be specified either through a bounding box (in the case of Twitter and Sina Weibo) or through a circle with a maximum radius of 5km (in the case of Instagram), according to the respective API documentations. For larger cities, multiple circles can be specified to cover wider regions. The incoming data are filtered, so that it is ensured they were generated by people and not by software bots. The filtered data are then stored in an open source relational database, namely PostgreSQL, to be further processed by specific modules that are designed to extract certain attributes (e.g. individual demographics, approximation of home location, calculation of the radius of gyration, origin-destination (OD) path extraction etc.). As stated in the previous section, the attributes are disaggregated at the level of individuals and POIs. Therefore, the component contains modules for the extraction of individual socio-demographic attributes from social media (e.g. age, gender, ethnicity, home location); the mapping of individual activity to specific POIs (e.g. to categories included in the Foursquare API); the calculation of mobility and spatial interaction metrics (e.g. radius of gyration, OD paths), and the analysis of semantics and sentiments included in a post. The extraction and analysis of the aforementioned attributes, with the exception of socio-demographic variables, is mainly based on the history of posts generated by an individual. Therefore, the corresponding modules can perform their tasks only after the various crawlers have received a sufficient amount of posts. The modules are analyzed in further detail in the following section.

Second, the *semantic enrichment and integration component* caters to the integration of the heterogeneous source data. The employed approach is based on domain ontologies that can be uploaded to the system. The incoming data are then mapped to ontology classes, following the methodology described in Chapter 3. The generated RDF triples are stored in a persistent repository and can be queried through a dedicated SPARQL endpoint. Following up on this step, the semantically annotated datasets (posts) are fed into modules of the first component, for further analysis. The data ingestion and analysis, together with the semantic enrichment and integration component comprise the backend of the *SocialGlass* system.

Third, the *data exploration and visualization component* is one of the two components that comprise the system's frontend. Its implementation consists of an online

interactive map-based user interface, which accommodates several options to visualize and filter the various collected data. The system organizes different types of visualization and data sources into layers (see Appendix C, Figures 32 – 34). This means that one can plot the data from a single or multiple sources of interest on top of the map illustrating the area in question, using different layers for each plot type. In this way, it is possible to compare multiple plots on top of one another, which may pertain to different sources and depict different attributes. The currently implemented plot types include dynamic point clusters, activity heat maps, OD paths, path routes, and various choropleth visualizations. The dynamic point cluster plots are more appropriate for illustrating the amount of social media posts generated from a given urban region, by dynamically distributing them over more disaggregate spatial divisions as one gradually zooms in (see Appendix C, Figure 35). Conversely, the activity heat maps show the aggregate intensity of social activity in geographic space, over certain periods of time. Through the use of time sliders, it is possible to explore the change in activity patterns in the course of a day (see Appendix C, Figure 36). With regard to exploring human movement, the system currently provides two plot types, namely OD paths and path routes (i.e. individual trajectories). The former creates a spatial network, where nodes represent POIs and links (arc edges) represent the interaction between the POIs, as extracted from the online visits of individuals. Besides interactions, the paths also accommodate flows, represented by the weights on the arcs of the graph (i.e. larger flows result in larger edge thickness and color density), indicating the amount of trips from a certain POI to another (see Appendix C, Figure 37). Conversely, path routes, plot the interactions between POIs (i.e. visiting patterns of an individual) on top of the street network of the city in question. Therefore, the edges of the network illustrate the exact route followed by a person from one POI to another (see Appendix C, Figure 38). Again, the line thickness and color density indicate the volume of connectivity (i.e. flows) between POIs. The system further supports choropleth maps, in which spatial subdivisions (i.e. neighborhood or district polygons) are shaded in proportion to the value of data attributes. These may refer to the geographic distribution of posts, POIs, and various socio-demographic attributes of individuals (e.g. age, gender, resident, commuter, foreign tourist etc.). Moreover, data can be filtered by POI category, time span, type of individual, gender, age, or queried by keywords. Also, several graph representations are supported to provide additional information about the daily distribution of social activities, the semantics of posts, and the socio-demographic composition of spatial subdivisions (see Appendix C, Figure 39).

Finally, the *real-time monitoring component* is complementary to the previous one, and specifically caters for the (near) real-time visualization of data streams retrieved from sensors and various online social networks. This particular component mainly consists of a dashboard that employs several graphical representations of the collected data (e.g. charts, timelines, word clouds, choropleths etc.), and operates in parallel to the map-based interface. In general, the real-time component is mostly

suited to continuous measurements that only relate to specific aspects of the city (e.g. human flows), rather than to urban analytics, which require further abstraction and aggregation. Overall, the manipulation of the system, including the initiation of new empirical experiments and the upload of custom data, is carried out through a dedicated admin interface.

§ 6.4.2 Organizing Proxies into Modules

Each of the above-described components comprises several modules that perform specific tasks. Besides data retrieval, the majority of the implemented modules infer attributes of social activity from social media data, by using the proxies listed in Table 15. In this section, the various modules that comprise the *SocialGlass* system are described in further detail (see Figure 25).

The key role among the modules is played by the *listeners* and *crawlers*, as they are responsible for feeding data into the system. The former retrieve posts from the APIs of online social networks, whereas the latter focus specifically on individuals (i.e. social media users) and their post history. The system currently combines data from three geo-enabled social media, namely Twitter, Instagram, and Sina Weibo. Respectively, three listeners and three crawlers are implemented. To initiate the data retrieval, a geographic area needs to be defined first. For Twitter, this would be a bounding box, surrounding the area in focus. For Instagram, a circle (or several circles, each) with a maximum radius of 5,000m is defined instead, according to the API requirements. Similarly, for Sina Weibo, a circle (or several circles, each) with a maximum radius of 11,132m has to be drawn. In the case of large cities or metropolitan regions, several overlapping circles have to be defined to cover the entire area in question.

Although Instagram and Sina Weibo allow the retrieval of the entire post history of the users within the predefined areas, the Twitter streaming API (i.e. publicly available data feed) provides a 1% sample of the entire set of public tweets included in the so-called 'firehose' (proprietary feed). Despite this, a recent study by (Morstatter et al., 2013) has shown that the Twitter streaming API returns almost the entire set of geo-tagged posts within the predefined area. Therefore, geo-tagged tweets, which are of interest to the system, are well represented in the given sample.

The collected posts are subsequently filtered, in order to exclude those that have been generated by software bots. Then, based on each unique user ID (detected by the listeners) from each one of the three social media platforms, the corresponding crawler modules perform two types of data crawling; namely, forward and backward. The former, retrieves posts that are generated by a person after the start of the data

collection period. Contrariwise, in backward crawling the focus is on the person's post history, that is, the posts that have been generated prior to the starting point of the data collection period. The collected posts are sent by means of messages to one or several queues for the extraction of specific attributes and the calculation of activity-related spatial metrics.

The remaining modules of the system, mainly under the data ingestion and analysis component, are dedicated to extracting attributes of an individual and of the distribution of her/his activity over space and time. These attributes include the person's demographics, role, and home location. Also, the modules extract the places s/he has visited and the links (paths) between these places, calculate the radius of gyration, and analyze topics and sentiments about the various activities performed. However, the current implementation of the system has no specific modules, dedicated to deriving the social contacts of an individual.

At the level of individuals, the system approximates *demographic* attributes, such as a person's age range, gender, ethnicity, home location, and role (see also Table 15). The first three attributes (i.e. age, gender, and ethnicity) can be inferred immediately after the listeners identify a new user. On the contrary, the remaining attributes require a certain amount of posts to be retrieved through backward crawling (i.e. a person's post history), before being estimated and assigned to the person in focus. For the estimation of an individual's age and gender, the demographics module uses as proxies the first name and profile picture of a person's social media account. This information is sent to three external open-source web services[53] that respectively determine the gender, age range, and country of origin, along with a confidence interval, and return the results back to the system. For privacy issues, the name and profile picture do not remain into the system's database and only the values of the returned results are taken into account for further processing. Thus, individuals are solely identified by an anonymized user ID and the values of the three aforementioned attributes.

For the estimation of a person's *home location*, the corresponding module carrying out this task implements the method defined in Sect. 5.3.1.1 and, in particular, the recursive grid search by means of geohashes (see also Figures 20 – 21). The iterative

53 To determine the gender of an individual, the demographics module sends the first name to the open-source web service *Genderize* (https://genderize.io). In addition, the module sends the profile picture of a person's account to an open-source web service for face detection with the name *Face++* (http://www.faceplusplus.com). The latter gives back an age range and gender estimation with high estimation precision (generally >90%). As regards the gender estimation, in the case of conflicting results between the two services, the corresponding values are first sent to a *Decision Tree* that is trained on human-annotated data. Then, only the outcome of this process is sent back to the demographics module. With regard to the estimation of the country of origin, the module retrieves the "country" field from a person's account and sends it to *GeoNames* (http://www.geonames.org) (see also Chapter 3) to ensure that the retrieved value corresponds to a real place.

clustering of a person's geo-tagged posts (retrieved from the post history through backward crawling) into geohashes of increasing granularity allows finer accuracy than simply considering home location as the center of mass of all posts a person generates. In addition, the module takes mostly into account the posts that have been created between 6pm and 8am, which would most likely originate from a person's actual home location. Drawing on the approximated home location, an additional module receives this piece of information and determines the *role* of an individual in relation to the city in focus. The system enables a classification into three different roles, namely residents, non-residents, and foreign visitors. When the estimated home location is placed within the predefined boundaries, then a person is inferred a resident of this region. In any other case, the person in focus is considered a non-resident, when the approximated home location is within the same country that the predefined area belongs to; or a foreign visitor when none of the previous two conditions applies.

The second set of attributes pertains to places and, in particular, *POIs*. Besides home location, which is considered a "first place" in an individual's activity space, POIs constitute "third places", where people most likely interact with one another. These places are the major condensers of social activity in cities and, thereby, constitute an essential indicator of social life and activity behavior in urban space (Rosenbaum, 2006). At the moment, places of employment or second places can only be inferred indirectly by the system, through text mapping and topical analysis, which will be later discussed. The module responsible for the extraction of POIs from the collected posts, retrieves relevant information from four different geo-enabled and location-based social media; in particular, Twitter, Instagram, Sina Weibo, and Foursquare. This information comprises the set of latitude and longitude coordinates of a POI, the POI name and category, as well as its popularity among social media users (based on a ranking algorithm). The aforementioned attributes are assigned to a unique POI ID. The module implements mappings (of posts) to certain POIs, by means of various mappers (e.g. a PostGIS mapper that maps posts on the basis of PostGIS queries to the PostgreSQL database, two foursquare mappers for Twitter and Instagram posts respectively, and a Weibo mapper). The extracted information is then sent to other modules in the pipeline that carry out more specialized tasks.

The extraction of POIs from the various collected posts allows the measurement of mobility-related aspects of human activity. These measurements provide further insights into the geographical extent of individual activity spaces and the volume of connectivity between places. Respectively, the system accommodates two modules for the calculation of the radius of gyration and for the extraction of OD paths and flows. The *radius of gyration* is calculated by the corresponding module, according to the equation defined in Sect. 5.3.3.3, in particular:

$$R_g^u(t) = \sqrt{\frac{1}{n_p^u(t)} \sum_{i=1}^{n_p^u(t)} (x_i - x_c)^2 + (y_i - y_c)^2} \tag{6.1}$$

Where, $n_p^u(t)$ represents the total number of posts generated by an individual within a period of time t, x_i and y_i are the coordinates of a post in location i, and x_c and y_c represent the center of the trajectory that is created by joining the single posts together, and are defined as follows:

$$x_c = \frac{1}{n_p^u(t)} \sum_{i=1}^{n_p^u(t)} x_i \tag{6.2}$$

$$y_c = \frac{1}{n_p^u(t)} \sum_{i=1}^{n_p^u(t)} y_i \tag{6.3}$$

As can be drawn from the equations above, the generation of new posts by an individual largely influences the radius of gyration. Therefore, the module iteratively recalculates the radius of gyration, each time a new post is created by the person in focus (see also Figure 24).

Aside from the radius of gyration, the system further extracts the set of individual *daily trajectories* between POIs (see also Figure 22). The trajectories are constructed from the sequence of consecutive posts generated by a person on a given day. Each time an individual creates a geo-tagged post in any of the social media platforms involved, the system records the respective geo-location and is, therefore, able to reconstruct the evolution of an individual's trajectory over time. In addition, the corresponding module measures the intensity of the interaction between POIs, i.e. the flows, by calculating the amount of individuals moving from one place to another (as inferred by their posting activity). However, the trajectories are limited within the area defined by the bounding box. The extracted trajectories and flows are, subsequently, visualized as paths (edges) that either connect directly the various POIs together (curved edges) or follow the street network of a city, as described in the previous section. Taking advantage of both forward and backward crawling, it is possible to extract the entire set of trajectories from a person's post history, as well as those created after the start of the data collection period. One may also configure the level of temporal (dis)aggregation, so that the extracted trajectories represent the hourly or weekly flows. In embedding the

trajectories in the street network of a city (instead of simple connector lines between POIs), the module estimates a number of intermediate waypoints[54] between the origin and destination POIs (see also Sect. 5.3.3.1). In this way, it is possible to make a rough estimate of the route followed by a person, or group of people, when moving from one place to another. However, the accuracy of the route estimation decreases substantially, when the distance and the number of street intersections between two consecutive POIs increase.

Although individual movement patterns are essential to understanding the spatial interactions between places and their evolution over time, it is equally important to gain insight into the type of activity performed in these places. As stated previously, POI categories are indicators of different activities, but may not necessarily reflect the actual activity that is carried out by an individual, or group of people, in any of the POIs. In addressing this challenge, additional modules are implemented, dedicated to *semantic* and *sentiment* analysis. The semantic analysis module extracts words and topics from user-generated posts and maps them to a related type of activity[55], whereas the sentiment analysis module identifies sentiments from the post content that range from highly positive to highly negative[56]. Prior to performing these tasks, a supplementary module first detects the language[57] used in a user's post. This piece of information is subsequently fed into the semantic and sentiment analysis modules. However, in the current implementation, semantic analysis is supported for content that is written only in English. Although, there are fluctuations in terms of accuracy levels of both semantic and sentiment approximations, they have potential to increase the understanding of what type of activities people perform in various places and times of the day.

54 In estimating the intermediate waypoints, the module calls the open-source Google Directions API (https://developers.google.com/maps/documentation/directions/). However, the free and open version of the API allows a maximum of 8 waypoints to be estimated between an origin and a destination. Therefore, complex street networks with multiple intersections largely influence the estimation accuracy.

55 The extracted words are mapped to relevant entities in *DBPedia Spotlight* (https://github.com/dbpedia-spotlight/dbpedia-spotlight). This allows the creation of topic profiles, as well as the inference of activities from the textual structure of posts.

56 At present, sentiment analysis makes use of the *SentiStrength* software (http://sentistrength.wlv.ac.uk), which is freely available for academic use.

57 In identifying the language used in posts, the system makes use of the *shuyo* language detection Java library (https://github.com/shuyo/language-detection).

§ 6.5 Exploring and Analyzing the Distribution of Social Activity over Space and Time

At an intra-urban level, human activity and mobility behavior may largely vary, depending on the time of the day or night and the city district or neighborhood. Correspondingly, city-scale social events potentially generate spatial and temporal fluctuations in the distribution of activity and mobility patterns. Therefore, they can be good examples to test the capacity of the system and its various components for the potential detection of such fluctuations, based on insights from different sources of social urban data. The Amsterdam Light Festival (ALF) 2015, an art-related city-scale event with a duration of two months, is used as a case study to explore the potential impact of the event on the daily activity and mobility behavior of different groups of people. Social media data from heterogeneous sources are used as proxies for the social activity behavior of different social categories and its distribution over space and time. In addition, they are used to explore the potential formation of spatial and temporal patterns of human activities before, during, and after the event. To this end, the system is configured to continuously crawl data not only throughout the event period, but also two weeks before the start and two weeks after the end of the event. This is to detect any variations in the intensity of activity that are likely to occur.

The main hypothesis is that the event influences the overall activity and mobility behavior of all social categories (i.e. residents, non-residents, and foreign tourists), especially in the areas where the event takes place. In particular, it is expected that the volume of activity will increase, in comparison with its corresponding intensity the weeks before and after the event. To test this hypothesis, the system collects and combines data from various social media (*listener* and *crawler* modules) and distributes them spatially and temporally over the postcode area districts of the city of Amsterdam. Subsequently, each of the modules extracts or estimates certain attributes from the collected data, throughout the monitoring period, as discussed in the previous section. Then, the extracted or estimated attribute values are visualized in the system's frontend, providing several possibilities for visual cluster identification, spatial autocorrelation, and temporal fluctuation. Besides these visual exploratory approaches, an additional spatial autocorrelation analysis is performed on the empirical data, to further measure the degree of activity clustering.

§ 6.5.1 Dataset

In exploring the spatial and temporal organization of human activities over a short time period, it is generally difficult to extract relevant information from traditional data sources, such as the census or travel surveys. Although they are highly reliable

sources of information about cities, their rather infrequent updates (e.g. once in ten years) pose a major limitation to the issue in question. Therefore, alternative data sources are needed that are able to relate information about human activities with their distribution over geographical space, at short time scales. Data from LBSNs (e.g. Foursquare) and geo-enabled social media (e.g. Twitter, Instagram) comply with this requirement, although they mainly provide indications of the whereabouts of (a sample of) people, as broached throughout this thesis. Despite this, the public availability of these datasets (by means of calls to the corresponding APIs), which is in agreement with the general principle of openness in terms of the data and technologies used in this thesis, is a considerable advantage. Especially in comparison with related data sources, such as CDRs, which may also approximate activity in geographic space over short time periods, yet at a high cost and with insignificant semantic information (see also Chapter 2). As a result, data from Twitter, Foursquare, and Instagram are used as proxies for the spatiotemporal distribution of human activities in the case study.

The data collection is carried out using the corresponding listener and crawler modules, under the data ingestion and analysis component of the *SocialGlass* system. In total, the observation period covers three months, which correspond to the two-month duration of the event, the two weeks prior to its start date, and the two weeks after the event finishes. More specifically, the overall period is between November 13, 2014 and January 31, 2015, with the starting date of the event on November 27, 2014 and the end date on January 18, 2015. The system collects solely geo-located data that are generated within the city region of Amsterdam, as inferred by the geo-tag or the accompanying POI location, for the aforementioned time period. The resulting datasets comprise 26,740,669 geo-tagged posts from Twitter (linked to Foursquare POIs) and 15,959,566 posts from Instagram. The collected records generally consist of the anonymized user ID, the latitude and longitude coordinates of the POI (retrieved from Twitter and Instagram posts, and aligned with POI information from Foursquare), the POI category, the timestamp, and demographic information, such as gender and city of residence, where available. In the cases where demographic information is missing, it is approximated by the corresponding modules of the system, regarding an individual's home location, role, gender, age range, and country of origin.

In addition to the above, demographic and socio-economic data from the 2011 census (Statistiek, 2011) are also integrated. These include the total population, the number of male and female residents, the age range, and income in the city region of Amsterdam. This information allows comparisons with the records that are collected from the different social media sources, especially in terms of representativeness. For privacy reasons, the demographic and socio-economic data are aggregated into spatial divisions with varying levels of detail. The latter range from the municipal level to the district and neighborhood level (Statistiek, 2014). The chosen level of spatial division for the case in focus is the neighborhood level (i.e. postcode area, based on the

Dutch postcode system), which corresponds to the maximum granularity available. Especially for the performance of spatial autocorrelation analysis (see Sect. 6.5.3), the observations from social media are also aggregated into the same spatial divisions.

§ 6.5.2 Visual Exploratory Analysis of Spatiotemporal Activity and Movement Behavior

The frontend of the system provides several interactive visualization and data filtering possibilities that enable the visual exploratory analysis of spatial and temporal phenomena, related to certain social categories. In particular, a user may represent the collected (and/or uploaded) datasets in the form of dynamic point clusters, intensity heat maps, OD paths, path routes, choropleths, charts, and graphs based on the issue at hand. These visual representations can be filtered by data source, social category, time span, time frame within a day, POI category, age range, and gender type, within a zooming user interface. Each type of visualization is stored in different layers, on top of the base map, to enable the simultaneous observation and exploration of several variables. The currently available visualization types and data filters are listed in Table 16.

Here, the focus is on testing the capacities of the *SocialGlass* components, as regards the visual exploratory analysis of the activity and movement behavior of people over space and time, on the basis of the main hypothesis. Drawing on the collected datasets, described previously, the first step is to choose the data source of interest. The system is built in such a way that it allows a user to investigate and compare the same set of variables from different sources. In the examined case, social activity is approximated by Twitter and Instagram data, in addition to POI-related information extracted from Foursquare. Therefore, these are the main sources in focus.

From the implemented visualization tools, heatmaps, choropleths, charts, and timelines, are used for the exploration of social activity distribution over space and time, whereas path routes are used for the investigation of flows (i.e. human movement behavior). In order to test the main hypothesis, the datasets are further filtered by social category (i.e. residents, non-residents, foreign tourists) and time span (i.e. before, during, and after the event).

Heatmaps provide an overview of hot and cold spots, respectively signifying locations of high or low intensity of social activity. The accompanying time sliders, enable users to study the hourly distribution of activities. In turn, choropleths illustrate this aspect by aggregating data into the geographic divisions (here, postcode areas) of the city, assigning a shade in proportion to the intensity of activity (i.e. number of posts per postcode area). Path routes represent the flows between locations, projected

on the street network of the city (instead of drawing a straight connector between the origins and destinations), using different opacity levels to indicate the volume of movement. Charts and timeline graphs, provide additional information about the temporal distribution of posts, the daily social activity, and metadata about the variable in question (e.g. age range, nationalities, popular venues, semantics etc.). The exploratory results can be exported in GeoJSON format for integration into external GIS or statistical spatial analysis platforms.

TABLE 16 Visualization types and data filters.

Visualization Types		
Type	**Variable**	**Function**
Point cluster	Post	Visualizes point data (i.e. geo-referenced posts) as marker clusters that are dynamically redistributed as one zooms in or out. A time slider is also available to show the distribution of posts on an hourly basis
Path	OD path (arc)	Creates a network of nodes (POIs) and links (arc edges) to represent the interaction between origin and destination POIs. Thickness indicates the volume of flows
	Path route	Creates a network that follows the footprint of streets, indicating the exact route followed by an individual (or group of people) from one POI location to another. Line thickness and color density indicate the volume of flows
Heatmap	Post	Shows the aggregate fluctuations of social activity intensity in geographic space. A time slider illustrates the variations on an hourly basis
Choropleth	Post	Creates a thematic map in which areas are shaded in proportion to the amount of posts. Time sliders are available in all choropleth maps
	POI category	Creates a thematic map in which areas are shaded in proportion to the POI category
	Individual role	Creates a thematic map in which areas are shaded in proportion to the social category (i.e. residents, non-residents, foreign tourists)
	Gender	Creates a thematic map in which areas are shaded in proportion to the gender type
	Age	Creates a thematic map in which areas are shaded in proportion to the age range
Charts, Graphs	Various	A set of auxiliary pie charts, histograms and timelines that present additional information about the percentage of individuals from a social category, the percentage of gender and age rate, the percentage of different nationalities, POI categories, POI ranking based on popularity, daily amount of posts, semantics etc.
Data Filters		
Filter		**Function**
Timespan		Filters the selected data according to a custom time span
Time of day		Filters the selected data according to a time frame within a day
POI category		Filters the selected data according to a POI category
Individual role		Filters the selected data according to a specific social category (i.e. residents, non-residents, foreign tourists)
Gender		Filters the selected data according to the gender type
Age		Filters the selected data according to the age range
POI source		Filters the selected data according to the source from which POIs are extracted (e.g. Foursquare)
Text query		Filters the selected data according to a set of words included in a post (e.g. a tweet)

Residents' activity in the period before (i.e. November 13, 2014 – November 26, 2014) and after the event (i.e. January 19, 2015 – January 31, 2015), as inferred by both Twitter and Instagram, presents similar patterns of intensity in the heatmap illustrations. Conversely, the period during the event is characterized by an exponential increase in the volume of social activity, especially observed in Instagram data. Despite this dissimilarity, clusters of hot spots tend to gather around the areas of the city center, regardless the time period and the differences in terms of overall intensity. However, smaller hot spots also appear in the eastern, southern, and western outskirts of the city, suggesting a slightly dispersed distribution of activity (Appendix D, Figures 40 – 47). This only signifies an overall tendency of activity behavior, which will be specified quantitatively in the spatial autocorrelation analysis, discussed in the following section. By overlaying heatmaps on top of the corresponding choropleth visualizations, one may detect the spatial units around which activity tends to cluster, as well as to have an overview of the degree of dispersion for each social category. A user can also activate or de-activate as many layers as needed in order to visually explore the spatial and temporal distribution of different variables. Besides, the accompanying timelines insight into the daily fluctuation of social activity over the period of study.

Using the same set of visualization tools, the social activity of non-residents presents a much smaller degree of intensity than that of residents (see Appendix D, Figures 46 – 51). Although in this social category, the volume of activity also becomes larger in absolute numbers, its spatial distribution gives evidence of a strong cluster around the areas of the city center. More specifically, almost the entire activity of non-residents appears to concentrate in the districts of Burgwallen-Nieuwe and Oude Zijde, Grachtengordel Zuid, Weteringschans, and Nieuwmarkt. In an interesting way, non-residents appear to collocate in these areas, not only during but also before and after the ALF event.

Evidence of similar clusters around the aforementioned areas, though of much higher intensity, is also given for foreign tourists (see Appendix D, Figures 52 – 57). This activity of this particular social category appears to form agglomerations around POIs of the city center, though on a much larger scale than non-residents and residents. Its spatial dispersion is almost negligible. Yet, what specifically diversifies foreign tourists from the other two social categories, is the fluctuation of activity intensity over the three main periods of study. The spatial footprint of foreign tourists' activity is rather weak in the period before the event, grows exponentially during the event, and remains sufficiently large in the period after the event, as inferred by both Twitter and Instagram. An indicative comparison between the social activity patterns of residents and foreign tourists for the entire period (i.e. November 13, 2014 till January 31, 2015), focusing specifically on their activity dynamics between 6pm and 9pm, is illustrated in Figure 26.

[a] ALF – Platform: Instagram | User type: Resident | Time Period: 18–21h

[b] ALF – Platform: Instagram | User type: Foreign Tourist | Time Period: 18–21h

FIGURE 26 Average activity patterns of (a) *residents* and (b) *foreign tourists* for the entire period before, during, and after the ALF event (between 6pm and 9pm). Residents appear to have a more dispersed activity over space, compared to foreign tourists who tend to cluster around the central districts of Amsterdam (as inferred from Instagram). Moreover, residents' activity appears more balanced throughout the period in focus, whereas in the case of foreign tourists, a steep increase in volume occurs, especially around the Christmas period.

The observations on the spatial and temporal distribution of social activity are also reflected in the movement patterns of the three social categories (Figures 27 – 29). The routes followed by residents and foreign tourists traverse more neighborhoods than those of non-residents, yet the strength of residents' flows is higher (represented in thicker red lines in the flow maps) and flow lines reach areas much farther than the city center. Especially with regard to flows and their representation, the zoom function of the UI enables users to explore different levels of detail, mitigating the negative effects of aggregate flow maps, which become easily cluttered.

The visual exploratory analysis of the collected datasets, using a set of visualization and data filtering tools implemented in *SocialGlass*, allowed to elucidate how the social activity of different groups of people is distributed over the geographic space and over time. With reference to the main hypothesis, it could be inferred that the event has an influence on the volume of activity, but there exist several dissimilarities between the social categories. Therefore, the intensity indeed increases, but what has yet to be tested is whether the spatial distribution of activity in the social categories tends to agglomerate around similar or different neighborhoods over the three examined time periods. This requires to extract the collected data, metadata, and results from the system, to perform further statistical spatial analyses.

FIGURE 27 Movement trajectories of residents throughout the entire period (i.e. November 13, 2014 – January 31, 2015).

FIGURE 28 Movement trajectories of non-residents throughout the entire period (i.e. November 13, 2014 – January 31, 2015).

FIGURE 29 Movement trajectories of foreign tourists throughout the entire period (i.e. November 13, 2014 – January 31, 2015).

§ 6.5.3 Spatial Autocorrelation Analysis

The visual exploratory analysis of several variables in the system's frontend indicated the existence of spatial and temporal clusters of social activity. It also demonstrated differences in the agglomeration of activity between the social categories in focus, that were largely dependent on the time period (i.e. before, during, or after the event). To further quantify the distribution of social activity over space and time, and to assess whether and where the activities of different social categories tend to concentrate around certain areas, a spatial autocorrelation analysis is performed, using the data collected by *SocialGlass*. This exploratory spatial data analysis (ESDA) (Bivand, 2010; Haining, 2003) method is also used to test the main hypothesis, concerning the degree to which ALF influences the activity behavior of different social categories. Indices and tests of autocorrelation, as well as local indicators of spatial association (LISA) are employed to determine potential spatial patterns of similar values in the examined variables. Moreover, they are used to explore the extent to which these patterns are significant from a statistical point of view. Such patterns and associations between values are often difficult to detect in the choropleth or heatmap visualizations, described in the previous section. This analysis is performed by means of the open-source R statistical language (R, 2008), using several packages[58] for spatial analysis (see also Appendix G).

First, the *global Moran's I coefficient* (Cliff & Ord, 1973, 1981; Moran, 1950) is used to identify spatial clusters of high (or low) POI density over the 96 areal units (i.e. postcode areas) of the city of Amsterdam. The POI records derived from Foursquare are used as proxies for the entire set of POIs in the city. Subsequently, the Moran I statistic is applied to the empirical data from Twitter and Instagram, to quantify spatial clusters and outliers of human activity concentration. It is specifically calculated for 27 variables (in addition to the POI density variable, making up a total of 28 different variables) – listed in Table 17 – which reflect the intensity of activity (i.e. number of visits to POIs), normalized by the total area of each spatial unit. The variables address the three social categories examined here (i.e. residents, non-residents, and foreign tourists) and cover different periods of time (i.e. before, during, and after the event). The obtained I-values are then tested against the null hypothesis of complete spatial randomness (CSR), that is, the assumption that the variables are completely spatially independent. These tests assess the likeliness of obtaining I-values that are equal to, or larger than, the calculated ones, in a hypothetical case of no spatial autocorrelation. Finally, local indicators of spatial association (LISAs) (Anselin, 1995), and in particular the *local*

58 For the purposes of the analysis, the following open-source R-packages are used: *GISTools* (http://bit.ly/23o-MYKm), *spdep* (http://bit.ly/1SNsnu2), *lctools* (http://bit.ly/1UyDgVm), *ggplot2* (http://bit.ly/1UyDjAv), *rgeos* (http://bit.ly/1W7kiX5).

Moran's I_i and the Getis-Ord G_i^* statistics, are calculated for each one of the total 28 variables. These local statistical indices provide further indications of significant spatial clusters of either high or low values around each observation (i.e. areal unit).

§ 6.5.3.1 Global spatial autocorrelation statistics and tests

The commonly used mathematical definition of the *global Moran's I* coefficient is the one defined by Cliff and Ord (Cliff & Ord, 1973, 1981), which slightly deviates from the initial version proposed by Moran (Moran, 1950), and is given by the following expression:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{6.4}$$

Where n is the number of areal units (i.e. observations), z_i, z_j are the values of a variable $Z = \{z_1, z_2, \ldots, z_i, z_j, \ldots, z_n\}$ of interest for two neighboring areas i and j respectively, $\bar{z}$ is the mean of the z_i values, and W_{ij} is a n x n weight matrix, specifying the degree of dependency between the areal units i and j. The I-values lie in the [-1, +1] range, where values close to +1 indicate strong positive autocorrelation (i.e. spatial clusters of similarly high or low values of the examined variable), values close to -1 indicate strong negative autocorrelation (i.e. areal units of high z_i-values close to districts of low z_j-values), and values close to 0 indicate the lack of any spatial autocorrelation and, hence, spatial patterns.

In general, there are several methods for calculating the W_{ij} weight matrix that mostly rely on the contiguity and (Euclidean) distance between the observations (Fotheringham, Brunsdon, & Charlton, 2002; Goodchild, 1986; Rogerson, 2010). Here, the weighting scheme used to compute the W_{ij} matrix is based on a number of k-nearest neighbors, and the Euclidean distance between the centroids of each postcode area i and $j(k) \neq i$. Subsequently, the value of the W_{ij} matrix equals to 1, when $d_{ij} \leq H_k$ (where H_k corresponds to the k-nearest distance between i and its k-nearest neighbor $j(k)$), and 0, when $d_{ij} > H_k$. This scheme is chosen for its ability to provide a minimum amount of neighbors for each one of the 96 observations, and especially for the southeastern areal units of the Amsterdam city region, which are disjoint from the main corpus of the city. However, in order to assess the sensitivity of the I-values and their corresponding significance to different weightings of the examined variables, every calculation of the I index is tested for nine different variations of k-nearest neighbors. The different Z variables, for which the global Moran's I is computed, in addition to the obtained results for every k-nearest neighbor variation, are listed in Table 17.

In order to assess the statistical significance of the obtained global I-values against the null hypothesis of zero spatial autocorrelation, two statistical inference tests are carried out. In the first one, the aforementioned hypothesis is further specified to assume that the distribution of I approximately approaches a Gaussian (i.e. normal) distribution, as the number of observations increases. Given that the 96 areal units are considered to represent a sufficiently large number of observations (Cliff & Ord, 1973, 1981; Rogerson, 2010)[59], it can be assumed that each z_i is drawn from a Gaussian distribution of random Z variables, so that the sample distribution of I-values is approximately normal. This particular approach is the *resampling hypothesis* of I (Goodchild, 1986), which may be used to determine a test statistic with a score of significance $z_N(I)$ that is given by the following mathematical expression:

$$z_N(I) = \frac{I - E_N(I)}{\sqrt{V_N(I)}} \tag{6.5}$$

Where $E_N(I)$ resembles the expected mean of the sampling distribution of I, when the null hypothesis is true, and is given by $E_N(I) = -1/(n-1)$. $V_N(I)$ is the variance of the distribution and the formulas for calculating it are provided in Appendix E. Based on this score, the probability (p-value) of obtaining a sample I statistic that is equal to, or larger than, the observed I-values may also be retrieved.

In the second test, the initial null hypothesis is further specified to assume that any random permutation of z_i-values over the entire set of areal units in question is equally possible. Therefore, instead of presuming a Gaussian distribution of the z_is, each of the observed I indices may be assessed in relation to the set of all possible values that could be derived from the random permutation of all z_is over the geographical divisions (Fischer & Wang, 2011). This approach is the *randomization hypothesis* of I (Goodchild, 1986) and the corresponding $z_R(I)$ statistic is given by:

$$z_R(I) = \frac{I - E_R(I)}{\sqrt{V_R(I)}} \tag{6.6}$$

Where, as before, $E_R(I) = -1/(n-1)$ is the expected mean value of I, and $V_R(I)$, whose mathematical definitions are also provided in Appendix E, resembles the variance of I under the randomization hypothesis (Fotheringham et al., 2000).

59 Cliff and Ord suggest that the number of observations should be higher than 16, whereas Rogerson suggests that they should be above 20, in order for them to asymptotically approach a Gaussian distribution.

From the definition of this hypothesis, if n is the number of areal units, then the possible permutations of z_i-values would be $n!$ or, specifically in the examined case, 96! = 9,917e+149 permutations. As it is virtually impossible to calculate such a large number of potential I-values, a Monte Carlo simulation is employed instead, which generates a certain amount of random permutations (here, 10,000 permutations are carried out). For each one of the simulated permutations, a global Moran's I is computed and tested against the observed I-value drawn from the empirical data, in order to assess the extent to which it deviates from the null hypothesis. This further allows to extract the corresponding p-value, which indicates how likely it is to get an I-value that is equal to, or greater than, the observed one, in the case of complete spatial randomness (Hope, 1968). For both tests, if the $z_N(I)$ or the $z_R(I)$ values are greater than +1.96 or smaller than -1.96, with a p-value < 0.05 (that is, more than 95% confidence interval), the null hypothesis is rejected and the obtained Moran's I index is considered statistically significant. The computed results of $z_N(I)$ and $z_R(I)$, along with the corresponding p-values for each one of the examined variables are listed in Table 17.

TABLE 17 Global Moran's I values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	k	Global Moran's I	$z_N(I)$ resampling	p-value resampling	$z_R(I)$ randomization	p-value randomization
POI density	2	0.619	6.875	6.203e-12	6.857	7.036e-12
	3	0.588	7.907	2.627e-15	7.887	3.101e-15
	4	0.590	9.127	7.069e-20	9.103	8.813e-20
	5	0.578	10.047	9.514e-24	10.020	1.242e-23
	6	0.574	10.990	4.283e-28	10.961	5.889e-23
	9	0.480	11.525	9.815e-31	11.495	1.392e-30
	12	0.419	12.103	1.015e-33	12.071	1.490e-33
	18	0.307	11.619	3.281e-31	11.590	4.649e-31
	24	0.222	10.447	1.509e-25	10.421	1.988e-25
Residents' Activity (*Twitter* – Entire period)	2	0.245	2.797	0.005	4.466	7.984e-06
	3	0.188	2.620	0.009	4.185	2.850e-05
	4	0.161	2.603	0.009	4.158	3.210e-05
	5	0.151	2.754	0.006	4.398	1.090e-05
	6	0.146	2.936	0.003	4.689	2.751e-06
	9	0.110	2.836	0.005	4.522	6.147e-06
	12	0.092	2.880	0.004	4.577	4.701e-06
	18	0.069	2.919	0.004	4.590	4.436e-06
	24	0.050	2.706	0.007	4.188	2.818e-05

>>>

TABLE 17 Global Moran's *I* values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	*k*	Global Moran's *I*	$z_N(I)$ resampling	*p*-value resampling	$z_R(I)$ randomization	*p*-value randomization
Residents' Activity (*Twitter* – 13/11/2014–26/11/2014)	2	0.196	2.257	0.024	3.909	9.271e-05
	3	0.154	2.170	0.030	3.763	0.000e+00
	4	0.134	2.191	0.028	3.798	0.000e+00
	5	0.129	2.383	0.017	4.130	3.624e-05
	6	0.131	2.659	0.007	4.606	4.105e-06
	9	0.098	2.549	0.011	4.409	1.037e-05
	12	0.080	2.548	0.010	4.389	1.139e-05
	18	0.061	2.601	0.009	4.419	9.922e-06
	24	0.038	2.184	0.029	3.637	0.000e+00
Residents' Activity (*Twitter* – ALF Event)	2	0.279	3.163	0.002	4.474	7.676e-06
	3	0.211	2.923	0.003	4.137	3.522e-05
	4	0.178	2.861	0.004	4.049	5.151e-05
	5	0.165	2.998	0.002	4.243	2.209e-05
	6	0.157	3.140	0.002	4.443	8.881e-06
	9	0.114	2.932	0.003	4.144	3.407e-05
	12	0.095	2.981	0.003	4.206	2.602e-05
	18	0.070	2.957	0.003	4.143	3.427e-05
	24	0.052	2.806	0.005	3.891	9.992e-05
Residents' Activity (*Twitter* – 19/01/2015–31/01/2015)	2	0.173	2.006	0.045	3.837	0.000
	3	0.136	1.941	0.052	3.717	0.000
	4	0.123	2.034	0.042	3.895	9.823e-05
	5	0.117	2.170	0.030	4.153	3.284e-05
	6	0.114	2.331	0.020	4.462	8.132e-06
	9	0.093	2.441	0.015	4.660	3.180e-06
	12	0.079	2.521	0.012	4.787	1.696e-06
	18	0.061	2.627	0.009	4.899	9.637e-07
	24	0.045	2.478	0.013	4.501	6.750e-06

>>>

TOC

TABLE 17 Global Moran's *I* values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	*k*	Global Moran's *I*	$z_N(I)$ resampling	*p*-value resampling	$z_R(I)$ randomization	*p*-value randomization
Residents' Activity (*Twitter* – Entire period | 6am – 9am)	2	0.232	2.655	0.008	6.230	4.655e-10
	3	0.184	2.570	0.010	6.042	1.519e-09
	4	0.152	2.466	0.017	5.799	6.685e-09
	5	0.148	2.708	0.007	6.364	1.960e-10
	6	0.147	2.957	0.003	6.946	3.761e-12
	9	0.104	2.700	0.007	6.316	2.688e-10
	12	0.085	2.674	0.007	6.201	5.618e-10
	18	0.061	2.621	0.009	5.897	3.708e-09
	24	0.040	2.265	0.024	4.883	1.045e-06
Residents' Activity (*Twitter* – Entire period | 12pm – 15pm)	2	0.326	3.677	0.000e+00	5.519	3.409e-08
	3	0.261	3.596	0.000e+00	5.399	6.677e-08
	4	0.224	3.566	0.000e+00	5.354	8.587e-08
	5	0.208	3.730	0.000e+00	5.600	2.142e-08
	6	0.203	4.018	5.866e-05	6.032	1.622e-09
	9	0.155	3.888	0.000e+00	5.830	5.533e-09
	12	0.127	3.871	0.000e+00	5.789	7.063e-09
	18	0.094	3.832	0.000e+00	5.682	1.335e-08
	24	0.067	3.498	0.000e+00	5.120	3.052e-07
Residents' Activity (*Twitter* – Entire period | 18pm – 21pm)	2	0.224	2.562	0.010	3.423	0.001
	3	0.164	2.310	0.021	3.086	0.002
	4	0.140	2.285	0.022	3.053	0.002
	5	0.131	2.424	0.015	3.238	0.001
	6	0.129	2.613	0.009	3.490	0.000
	9	0.098	2.550	0.012	3.404	0.001
	12	0.082	2.610	0.009	3.479	0.000
	18	0.065	2.786	0.005	3.693	0.000
	24	0.047	2.585	0.001	3.399	0.001
Non-Residents' Activity (*Twitter* – Entire period)	2	0.668	7.420	1.167e-13	8.681	3.932e-18
	3	0.519	7.008	2.419e-12	8.199	2.418e-16
	4	0.437	6.800	1.048e-11	7.956	1.781e-15
	5	0.400	7.007	2.434e-12	8.198	2.446e-16
	6	0.437	8.403	4.358e-17	9.831	8.320e-23
	9	0.376	9.090	9.923e-20	10.631	2.147e-26
	12	0.286	8.334	7.813e-17	9.739	2.046e-22
	18	0.221	8.474	2.365e-17	9.878	5.194e-23
	24	0.169	8.074	6.797e-16	9.375	6.932e-21

>>>

TABLE 17 Global Moran's *I* values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	*k*	Global Moran's *I*	$z_N(I)$ resampling	*p*-value resampling	$z_R(I)$ randomization	*p*-value randomization
Non-Residents' Activity (*Twitter* – 13/11/2014–26/11/2014)	2	0.643	7.145	8.992e-13	8.160	3.361e-16
	3	0.501	6.771	1.281e-11	7.734	1.050e-14
	4	0.430	6.702	2.059e-11	7.654	1.945e-14
	5	0.370	6.497	8.219e-11	7.419	1.175e-13
	6	0.419	8.063	7.475e-16	9.208	3.332e-20
	9	0.356	8.627	6.325e-18	9.849	6.918e-23
	12	0.269	7.857	3.927e-15	8.965	3.095e-19
	18	0.216	8.277	1.267e-16	9.424	4.353e-21
	24	0.168	8.037	9.145e-16	9.122	7.354e-20
Non-Residents' Activity (*Twitter* – ALF Event)	2	0.645	7.165	7.804e-13	8.364	6.078e-17
	3	0.498	6.728	1.721e-11	7.855	3.993e-15
	4	0.408	6.358	2.045e-10	7.423	1.145e-13
	5	0.383	6.723	1.783e-11	7.849	4.201e-15
	6	0.412	7.938	2.047e-15	9.268	1.903e-20
	9	0.354	8.571	1.022e-17	10.003	1.474e-23
	12	0.266	7.770	7.830e-15	9.061	1.284e-19
	18	0.206	7.911	2.556e-15	9.202	3.506e-20
	24	0.158	7.577	3.529e-14	8.780	1.629e-18
Non-Residents' Activity (*Twitter* – 19/01/2015–31/01/2015)	2	0.661	7.339	2.146e-13	8.452	2.862e-17
	3	0.522	7.036	1.976e-12	8.104	5.321e-16
	4	0.461	7.168	7.634e-13	8.255	1.517e-16
	5	0.419	7.327	2.361e-13	8.438	3.225e-17
	6	0.456	8.765	1.860e-18	10.095	5.823e-24
	9	0.396	9.559	1.186e-21	11.006	3.583e-28
	12	0.315	9.156	5.378e-20	10.534	5.978e-26
	18	0.231	8.830	1.047e-18	10.136	3.814e-24
	24	0.170	8.131	4.252e-16	9.302	1.381e-20
Foreign Tourists' Activity (*Twitter* – Entire period)	2	0.724	8.024	1.025e-15	9.449	3.413e-21
	3	0.593	7.978	1.483e-15	9.397	5.612e-21
	4	0.492	7.650	2.012e-14	9.010	2.060e-19
	5	0.420	7.348	2.018e-13	8.654	4.982e-18
	6	0.433	8.327	8.288e-17	9.807	1.052e-22
	9	0.342	8.292	1.115e-16	9.762	1.639e-22
	12	0.265	7.753	8.978e-15	9.120	7.487e-20
	18	0.186	7.207	5.713e-13	8.455	2.782e-17
	24	0.132	6.393	1.623e-10	7.470	8.020e-14

>>>

TOC

TABLE 17 Global Moran's *I* values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	*k*	Global Moran's *I*	$z_N(I)$ resampling	*p*-value resampling	$z_R(I)$ randomization	*p*-value randomization
Foreign Tourists' Activity (*Twitter* – 13/11/2014–26/11/2014)	2	0.773	8.557	1.155e-17	10.119	4.537e-24
	3	0.643	8.646	5.343e-18	10.225	1.526e-24
	4	0.536	8.308	9.721e-17	9.826	8.699e-23
	5	0.460	8.025	1.018e-15	9.490	2.303e-21
	6	0.475	9.118	7.680e-20	10.782	4.167e-27
	9	0.375	9.069	1.200e-19	10.721	8.097e-27
	12	0.289	8.436	3.277e-17	9.965	2.169e-23
	18	0.196	7.570	3.718e-14	8.917	4.772e-19
	24	0.135	6.526	6.748e-11	7.655	1.933e-14
Foreign Tourists' Activity (*Twitter* – ALF Event)	2	0.721	7.995	1.293e-15	9.386	6.216e-21
	3	0.589	7.925	2.284e-15	9.305	1.340e-20
	4	0.490	7.612	2.694e-14	8.938	3.967e-19
	5	0.417	7.301	2.847e-13	8.573	1.011e-17
	6	0.430	8.282	1.216e-16	9.723	2.406e-22
	9	0.340	8.251	1.570e-16	9.683	3.532e-22
	12	0.264	7.722	1.139e-14	9.057	1.345e-19
	18	0.187	7.226	4.966e-13	8.452	2.861e-17
	24	0.133	6.438	1.209e-10	7.500	6.370e-14
Foreign Tourists' Activity (*Twitter* – 19/01/2015–31/01/2015)	2	0.661	7.341	2.116e-13	8.620	6.717e-18
	3	0.537	7.246	4.282e-13	8.509	1.748e-17
	4	0.442	6.875	6.203e-12	8.073	6.847e-16
	5	0.376	6.601	4.097e-11	7.751	9.134e-15
	6	0.383	7.397	1.393e-13	8.686	3.769e-18
	9	0.301	7.325	2.397e-13	8.598	8.143e-18
	12	0.232	6.823	8.935e-12	8.002	1.221e-15
	18	0.165	6.423	1.339e-10	7.513	5.772e-14
	24	0.119	5.805	6.418e-09	6.764	1.341e-11
Residents' Activity (*Instagram* – Entire period)	2	0.813	8.998	2.305e-19	9.154	5.507e-20
	3	0.769	10.315	6.053e-25	10.493	9.256e-26
	4	0.733	11.307	1.206e-29	11.503	1.267e-30
	5	0.709	12.275	1.237e-34	12.488	8.721e-36
	6	0.729	13.899	6.391e-44	14.140	2.144e-45
	9	0.640	15.302	7.413e-53	15.567	1.225e-54
	12	0.554	15.896	6.726e-57	16.170	8.192e-59
	18	0.456	17.083	1.994e-65	17.373	1.325e-67
	24	0.364	16.831	1.449e-63	17.110	1.244e-65

>>>

TABLE 17 Global Moran's *I* values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	*k*	Global Moran's *I*	$z_N(I)$ resampling	*p*-value resampling	$z_R(I)$ randomization	*p*-value randomization
Residents' Activity (*Instagram* – 13/11/2014–26/11/2014)	2	0.799	8.842	9.343e-19	8.979	2.708e-19
	3	0.765	10.257	1.096e-24	10.417	2.078e-25
	4	0.724	11.164	6.092e-29	11.338	8.522e-30
	5	0.701	12.140	6.491e-34	12.328	6.363e-35
	6	0.720	13.733	6.411e-43	13.947	3.299e-44
	9	0.629	15.031	4.551e-51	15.265	1.314e-52
	12	0.540	15.504	3.266e-54	15.743	7.648e-56
	18	0.447	16.733	7.539e-63	16.988	1.016e-64
	24	0.359	16.611	5.812e-62	16.858	9.170e-64
Residents' Activity (*Instagram* – ALF Event)	2	0.811	8.980	2.718e-19	9.142	6.156e-20
	3	0.765	10.260	1.062e-24	10.446	1.534e-25
	4	0.730	11.269	1.858e-29	11.473	1.806e-30
	5	0.707	12.244	1.801e-34	12.465	1.154e-35
	6	0.727	13.864	1.046e-43	14.114	3.105e-45
	9	0.641	15.305	7.058e-53	15.581	9.826e-55
	12	0.555	15.920	4.632e-57	16.205	4.640e-59
	18	0.456	17.059	3.018e-65	17.360	1.655e-67
	24	0.363	16.773	3.856e-63	17.063	2.815e-65
Residents' Activity (*Instagram* – 19/01/2015–31/01/2015)	2	0.832	9.204	3.448e-20	9.420	4.508e-21
	3	0.767	10.284	8.318e-25	10.526	6.586e-26
	4	0.724	11.165	6.045e-29	11.427	3.055e-30
	5	0.685	11.879	1.518e-32	12.158	5.192e-34
	6	0.707	13.484	1.934e-41	13.810	2.517e-43
	9	0.610	14.579	3.811e-48	14.921	2.414e-50
	12	0.533	15.294	8.383e-53	15.651	3.277e-55
	18	0.451	16.889	5.420e-64	17.277	6.969e-67
	24	0.360	16.667	2.296e-62	17.041	4.073e-65
Non-Residents' Activity (*Instagram* – Entire period)	2	0.753	8.348	6.924e-17	9.277	1.746e-20
	3	0.632	8.492	2.035e-17	9.437	3.832e-21
	4	0.551	8.539	1.352e-17	9.490	2.317e-21
	5	0.502	8.753	2.083e-18	9.727	2.317e-22
	6	0.551	10.542	5.542e-26	11.715	1.071e-31
	9	0.473	11.369	5.951e-30	12.632	1.415e-36
	12	0.379	10.960	5.931e-28	12.171	4.416e-34
	18	0.289	10.965	5.659e-28	12.156	5.328e-34
	24	0.223	10.474	1.136e-25	11.583	5.000e-31

>>>

TOC

TABLE 17 Global Moran's *I* values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	*k*	Global Moran's *I*	$z_N(I)$ resampling	*p*-value resampling	$z_R(I)$ randomization	*p*-value randomization
Non-Residents' Activity (*Instagram* – 13/11/2014–26/11/2014)	2	0.733	8.123	4.537e-16	8.697	3.398e-18
	3	0.641	8.614	7.051e-18	9.223	2.880e-20
	4	0.563	8.717	2.848e-18	9.334	1.021e-20
	5	0.511	8.910	5.106e-19	9.540	1.429e-21
	6	0.553	10.593	3.217e-26	11.342	8.159e-30
	9	0.477	11.469	1.888e-30	12.278	1.187e-34
	12	0.384	11.104	1.196e-28	11.884	1.432e-32
	18	0.316	11.963	5.579e-33	12.790	1.875e-37
	24	0.253	11.842	2.374e-32	12.641	1.260e-36
Non-Residents' Activity (*Instagram* – ALF Event)	2	0.744	8.246	1.641e-16	9.202	3.514e-20
	3	0.617	8.303	1.019e-16	9.266	1.929e-20
	4	0.539	8.359	6.329e-17	9.329	1.068e-20
	5	0.491	8.568	1.049e-17	9.563	1.147e-21
	6	0.541	10.370	3.391e-25	11.573	5.639e-31
	9	0.464	11.148	7.308e-29	12.439	1.612e-35
	12	0.371	10.731	7.318e-27	11.967	5.309e-33
	18	0.278	10.570	4.108e-26	11.767	5.749e-32
	24	0.212	10.012	1.348e-23	11.117	1.033e-28
Non-Residents' Activity (*Instagram* – 19/01/2015–31/01/2015)	2	0.693	7.692	1.443e-14	9.061	1.297e-19
	3	0.577	7.764	8.240e-15	9.146	5.904e-20
	4	0.489	7.589	3.226e-14	8.940	3.893e-19
	5	0.444	7.763	8.271e-15	9.145	5.956e-20
	6	0.485	9.318	1.182e-20	10.976	4.967e-28
	9	0.421	10.152	3.262e-24	11.954	6.220e-33
	12	0.344	9.991	1.668e-23	11.755	6.631e-32
	18	0.258	9.810	1.015e-22	11.512	1.154e-30
	24	0.190	9.001	2.239e-19	10.518	7.102e-26
Foreign Tourists' Activity (*Instagram* – Entire period)	2	0.732	8.113	4.924e-16	9.707	2.813e-22
	3	0.604	8.129	4.325e-16	9.728	2.300e-22
	4	0.502	7.797	6.329e-15	9.330	1.056e-20
	5	0.433	7.563	3.928e-14	9.050	1.428e-19
	6	0.448	8.620	6.696e-18	10.314	6.096e-25
	9	0.358	8.674	4.149e-18	10.375	3.218e-25
	12	0.277	8.084	6.274e-16	9.660	4.450e-22
	18	0.193	7.455	8.975e-14	8.882	6.563e-19
	24	0.133	6.453	1.096e-10	7.653	1.960e-14

>>>

TABLE 17 Global Moran's *I* values of the examined variables, along with the scores of statistical significance for both the resampling and the randomization hypotheses.

Z variable	*k*	Global Moran's *I*	$z_N(I)$ resampling	*p*-value resampling	$z_R(I)$ randomization	*p*-value randomization
Foreign Tourists' Activity (*Instagram* – 13/11/2014–26/11/2014)	2	0.813	8.999	2.271e-19	10.127	4.186e-24
	3	0.688	9.233	2.634e-20	10.391	2.723e-25
	4	0.596	9.229	2.730e-20	10.387	2.849e-25
	5	0.537	9.351	8.651e-21	10.524	6.687e-26
	6	0.564	10.791	3.792e-27	12.144	6.174e-34
	9	0.472	11.355	6.999e-30	12.776	2.245e-37
	12	0.381	11.007	3.529e-28	12.377	3.461e-35
	18	0.276	10.487	9.944e-26	11.770	5.559e-32
	24	0.196	9.300	1.403e-20	10.409	2.258e-25
Foreign Tourists' Activity (*Instagram* – ALF Event)	2	0.728	8.075	6.751e-16	9.669	4.072e-22
	3	0.601	8.091	5.936e-16	9.690	3.334e-22
	4	0.499	7.749	9.289e-15	9.280	1.6934e-20
	5	0.429	7.505	6.140e-14	8.988	2.518e-19
	6	0.444	8.549	1.244e-17	10.237	1.348e-24
	9	0.354	8.580	9.491e-18	10.271	9.555e-25
	12	0.273	7.975	1.516e-15	9.539	1.443e-21
	18	0.190	7.345	2.057e-13	8.758	1.986e-18
	24	0.131	6.353	2.105e-10	7.541	4.660e-14
Foreign Tourists' Activity (*Instagram* – 19/01/2015–31/01/2015)	2	0.692	7.681	1.575e-14	9.675	3.867e-22
	3	0.561	7.553	4.247e-14	9.515	1.810e-21
	4	0.458	7.130	1.005e-12	8.982	2.659e-19
	5	0.387	6.778	1.219e-11	8.538	1.363e-17
	6	0.395	7.628	2.378e-14	9.609	7.340e-22
	9	0.314	7.632	2.318e-14	9.608	7.415e-22
	12	0.239	7.037	1.968e-12	8.848	8.882e-19
	18	0.164	6.404	1.518e-10	8.019	1.064e-15
	24	0.111	5.481	4.234e-08	6.821	9.010e-12

§ 6.5.3.2 Results of global spatial autocorrelation analysis

As mentioned previously, the first variable that is examined is the density of POI locations. The POI dataset comprises 12,198 locations of "third places" (Rosenbaum, 2006), such as restaurants, museums, parks, nightlife spots, cafés, cinemas, transportation hubs, and other related facilities, extracted from Foursquare.

Professional places ("second places") and residences ("first places") are excluded from the collected dataset, so as to resemble as much as possible the places where people are most likely to socialize. Subsequently, the POIs are aggregated into the various spatial divisions (i.e. postcode areas), and normalized by the total area of the latter, to obtain the POI density of each areal unit.

The spatial autocorrelation of the POI density levels aims to firstly identify spatial patterns of neighboring districts with high or low concentration of POI locations and, secondly, examine whether or not there is a correspondence between POI density and human social activity. The global Moran's I of POI density for various k-nearest neighbors (see Table 17) indicates a strong positive spatial autocorrelation. Moreover, there is sufficient evidence that the obtained I-values are statistically significant, under both the resampling and the randomization hypothesis, since the respective $z_N(I)$ and $z_R(I)$ scores are rather high (around +10.00 in all tested variations), with a greater than 99.99% confidence interval (p-values << 0.001). In addition, the observed global I index appears relatively stable, especially within the range of 2 to 9 nearest neighbors. The aforementioned results indicate the existence of spatial clusters of districts with similarly high or low values of physical POI density. Therefore, the null hypothesis of complete spatial randomness with regard to POI density values is rejected. To specifically identify which of the districts lie in these clusters, LISA statistics are used, as discussed later in the section and shown in the corresponding cluster maps (see Appendix F).

Although the above measures identify spatial patterns of areal units where physical POI locations are either densely or sparsely distributed, they do not reveal anything about the intensity of activity that takes place in these areas. To quantify spatial and temporal patterns of high or low social activity intensity, especially around the time period in focus, the global Moran's I statistic is calculated for 27 additional variables, listed in Table 17. The collected data from Twitter and Instagram are used as proxies of social activity, characterizing different groups of people at various time intervals. The *SocialGlass* system enables the identification of three distinct social categories in the dataset, namely residents, non-residents, and foreign tourists, whose activity patterns are studied here. In general, the variables are classified into three main sets, each one resembling the activity of a particular social category. Then, each set is divided into a subset of variables that represent four different time periods. These respectively refer to the entire monitoring period (from November 13, 2014 till January 31, 2015), the weeks before the ALF event (November 13, 2014 – November 26, 2014), the ALF event period (November 27, 2014 – January 18, 2015), and the weeks after the event (January 18, 2015 – January 31, 2015). Each of the aforementioned variables corresponds to social activity intensity, as inferred from either Twitter or Instagram. There is also an additional set of three variables that reflects the aggregated activity of residents at different time intervals within a day – respectively from 6am to 9am, from 12pm to 15pm, and from 18pm to 21pm – over the entire monitoring period, as inferred from

Twitter. As with the POI locations, the collected geo-referenced posts are aggregated into the postcode areas, and normalized by the total area of each spatial unit, in order to provide a social activity intensity rate for each geographic division in focus.

The obtained global Moran's I values that correspond to the intensity of social activity of residents on Twitter indicate weak positive spatial autocorrelation, accompanied by relatively low values of statistical significance, regardless the time period. These results suggest that residents do not cluster significantly in particular regions of the city when they socialize, but instead their activity is quite dispersed or random over the entire city region. This was also evident in the visual exploratory analysis of activity heatmaps, discussed in Sect. 6.5.2. On the contrary, the I-values of their activity on Instagram signify very strong positive spatial autocorrelation (e.g. values larger than +0.7 for k=6 nearest neighbors), which additionally yield high z-scores and confidence levels. This particular contradiction may derive from the different nature of the social media platforms in question. People tend to use each platform differently, to express, communicate, and share certain aspects of their daily lives. Apparently, this variance in the usage is reflected on the data values too. A semantic analysis on the corresponding posts would presumably yield insight into the nature of this discrepancy, yet it has not been carried out in the present study. However, what is particularly interesting, is that, regardless the data source, the values of the I statistic remain relatively stable – according to the sensitivity analysis using different weights – irrespective of the time period. This signifies that the activity behavior of the residents is not influenced by the event, which is in contrast to the main hypothesis. This would be tested further at the district level with the use of local indicators.

As regards the non-residents, the observed I-values suggest strong deviation from the random distribution. In other words, this particular social category appears to systematically form spatial patterns of activity behavior, reflected in both Twitter and Instagram data. In fact, the spatial autocorrelation of their activity yields almost equal positive values of high statistical significance, irrespective of the time period and data source. This also contradicts the initial hypothesis that the event would cause an increase in the intensity of activity. Moreover, it further contradicts the assumption that the event would have an effect on the activity behavior of all social categories. Already at this stage, substantial differences are detected between the social categories, in terms of activity behavior over space and time, yet these appear to be independent of the event.

Lastly, the activity of foreign tourists also appears to be organized around clusters of either high or low values of intensity rates. Strong positive autocorrelation is detected in both Twitter and Instagram data. Although, in general, the I-values remain relatively stable over the different time periods, it appears that the detected clusters are stronger prior to the event, with a tendency to become slightly weaker in the two remaining time periods that are studied here.

In total, the obtained results from the global autocorrelation analysis have shown that the social categories of non-residents and foreign tourists have a tendency to agglomerate around certain districts of the city. Unlike the group of residents, these two groups are more likely to collocate over space and time. For the moment, the event has almost negligible effects on the activity behavior of the three social categories, which could indicate that each group is characterized by quite regular activity patterns over time. However, the measures that have been carried out thus far reflect the entire city region. To further identify which specific areal units contribute strongly to the spatial clusters or outliers detected at the global level, a local autocorrelation analysis is performed and discussed in the following sections.

§ 6.5.3.3 Local spatial association statistics and tests

The local autocorrelation analysis is based on indicators, in particular, *local indicators of spatial association* (LISAs) (Anselin, 1995) that spatially decompose the global ones. This spatial decomposition enables the identification of localized phenomena (clusters or outliers) in each areal unit under consideration that further allow to assess which particular observations contribute most strongly to the patterns detected in the overall results. Instead of providing a general indication about whether autocorrelation occurs, LISAs assign to each data value of an areal unit a local measure that signifies where there is strong – or weak – clustering of similar values around that spatial unit. In other words, they allow to correlate the values of a chosen variable at each geographical division or location with the average value of the same variable at certain k-nearest neighboring locations. The results obtained from the local autocorrelation analysis give evidence of potential local clusters of either high (else called *hot spots*) or low values (or *cold spots*). In addition, they give a quantifiable indication of the extent to which any association occurs.

In the context of this study, LISAs are used to, first, identify which particular spatial units contribute most strongly to the clustering of POI-dense districts, detected in the global analysis, and to assess the extent and spatial distribution of these patterns. Second, they are applied to the various social activity variables to detect the corresponding spatial footprints of human activity patterns that characterize each social category over time. This would eventually allow to assess whether or not there is a relationship between dense POI locations and intense social activity rates. It would further enable to understand the spatial impact of the event on the social activity patterns of people.

To this end, the local version of the Moran's I index, in combination with the Getis-Ord G_i^* statistic are employed. The local Moran I_i statistic is given by the following expression:

$$I_i = \frac{z_i - \bar{z}}{1/n \sum_{i=1}^{n} (z_i - \bar{z})^2} \sum_{j=1}^{k} W_{ij}(z_j - \bar{z}) \ , \qquad j \neq i \tag{6.7}$$

The notation is similar to the global I index, given in Eq. (6.4). The weight matrix W_{ij} is calculated in the same way as for the global Moran's I, based on the Euclidean distance between an areal unit i and the k-nearest j neighbor. Although the interpretation of the local I_i statistic is similar to the global one, the range of values of the former does not lie within the [-1, +1] range of the global index. Accordingly, the significance of the obtained I_i-values can be assessed against the corresponding $z(I_i)$-scores of a random permutation test, under a null hypothesis of no spatial association, given by the following (Anselin, 1995):

$$z_R(I_i) = \frac{I_i - E_R(I_i)}{\sqrt{V_R(I_i)}} \tag{6.8}$$

The mathematical definitions of the expected mean $E_R(I_i)$ and the variance $V_R(I_i)$ are provided in Appendix E. With regard to the corresponding p-values, obtained from the above test, and in order to mitigate the risk of acquiring "false-positive results" (i.e. significant I_i-values in areas where the null hypothesis of no spatial autocorrelation is in fact true) they are adjusted according to the False Discovery Rate (FDR) test, introduced by (Benjamini & Hochberg, 1995). The test results are illustrated in the FDR choropleths, included in Appendix F.

In addition to the local index, the Getis-Ord G_i^* statistic (Getis & Ord, 1992; Ord & Getis, 1995) is calculated for each -value of the set of Z variables under consideration (see Table 17). Unlike the local Moran's I_i which indicates the existence of spatial clusters characterized by similar values, the Getis-Ord statistic can further detect local spatial units that comprise either high or low values of the variable in question. Here, the calculation of the local G_i^* index employs the updated definition of the statistic, described in (Ord & Getis, 1995), and given by:

$$G_i^*(d) = \frac{\sum_{j=1}^{n} W_{ij}(d) z_j - \bar{z} \sum_{j=1}^{n} W_{ij}(d)}{s \sqrt{\frac{n \sum_{j=1}^{n} W_{ij}^2(d) - \left[\sum_{j=1}^{n} W_{ij}(d)\right]^2}{n-1}}} \tag{6.9}$$

Where z_j resembles the value of the variable Z at a location (or spatial unit) j, which is a neighboring areal unit to i, $\bar{z}$ is the sample mean of Z, s is its standard deviation, and

$W_{ij}(d)$ represents the weight matrix, specifying the degree of dependency between the areal units i and j within a distance d. If $d_{ij} \leq d$ then $W_{ij}(d) = 1$, otherwise $W_{ij}(d) = 0$, where d_{ij} is the distance between the centroids of the spatial units i and j. The combination of both the local Moran's I_i and the Getis-Ord G_i^* allows the identification of spatially homogeneous or heterogeneous patterns of human activity, that are further characterized by agglomerations of significant positive or negative G_i^*-values.

§ 6.5.3.4 Findings of local spatial association analysis – Identifying local spatial clusters of social activity over time

The global autocorrelation analysis of the POI density over the various spatial units of Amsterdam has revealed that there exist significant clusters of neighboring districts, where POI locations tend to agglomerate or be dispersed (strong positive overall autocorrelation). Yet, it has not identified which are the particular spatial units that largely contribute to the overall outcome. For this, the local autocorrelation analysis results, in relation to POI density, suggest a prominent pattern of high I_i z-scores (high with high values – HH) at the central and southern parts of the city, forming a single cluster of districts[60] (see Appendix F, Figure 58). On the contrary, areas at the north-western (Waterland), north (Kadoelen), north-eastern (Weestelijk Havengebied, Spieringerhorn, De Eendracht), and south-eastern outskirts (Bijlmer Centrum, Bullewijk, Holendrecht etc.) form clusters of low I_i-values (low with low values – LL), yet with significant z-scores. The spatially homogeneous pattern at the central and southern parts signifies that areas of high POI density tend to concentrate into a single region of the city. The dominance of this region is further confirmed by the corresponding G_i^* statistic. According to the Getis-Ord cluster map (see Appendix F, Figure 58), the HH I_i units also yield significant G_i^*-values, with more than 99% confidence interval (p-values < 0.01, using a Monte Carlo simulation with 10,000 permutations). This means that areas with high POI density in the city of Amsterdam are closer together than areas with low POI density. What remains, is to examine whether or not these clusters are related to patterns of areas with high social activity intensity. Also, to examine whether or not different patterns of spatial association, in terms of social activity intensity, appear before, during, and after the ALF event.

[60] The districts that comprise the cluster are namely: Hoofddorppleinbuurt, Westlandgracht, Westindische buurt, Willemspark, Overtoomse Sluis, Schinkelbuurt, Apollobuurt, Scheldebuurt, Nieuwe Pijp, Oude Pijp, Duivelseiland, IJsselbuurt, Weteringschans, Weesperzijde, Transvaalbuurt, and Frankenbuurt.

The Moran scatterplot of residents' activity (see Appendix F, Figure 59)[61] as inferred from Twitter for the entire period in question, generally suggests a positive autocorrelation. More specifically, areas where local I_i z-scores are below average, tend to cluster with neighboring districts that have equally low average values (lower-left quadrant of the scatterplot). There are only a few high activity areas that cluster with other high activity neighboring districts (upper-right quadrant). These districts are the ones that contribute strongly to the global result, which indeed indicated a weak positive spatial autocorrelation. Looking at the resulting cluster map, it appears that the social activity of residents clusters around the central train station, at the northern part of Amsterdam (Figure 30). Conversely, the corresponding activity inferred from Instagram generates a wider and dominant cluster of areas with high z-scores, covering the central districts of the city (Figure 31). The different time periods examined here, show minor variations in the pattern formation for both Twitter and Instagram activity. Moreover, the local autocorrelation analysis of the aggregate daily activity in different time frames (i.e. 6am – 9am; 12pm – 15pm; 18pm – 21pm) also indicates minor fluctuations in the structure of clusters (see Appendix F, Figure 67). Yet, a variance is detected in terms of the significance of the obtained values, as observed in the respective scatterplots. Especially regarding the intensity of activity from 6am to 9am, the resulting z-scores accumulate either in HH or LL regions. This implies the existence of certain dominant activity hubs (HH regions, around the central station), where residents tend to co-locate, resulting in a contrasting pattern between highly vibrant and rather idle neighborhoods.

Although the neighborhoods belonging to the HH I_i z-score regions – which are also characterized by significant G_i^* values (see Appendix F, Figures 65 – 66) – coincide with the areas where ALF installations are mostly gathered, it appears that the activity clusters are not so much related to the event, as almost the same patterns occur in the periods before and after ALF. What is interesting, though, is that the high activity hubs barely relate to those characterized by high POI density. This could imply – at least for the case of Amsterdam – that a high number of POI venues in a neighborhood is not necessarily linked to equally intense social activity.

As regards the activity of both non-residents and foreign tourists, the local autocorrelation outcomes suggest almost identical patterns of either HH or LL regions (Figures 30 – 31). Unlike residents, whose activity appears more dispersed, non-residents and foreign tourists tend to strongly co-locate around the city center. For both social categories, the neighborhoods that contribute significantly to the strong positive autocorrelation, observed in the global analysis results, gather around the central station and cover the neighborhood comprising Amsterdam's city center.

61 The entire set of Moran scatterplots, choropleths of local Moran's I_i-values, and the corresponding cluster maps of Moran's I_i and Getis-Ord G_i^*-statistics are included in Appendix F.

Although, in absolute numbers, the intensity of activity (i.e. the number of posts per POI) during the event is higher than the period before or after ALF – especially for tourists – it appears that it always tends to concentrate around the aforementioned districts, throughout the three-month period. This spatially homogeneous pattern of social activity, observed in the local autocorrelation results, is further strengthened by the high G_i^* values, characterizing the areas within the aforementioned cluster (see Appendix F, Figures 59 – 66).

Overall, both global and local autocorrelation analyses suggest non-random distribution of the activity behavior of the three social categories in question. This further indicates that the spatial manifestation of social activity, in the examined case, is characterized by a tendency to concentrate around specific areas of the city. However, the distribution of social activity over space and time demonstrates different patterns, according to the group of people it relates to. Residents' activity behavior appears more dispersed than the one of non-residents and foreign tourists. This was also observed initially in the visual exploratory analysis of spatiotemporal activity patterns (see Sect. 6.5.2), provided by the system's frontend. Yet, the spatial autocorrelation analysis further allowed to quantify and assess the extent of these clusters for each social category.

With regard to the main hypothesis, where it was assumed that the event would influence the overall activity behavior of all social categories especially in the areas around the event's installations, the analysis gives evidence to reject it. Although the volume of activity increases in absolute numbers for some social categories (i.e. foreign tourists and non-residents) during the event, the spatial footprint of activity behavior of each group presents only slight differences before, during, and after the event. In the city of Amsterdam, regardless the social category, areas characterized by high activity tend to form spatially homogeneous clusters, which – in this case – are observed around the central train station. In turn, the geographic extent and strength of these clusters depends on the social category in focus, and gives and indication of the areas where people belonging to different groups are more likely to co-locate. In either way, intense social activity tends to agglomerate around approximately 8-10 districts surrounding the central train station. This spatially concentrated distribution of social activity around a single region, is an indication of a dominant monocentric urban structure, both functionally and morphologically.

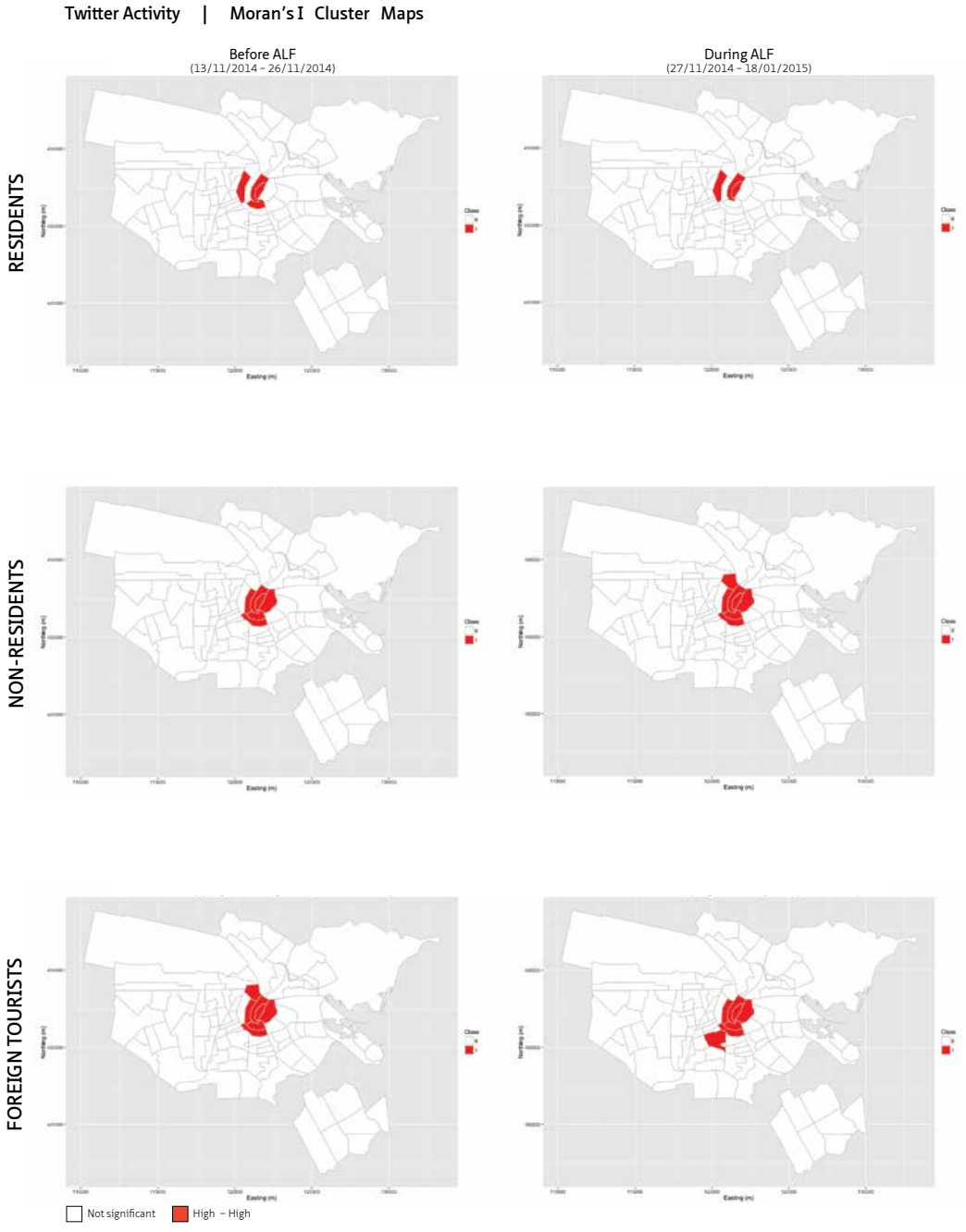
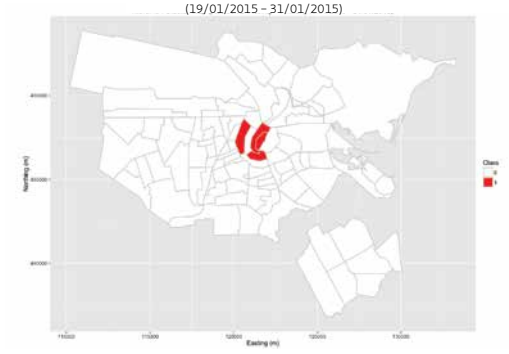
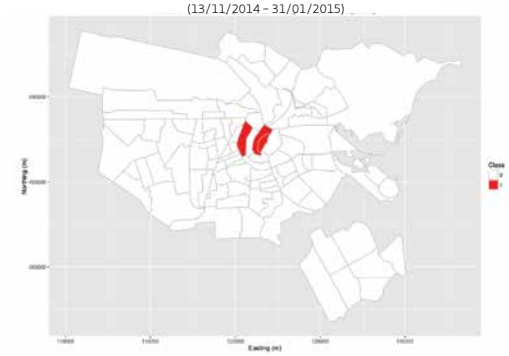


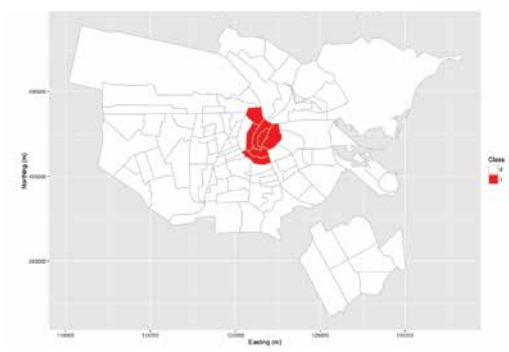
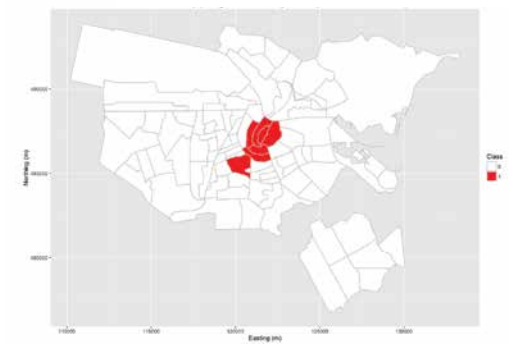
FIGURE 30 Local Moran's *I* cluster maps of social activity, referring to different social categories of people during different time periods, as inferred from Twitter. Red-colored districts indicate clusters of neighboring areas with high values of social activity.

TOC

After ALF
(19/01/2015 – 31/01/2015)

Entire Period
(13/11/2014 – 31/01/2015)

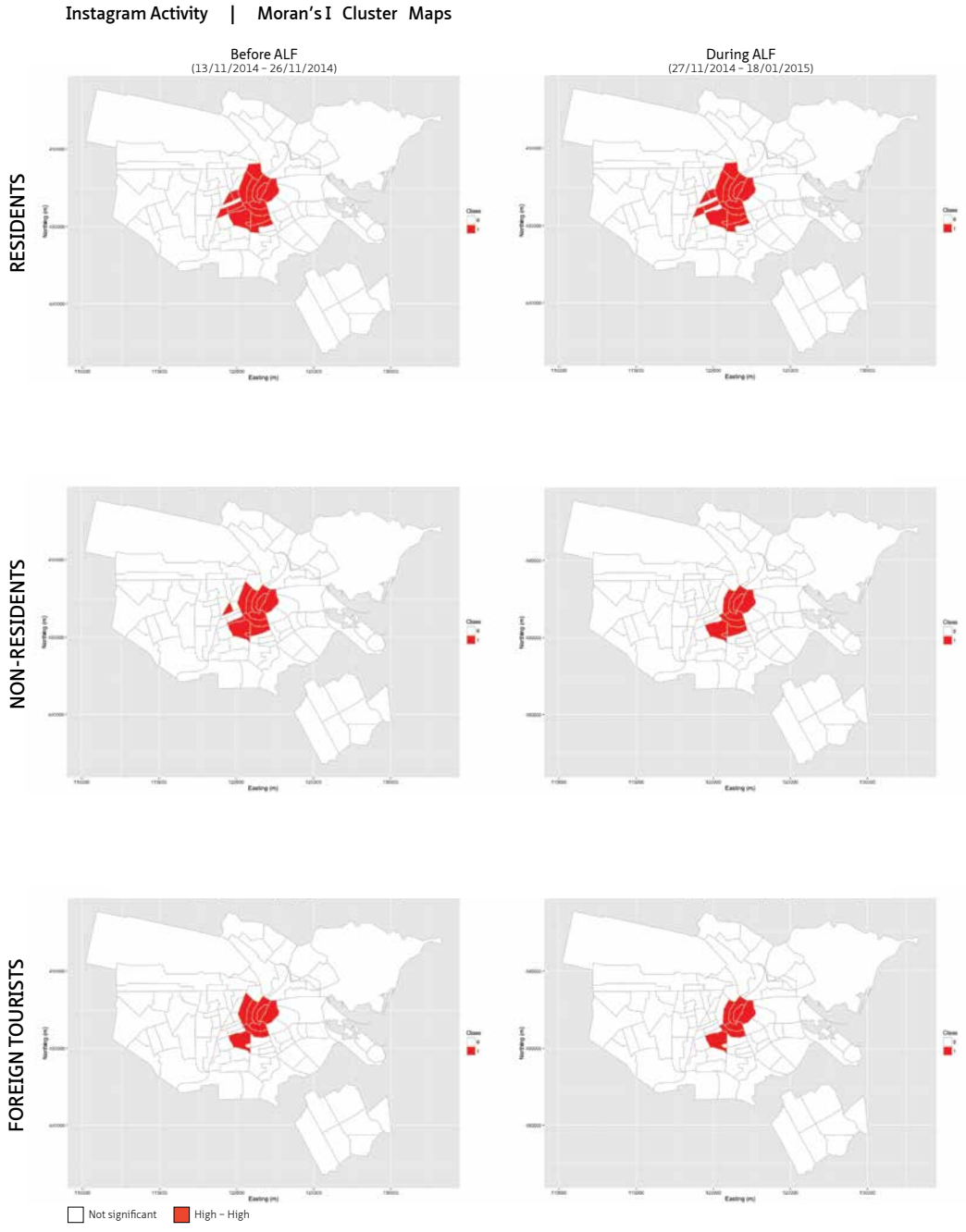
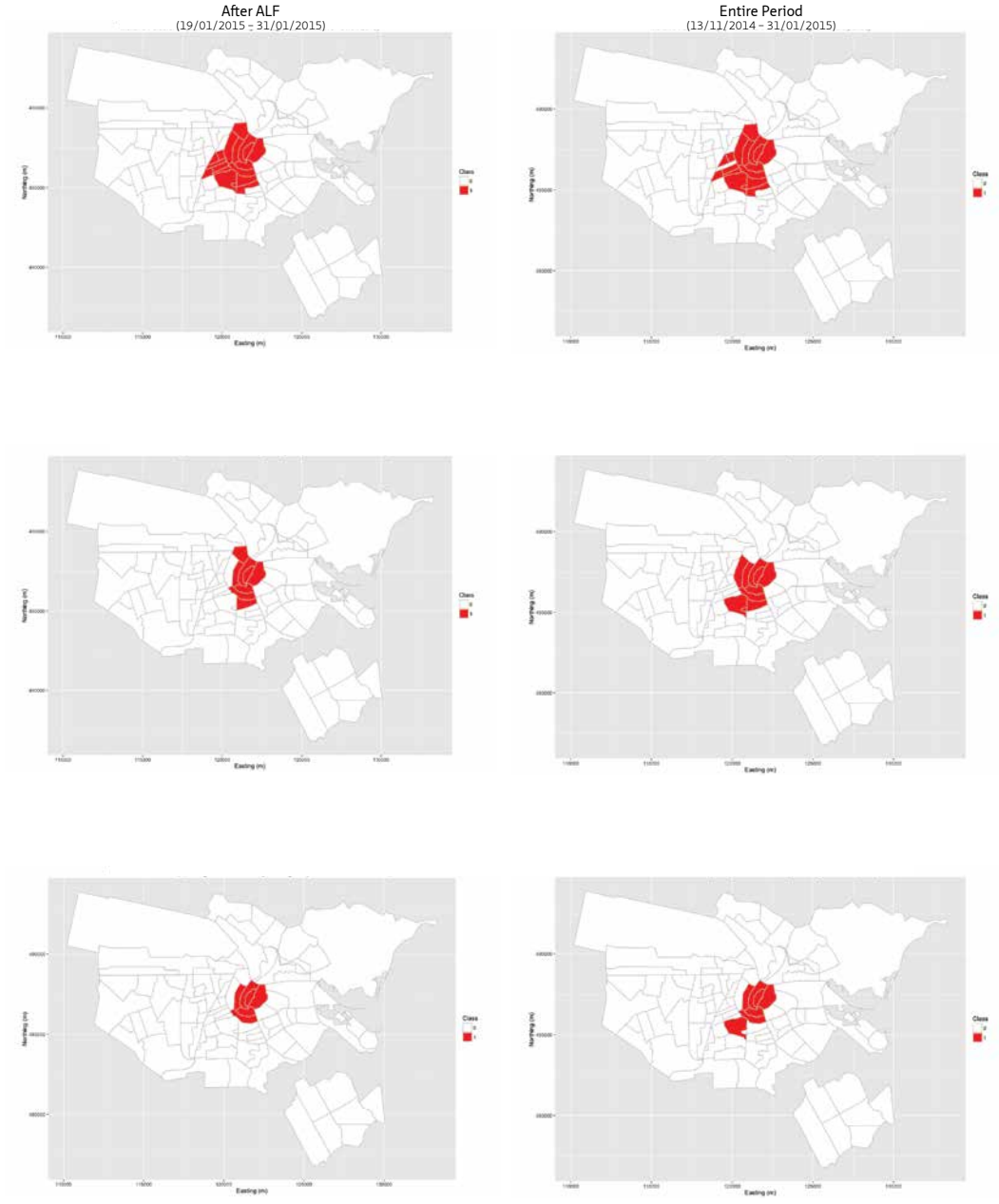



FIGURE 31 Local Moran's *I* cluster maps of social activity, referring to different social categories of people during different time periods, as inferred from Instagram.


After ALF
(19/01/2015 – 31/01/2015)
Entire Period
(13/11/2014 – 31/01/2015)

§ 6.6 Discussion and Conclusions

This Chapter introduced a web-based system (*SocialGlass*) that enables the visualization and exploratory analysis of dynamic urban phenomena related to social activity and human movement, by integrating heterogeneous data sources. It presented the features and functionality of the various components and modules, comprising the overall system architecture, and discussed the strategy employed to infer attributes of social agents and their activity behavior in cities from the datasets it incorporates. To demonstrate how these can be put to use in understanding the spatial distribution of social activity over time, a select number of the tools were implemented in a real-world case study. The latter involved the investigation of the potential influence a city-scale event may have on the everyday activity and mobility patterns of different social categories. Along with these exploratory goals, the case study is also used as a means to assess the capacities and limitations of the implemented system. The following reflections concern both the system and the study findings, and highlight the potential flaws of the chosen data sources, methods, approximations, and software development strategies.

With regard to the data sources, *SocialGlass* primarily operates with data from LBSNs (e.g. Foursquare) and geo-enabled online social networks (e.g. Twitter, Instagram, Sina Weibo), as proxies for social activity and human movement. Although it incorporates other sources such as census and socio-economic records, and also provides tools that allow custom (spatial) data to be uploaded to the system, the majority of built-in modules engage with aspects related to data from social media. On the one hand, that is because these particular data sources have only recently been used in urban analytics; much less than authoritative census data and CDRs. On the other hand, they are the most challenging sources in terms of ambiguity, semantic uncertainty, lack of structure, as well as limited and often unbalanced representativeness (see also Chapter 2). Many of the attributes discussed in this Chapter (e.g. home location, age range, gender, ethnicity, sentiments etc.) are not readily extractable from the raw data. This explains why there is an entire pipeline between data collection and availability in the frontend, with all the aforementioned attributes already approximated and provided to the end-user. At the same time, though, the entire range of interim processes is prone to the aforementioned limitations of human-generated data.

One of these limitations pertains to the representativeness of the sample, in terms of both individuals and POIs. Although the system identifies the amount of individual users that correspond to the total number of collected posts, and classifies them according to their home location, age range, gender, and ethnicity, it is difficult to cross-validate these attributes against a more reliable source, such as census data. The reason is that the attributes are in fact approximated and, therefore, no concrete measures can be made as to how representative they are of the population they refer

to. Similarly, POIs and their corresponding categories are extracted from Foursquare and, as such, only resemble a sample of the actual POI locations in a given city. Publicly-available data of functions at the level of individual buildings are generally scarce. An alternative source to correlate with, that is widely used in urban studies, is OpenStreetMap. Yet again, OSM is a VGI source that is also characterized by a degree of bias (Goodchild & Li, 2012). Presumably, the larger the penetration rate of social media technologies, the larger the amount of users and the corresponding sample, but how representative it is, is an issue to be studied further.

Another related issue is the untrustworthiness that generally characterizes human-generated data, which may affect the conclusions one could reach. One way to mitigate this problem is by employing crowd-sourcing or human computation techniques to carry out on-demand service requests to social media users for data cleansing and linkage. This is one of the main directions for future improvement of the system. In addition, ontology-based data integration methods (see Chapter 3) and tools (see Chapter 4) are currently being implemented in the semantic enrichment and integration component. In this way, a larger amount of (semantic) links could be established between the various datasets that are collected by or uploaded to the system. There is also room for improvement as regards the development of modules that integrate the potential interpersonal relationships between the users identified in the collected datasets (as derived from their online contacts) into the map-based frontend, as suggested by (Andris, 2016). This addition could largely benefit the understanding of human activity behavior over space and time.

The empirical validation of an instance of *SocialGlass* in the case study offers further insights into both the functionality of the system and the dynamics of social activity in cities. In particular, the datasets that were used in the study, in combination with the attributes that were extracted by the system's modules, enabled the exploration of the spatial and temporal organization of social activities over much shorter time scales (e.g. daily, hourly etc.) than it is possible with traditional urban data. Moreover, they allowed for disaggregation into different social categories, providing information that is hardly available in official records. The results of both the visual exploratory analysis using the system's frontend and the spatial autocorrelation analysis suggest that different groups of people make different use of urban space over time. Therefore, in studying the behavior and distribution of human activity, using aggregate populations that ignore internal diversities may largely affect the observations that could be made. Also, the data collection period plays a crucial role. The example examined here indicates that large-scale public events may influence the intensity of social activity and, presumably, the routine of human movement. The reflection of such fluctuations in the collected datasets could cause several anomalies that need to be taken into account in similar studies. Discrepancies were also detected between the data sources used. To a certain degree and for some social categories (e.g. the residents), Twitter data reflect different values of activity intensity from those inferred from Instagram. The ability of the system

to integrate heterogeneous sources and to also accommodate their diversity (e.g. identification of different social categories instead of considering aggregate uniform populations, use of different social media platforms etc.) proves beneficial in this regard. It also helps mitigate biases in the interpretation of the obtained results. In addition, the multiple types of visualization and data filters provide many opportunities to examine several variables in parallel.

What is still lacking, though, are modules for data analysis. As demonstrated in the case study, only the visual exploratory analysis is carried out using the system's frontend. To perform additional statistical analysis on the findings, external platforms are employed (here, the R statistical language). Moreover, the system does not introduce new social or spatial statistics, but does implement select metrics and methods for attribute extraction, described in Chapter 5, so as to infer attributes of individuals and their activity behavior.

In studying the dynamics of human activity in cities *SocialGlass* can be employed to create new experiments, define the data collection period and the area of study, invoke data crawling from different social media, upload custom datasets, extract related attributes, visualize and filter the collected datasets, perform interactive visual exploratory analysis, and export the results for integration into other specialized tools for further analysis. Therefore, *SocialGlass* can be used in combination with other software platforms to gain detailed insights into urban dynamics.

7 Discussion and Conclusions

§ 7.1 Introduction

The study of dynamic social and spatial phenomena in cities has evolved rapidly in the recent years, yielding new insights into the notion of "urban dynamics", as being first introduced by (Forrester, 1969). Although the term initially focused on the growth and economic interactions in urban systems, it currently also encompasses human mobility, flows of individuals and goods, and the distribution of social activity over space and time, among others. This evolution is strongly related to the concurrent emergence of data sources, which have potential to capture dimensions of social and geographic systems that are difficult to detect in traditional urban data (e.g. census data). The majority of datasets that are generated from these new sources (e.g. sensors, mobile phones, social media etc.) are spatially and temporally disaggregated, addressing short time intervals and individual locations of places and social agents. This particular characteristic further signifies their main novelty in comparison with conventional data for cities.

At the same time, as the available data sources increase in number, the produced datasets increase in diversity. Although the current capabilities of computing systems allow the storage, processing, analysis, and visualization of large-scale data, integration remains a challenge. Even though advances have been made in other scientific fields (e.g. computational biology, medicine, artificial intelligence etc.), the domains of urban geography, planning, and spatial analysis have limited contribution, in this regard. In tackling the multifarious social, economic, and environmental challenges of rapid urbanization, planners and policy makers need supporting frameworks to capitalize on the new possibilities given by emerging sources of social urban data. This calls for new methods, tools, and theories to decipher the potential and limitations of these sources, to enable their integration, and to potentially improve the understanding of complex urban dynamics. This thesis makes a step towards this goal.

The thesis proposes a framework of methods and tools for the integration, visualization, and exploratory analysis of large-scale heterogeneous urban data with spatial and temporal attributes, so as to contribute to the current research on urban dynamics. In the first place, it introduces the concept of 'social urban data' to describe the spatiotemporal datasets that originate from emerging sources (e.g. sensors, mobile phones, geo-enabled social media, and LBSNs) and have potential to

represent aspects of social and spatial networks in cities, but are also characterized by semantic ambiguity. After defining their distinguishing characteristics and identifying their strengths and weaknesses (Chapter 2), the data integration component of the proposed framework is introduced (Chapter 3). This comprises a methodology for the semantic integration of (social) urban data from heterogeneous sources and their transformation into multidimensional linked urban data. Following an ontology-based approach to data integration, in addition to adopting Semantic Web and Linked Data technologies, the methodology is specifically oriented to the domain of urban analytics and aims to counter the lack of domain-specific guidelines for generating and publishing Linked Data (Radulovic et al., 2015; Villazón-Terrazas et al., 2012; Villazón-Terrazas et al., 2011). An implementation of this methodology is also demonstrated with a comprehensive example of data transformation into LOD, concerning a large-scale dataset about the entire public transportation system of Athens, Greece. In facilitating the semantic interoperability between social urban data, a domain ontology of public transportation systems (Chapter 3) and an upper-level ontology of urban networks (Chapter 4) are also developed. To support the adoption of the methodology by planners and policy makers, the framework additionally comprises a set of web-based tools for the visual exploration of ontologies and linked urban data, supported by interactive user interfaces (Chapter 4). The attributes that can be sourced from emerging social urban data are then used as proxies to estimate socio-demographic variables of individuals (e.g. home location, age range, gender etc.), aspects of human movement (e.g. radius of gyration, flows on the road network), and social activity (e.g. POI visits, activity spaces, activity-related semantics and sentiments etc.). Therefore, a set of methods is introduced to source this information from various social urban data. The derived attributes are used to enrich measurements of functional density and diversity with disaggregate human activity attributes (Chapter 5). Finally, the proposed framework is complemented by a web-based system that is developed to accommodate and combine the methods introduced in the thesis under a single platform. It comprises a set of components and modules that specifically enable the visualization and exploratory analysis of human activity and movement patterns in space and time, and also support the integration of traditional and emerging social urban data from multiple sources. An instance of the platform is also put to use in understanding the spatial distribution of activity and movement patterns of different social categories before, during, and after a city-scale event in the city of Amsterdam (Chapter 6).

In the remainder of this Chapter, the findings of the research are, first, discussed by revisiting the research questions formulated in Chapter 1. Next, the limitations of the research are highlighted and classified into four main categories. Following this, Sect. 7.3 presents the overall conclusions by first answering the main research question, followed by a summary of the major findings. Afterwards, it presents potential applications to practice and research and, finally, concludes with pointers to future research.

§ 7.2 Discussion of the Research Findings

§ 7.2.1 Revisiting the Research Questions

This section discusses the major findings of the research, arranged according to the various research questions formulated in the Introduction (Chapter 1).

What are the characteristics that distinguish emerging social urban data from traditional ones?

In Chapter 2, traditional data for cities are juxtaposed with data deriving from emerging sources (i.e. sensors, mobile phones, geo-enabled social media, and LBSNs), based on eight characteristics to elicit what distinguishes the latter from the former. These characteristics are namely: *diversity*, *scale*, *timeliness*, *structure*, *spatiotemporal resolution*, *semantic expressiveness*, *representativeness*, and *veracity*. Unlike existing literature that assigns some of these attributes only to large-scale (or 'big') data (e.g. (Boyd & Crawford, 2012; Kitchin, 2014a; Mayer-Schönberger & Cukier, 2013; Zikopoulos et al., 2012), it is argued here that the eight aforementioned characteristics are in fact inherent to all types of datasets, regardless the source. The difference lies in the extent to which each of the characteristics typifies a certain data type or source. Therefore, the eight characteristics described in Chapter 2 serve as a framework to assess the strengths and weaknesses of each source, either traditional or emerging, against each feature.

Current literature elaborates primarily on generic definitions and characteristics, classifying new data sources under the – also generic – 'Big Data' umbrella, which lacks distinct context. Arguing that this term is insufficient to describe the specificities of data sources about cities, the research introduces the term 'social urban data' instead. The latter encompasses the data that (a) are generated either directly from people or indirectly from people's actions, (b) derive from sources such as sensors, mobile phones, geo-enabled social media, and LBSNs, (c) are multidimensional in nature, meaning that they are spatially and temporally referenced, and as such (d) they can be used to infer spatial, temporal, and social aspects of human movement, activity, and social connectivity, at the disaggregate level, (e) but are less structured and more semantically ambiguous than traditional urban data.

However, social urban data should not be regarded as a unified category of data. In fact, there exist several differences within and between the sources producing social urban data. As it is discussed in Chapter 2, and shown in the case examples in Chapters

3 and 6, according to the source that generates them, social urban data may have varied levels of structure, semantic expressiveness, spatiotemporal resolution, and also represent different population samples. For instance, the amount of geo-referenced data on Twitter is generally smaller than the corresponding one on Instagram. At the same time, different usage patterns in different cities (or countries) drastically affect the sample of data that can be sourced. In some cases, (e.g. in China) none of the two aforementioned social media can be used as proxies to infer aspects of urban dynamics, since the penetration rate is negligible. Instead, one should seek alternative sources (e.g. Weibo, in the case of Chinese cities) that make their data available through APIs. Similarly, the penetration rates of mobile phone providers, as well as the density of devices in a sensor network arrangement, have a strong influence on the representativeness of the obtained sample. This also explains why putting everything under the generic umbrella of 'Big Data' can be rather misleading, especially when it comes to understanding their implications.

In answering the above research question, the most distinguishing characteristic that differentiates emerging social urban data in general from traditional ones, is the purpose guiding their generation. Conventional data about cities (e.g. data from censuses, household travel surveys etc.) are generated ad hoc, i.e. to serve a specific purpose, which in turn drives the sampling frame, the data model, the associated meta-data, the frequency of updates, and the resolution. This subsequently results in high quality data, characterized by clean structure and high levels of trustworthiness in regards to the information included. On the contrary, the majority of social urban data are generated spontaneously, without a particular data model, while the purpose they serve may vary substantially. This generally results in muddled data structures and ambiguous semantics, which in turn lead to lower quality data compared to traditional ones. This explains why data from emerging sources are not directly applicable to urban analytics, but instead require intermediate processing in order for aspects of individuals or urban environments to be derived or inferred (e.g. home location, age and gender estimation, travel or migration trajectories, social networks, type of activity etc.), as described in Chapters 5 and 6.

Overall, the present-day urban data landscape is characterized by a large diversity of sources, compared to previous periods. In addition, the frequent update rates of emerging data sources allow the analysis of shorter time intervals than it was possible hitherto. At the same time, though, the lack of structure and the ambiguity of accompanying semantics still constitute major challenges in taking full advantage of the capacity of new sources of urban data. The suitability of each source – or set of sources – to be used in the study of urban dynamics, primarily depends on the context of the case in question. In general, traditional urban data are characterized by high levels of structure, semantic expressiveness, representativeness, and veracity, but are weaker in terms of update frequency (timeliness) and spatiotemporal resolution. Sensor data are updated frequently, are highly structured, have high levels

of spatiotemporal resolution and veracity, and may vary substantially, based on the aspect they measure. However, they have almost negligible semantic information associated with them, while their representativeness depends on the coverage and density of devices comprising a sensor network. The same applies to mobile phone data (or CDRs), with the exception of diversity, since data records usually come in standard file formats (e.g. tabular data in csv, xml, json or similar formats) (e.g. (Calabrese, Smoreda, et al., 2011; Grauwin et al., 2015; Wang et al., 2015). Finally, data from geo-enabled social media (e.g. Twitter, Instagram, Sina Weibo etc.) and LBSNs (e.g. Foursquare) also have frequent update rates, usually high spatiotemporal resolution and semantic expressiveness, but are very weak in terms of structure, representativeness, and veracity.

How to transform heterogeneous data for cities into multidimensional linked urban data?

One of the major challenges in exploring complex aspects of urban systems is the integration of data deriving from multiple and diverse sources. Although the combination of various types of data could be valuable in filling in missing (values of) attributes, as well as in mitigating the weaknesses of one source by leveraging the strengths of another, the heterogeneities in data format, schema, structure, resolution, naming conventions, and level of aggregation raise several issues. In answering this challenge, a methodology for the semantic integration and transformation of heterogeneous urban data into multidimensional linked urban data is designed in Chapter 3. The methodology follows an ontology-based data integration approach and accommodates a variety of semantic (web) and linked data technologies. Overall, the methodology comprises three main processes, i.e. semantic integration, linked data generation, and publication to the LOD cloud, with each one consisting of several sub-processes, as follows:

- Semantic integration:
    - Selection of data sources and data preprocessing
    - Data analysis and modeling
        - Schema extraction
        - Resource naming strategy definition
        - Ontology design and development
            - Terms extraction
            - Reuse of existing ontologies and external structured vocabularies
            - Terms hierarchy and ontology conceptualization
            - Ontology evaluation
    - Mapping source data to the ontology (data transformation)
- Transformation into multidimensional linked urban data:
    - Establishing links with other sources
- Publication to the LOD cloud:

- Ontology and RDF dataset publication on the Web
- Documentation accessibility (human-readable and machine-processable)
- Registration into a Linked Data catalog and publication to the LOD cloud

Each one of the three main processes comprising the methodology resembles a different level of data openness, reusability, reproducibility, connectivity, and retrieval. In particular, the semantic integration process addresses the fusion of local data that can be either open or proprietary. In the linked data generation (data transformation) process, the integrated dataset that has resulted from the previous step, can be linked with datasets from other sources. Yet again, the obtained linked data can be either open or proprietary (i.e. exploited only within a certain group of stakeholders). On the contrary, the publication process to the LOD cloud is only applicable to linked open data that can be publicly retrieved, reused, republished, transformed, connected further with other datasets, and be exploited in various applications. Therefore, the degree of openness, reproducibility, and reusability scales linearly from one process to another.

This methodology can be replicated with relatively low effort and be applied (with minor adjustments) to different types of urban data, irrespective of the chosen sources. Moreover, the fact that it is based on ontologies enables the semi-automatic iteration of the data mapping for any future updates of the source data, provided that the latter maintain their initial schemas. This can be beneficial for contemporary social urban data, which are characterized by very frequent update rates. Besides minimizing data redundancy and ensuring semantic interoperability, ontologies can also be used as a basis for querying and retrieving the resulting linked datasets, e.g. from SPARQL endpoints. Although part of the methodology draws on generic methods for generating and publishing LOD, such as the one proposed by (Heath & Bizer, 2011a), it is specifically developed to cater to the domain of urban analytics.

A comprehensive example of urban data integration, transformation into multidimensional linked urban data, and publication to the LOD cloud is also presented in Chapter 3 to demonstrate the applicability of the methodology to large-scale spatiotemporal urban data from different sources. To support the interoperability between the data in question (i.e. regarding the entire public transport network of the city of Athens), a domain-specific ontology of public transportation systems (ROUTE Ontology) is also developed. The ontology can be used in the integration and linked data generation processes of similar data from the transportation and mobility domains. Although the example makes use of mostly authoritative data, rather than social urban data (an issue that is further discussed in Sect. 7.2.2, later in this chapter), the demonstrated methodology for data integration and transformation into LOD can be adjusted to serve data from emerging sources as well. This replicable methodology can potentially facilitate the generation and publication of LOD in the domains of urban geography, planning, and analytics, which are currently very scarcely represented on the LOD cloud, compared to other scientific domains.

How could urban planners, researchers and policy makers leverage the potential of multidimensional linked data in city analytics?

Although the methodology that is introduced in Chapter 3 provides a means to fuse data and information from multiple sources, it requires one to be familiar with the formalisms of ontologies and other semantic technologies. However, this is rarely the case with policy makers, urban planners, and researchers alike. The answer to the above question is covered in Chapter 4, where a set of web-based tools (the *OSMoSys* framework) for the visual representation of ontologies and multidimensional linked urban data is designed and demonstrated through benchmark cases. Having in mind that policy makers, urban planners, and researchers will increasingly be facing the need to integrate multidimensional data into urban models, they will also progressively come into contact with ontologies and linked data. Acknowledging this necessity, the proposed computational tools provide graphical user interfaces, in combination with navigational aids for browsing though and filtering interlinked data and knowledge models (i.e. ontologies), without requiring previous experience with the technologies involved. Unlike related examples (e.g. (S. M. Falconer et al., 2010; Heim et al., 2009; Stuhr et al., 2011) that depend on specialized software or require installation, the proposed tools are fully accessible through the Web and rely on open-source technology.

The *OSMoSys* set of tools uses force-directed graphs to visualize linked datasets and their underlying ontologies. Nodes represent the data instances, whereas links (or edges) represent the relationships between them. As data instances may correspond to any real-world entity (e.g. a city, a district, a POI, an individual etc.), the interactive graph representation enables the visual exploration of relationships between instances of different dimensions (i.e. spatial, social, temporal). For instance, a pair of social contacts (e.g. derived from a social media platform or inferred from mobile phone data), with each individual being an instance of the *foaf:Agent* class, could be linked with a specific POI location (i.e. an instance of the *osmosys:PointOfInterest* class) in a city (e.g. derived from GeoNames), and be further linked with a type of activity at a certain time frame, in the same graph.

It was found, though, that such multidimensional networks can easily become muddled as the number of nodes and links increases. To increase the readability and exploration potential, the proposed tools incorporate functions such as varied node sizes based on a node's centrality (i.e. the amount of instances it connects with), node clustering, semantic zooming, grouping (e.g. based on feature type), keyword search, and isolated views of local graphs (i.e. comprising only the nodes that a data instance is directly linked to). Besides the graph visualization, the proposed framework also provides a web ontology browser to browse through the components of an ontology that is uploaded to the system. This particular feature is not provided in the visualization of linked data.

The proposed tools can also be used by domain experts as a basis to evaluate ontologies – generally created by ontology experts who do not possess the expertise of the domain that is represented in the model – during the various development stages. In addition, they supplement the data integration methodology presented in Chapter 3 with visual means that deviate from the formalisms of semantic technologies.

What types of attributes can be derived from social urban data in relation to the dynamics of human activity?

The answer to this question is mainly covered in Chapter 5, where a set of methods and techniques are described, pertinent to the extraction of socio-demographic attributes of individuals, functional attributes of places, individual spatial movement patterns, and topical attributes of human activity from social urban data. The chapter also addresses how these approximated attributes help measure the functional density and diversity of urban areas, as well as the geographical extents of activity spaces over different periods of time. A part of the answer is also covered in Chapter 6, where it is demonstrated how these methods and techniques can be integrated into a software platform for urban analytics and implemented in the study of human activity patterns in a real-world case study.

The types of attributes that can be derived from a variety of social urban data sources pertain to: the home location of an individual, socio-demographic features such as gender, age range, and ethnicity, individual trajectories and activity spaces, land uses of POIs, as well as topics and sentiments about a certain type of activity. As stated, these attributes are in fact approximations of the actual ones and, therefore, cannot be considered fully accurate. The level of accuracy, as well as the type of attribute that can be derived, largely depend on the source and the quality of data that are generated.

From the above-mentioned attributes, data from sensors (e.g. GPS trackers) can only be used to derive individual trajectories. Therefore, they can facilitate the understanding of human movement at the disaggregate level, but are ineffective as regards the dynamics of social activity. Conversely, one can derive the approximate home location, the trajectory, and activity space of an individual or a group of people from mobile phone data (e.g. CDRs). Mapping phone activity (e.g. a call or SMS) to a certain POI location is possible by triangulating the closest cell phone towers, yet this estimation can be highly ambiguous. Moreover, due to the negligible semantics that accompany such datasets, it is difficult to make any estimate on the type of activity that is carried out or the demographics of the sample in question.

The sources that can be employed to derive the entire set of the above-mentioned attributes are data from geo-enabled social media and LBSNs. Their accuracy is lower compared to sensor data and CDRs, yet they outperform all other sources of social urban data in terms of semantic richness. For this reason, they can valuable proxies for

the spatial distribution of social activity patterns, as it is demonstrated in Chapter 6. However, the accuracy (and representativeness) varies from one social media platform to another. Also, especially in relation to the approximation of land uses, the obtained observations are limited by the set of POIs that are listed in these platforms (e.g. Foursquare) and their aggregation into predefined categories. Additional limitations with regard to the approximation of socio-demographic attributes, topical semantics, and sentiments are further discussed in Sect. 7.2.2. It is, therefore, suggested that (if available) multiple social urban data sources are combined together – using also the data integration methods that are presented in Chapter 3 – and cross-validated against reliable sources of urban data (e.g. censuses, surveys).

The incorporation of these attributes into urban analytics helps deviate from traditional approaches, in which people and places are usually perceived as aggregate (i.e. average, mean, or summed values) parameters within spatial subdivisions (e.g. census tracts).

How do different sources of social urban data influence the understanding of the spatiotemporal dynamics of human activity in cities?

After introducing methods and tools for data integration (Chapter 3), visual exploration of linked spatiotemporal data (Chapter 4), and derivation of various attributes of people and places from different social urban data (Chapter 5), it is examined how they can all be combined into a single platform and put to use in understanding spatiotemporal patterns of human activity in cities. In particular, Chapter 6 presents a novel web-based system (*SocialGlass*) that enables this combination and further allows to explore how different social categories of people use urban space over time, using their online social activity as proxy. The system combines data from various geo-enabled social media and LBSNs (i.e. Twitter, Instagram, Sina Weibo, Foursquare), sensor networks (i.e. GPS trackers, Wi-Fi cameras), and conventional socio-economic urban records, but also has the potential to employ custom datasets from other sources. The previously discussed methods for data integration and approximation of attributes are implemented in several modules of the system architecture and support the frontend GUI. The latter further accommodates a variety of visualization types and data filters to support the visual exploratory analysis of human activity in cities.

A real-world case study is also analyzed and used as a demonstrator of the capacities of the proposed web-based system in the study of urban dynamics (Chapter 6). The case study explores the potential impact of a city-scale event (i.e. the Amsterdam Light festival 2015) on the activity and movement patterns of different social categories (i.e. residents, non-residents, foreign tourists), as compared to their daily and hourly routines in the periods before and after the event. Besides assessing the deployment of the system in a real-world use case, the aim of this empirical exploration was also to evaluate the knowledge gained from different sources of social urban data, as regards the dynamics of human activity.

By analyzing 28 different variables, derived from different geo-enabled social media, through both visual exploratory analysis (using the *SocialGlass* GUI) and – global and local – spatial autocorrelation analysis, it is found that different social categories of people use urban space differently over time. In particular, residents' activity behavior – as inferred from their online social activity – appears more dispersed than the one of non-residents and foreign tourists, who tend to concentrate around specific districts. This is evident in the results of both visual exploratory and spatial autocorrelation analyses. Another interesting finding is that although the volume of activity for some social categories (i.e. foreign tourists and non-residents) increases in absolute numbers during the event, the spatial footprint of activity behavior of each group presents only slight differences over time. The observations on the spatial and temporal distribution of social activity are also reflected in the movement patterns of the three social categories in question (as inferred from the *SocialGlass* system). The trajectories followed by residents and foreign tourists traverse more neighborhoods than those of non-residents. Yet, the volume of resident flows is larger and their trajectories cover a wider area compared to the other social categories. The overall findings suggest that the event had an impact on the intensity of activity of non-residents and foreign tourists, but had almost no effect on its spatial distribution over time. Moreover, the local spatial autocorrelation analysis revealed that – in the case of Amsterdam – the places were activity tends to concentrate, irrespective of the social category, do not coincide with the places that are characterized by high density of POI locations.

Overall, the above findings suggest that it is necessary to consider different social categories, rather than aggregate populations, when studying the dynamics of human activity and mobility behavior in cities. Moreover, the data collection period and the data source play a crucial role, when it comes to anomalies that could be reflected in the collected data. In Sect. 6.5.2 – 6.5.4, it was shown that the consideration of a single source (e.g. Twitter or Instagram) could yield entirely different outcomes and, hence, biased interpretation of the result. Therefore, the ability of the system to integrate data from heterogeneous sources, along with the multiple visualization types and data filters it provides, prove beneficial in the spatiotemporal exploration of human activity in cities. Although this research presents a single case study regarding the city of Amsterdam, the platform has already been tested successfully for similar purposes (i.e. urban dynamics, with a focus on spatiotemporal human activity and movement behavior) in several cities worldwide, i.e. London, Milan, Paris, Rome (Psyllidis et al., 2015a); Como (Psyllidis, Bozzon, Bocconi, & Bolivar, 2015b); Rotterdam, Shenzhen (Gong, Bozzon, Psyllidis, & Yang, 2016); Boston, Jakarta, Singapore, Sydney, and Zurich (Psyllidis, Bozzon, Yang, & Mesbah, 2016 (in press)).

§ 7.2.2 Limitations of the Research

The thesis has introduced a variety of new computational methods and tools to address the research challenge, objectives, and questions, and has further employed data from a range of emerging sources that have only recently been considered in urban analytics. As explicitly stated in this research, emerging social urban data suffer from several ambiguities at – among others – the semantic, structural, and sampling level. These ambiguities lead to several biases that could be reflected in the tools and methods and, subsequently, affect the interpretation of the results. Other assumptions made throughout the various stages of research may also have similar effects in the findings. This section highlights these limitations and classifies them into the following categories: *limitations of social urban data as proxies for the analysis of urban dynamics*, *limitations of data integration and interlinkage methods*, *limitations of the proposed tools*, and *ethics and privacy*.

Limitations of social urban data as proxies for the analysis of urban dynamics

The use of emerging sources such as sensors, mobile phones, geo-enabled social media and LBSNs in the analysis of urban dynamics is associated with issues of data quality and representativeness. Quality is primarily affected by the context in which these types of data are produced. The context in turn impacts the level of precision and the semantics that accompany these data, which in turn influence the interpretation of the results. In the history of spatial analysis, geography, and urban and regional studies, it is the first time that researchers gradually employ data from sources that are designed to serve different purposes (e.g. to socialize on the Web, to promote one's work, to call contacts etc.) than the ones for which they are used (i.e. urban analytics). Moreover, one has almost no control on the way these data are generated. In mitigating the resulting data 'noise', data 'cleaning' processes are usually applied. Yet the decisions on what to include and what to omit from the analysis also carries some bias (Boyd & Crawford, 2012). This limitation also applies to this research, where various assumptions are made, especially in the attributes (e.g. home location, age range, gender, ethnicity etc.) derived from the social media data (Chapters 5 and 6).

Another limitation is related to sampling biases, which are pertinent not only to this research, but also to relevant studies using these emerging sources. This particular issue is crucial for assessing the representativeness of the considered datasets. In the case of sensor data, the sample of observations is generally dependent on the coverage of the sensor network in focus. However, it is almost impossible to make any assumptions about the demographics of the sample (with the exception of camera-generated data, but this raises issues of privacy). Therefore, the obtained observations cannot be disaggregated by age, gender, or any other socio-economic attribute,

meaning that only questions about *how* an activity is performed can be addressed, rather than *who* performs this activity. In the case of mobile phone data (e.g. CDRs), the representativeness of the sample depends on the penetration of the various mobile phone providers and the extent of the dataset they are willing to provide. However, it is generally difficult to identify the actual amount of unique users within the dataset, as one person may own more than one phone number with the same provider. As is the case with sensor data, it is also difficult to infer any socio-demographic characteristics of the population represented in the sample.

This research primarily employs data from geo-enabled social media and LBSNs. Mobile phone and sensor data are not used in the case studies, since the former are proprietary and the latter require extensive resources in sensing devices. Therefore, the datasets used here are limited to only those that are publicly available and can be retrieved from each platform's API (i.e. Twitter, Instagram, and Foursquare APIs). Moreover, the collected sample is also affected by the limits imposed by each API, which may relate to restrictions in terms of the amount of data that can be derived. For instance, in the case of Twitter, the publicly available feed provided by the steaming API is used instead of the proprietary 'firehose'. This subsequently limits the collected sample to only 1% of the entire set of public tweets, out of which only the geo-referenced posts are selected. However, it has been shown by (Morstatter et al., 2013) that the streaming API returns almost the entire set of geo-tagged posts within a predefined bounding box. In regards to the latter (i.e. the predefined area within which one is allowed to retrieved geo-referenced posts), the Twitter API allows large areas to be covered at once. Instead of a bounding box, the APIs of Instagram and Sina Weibo require a center point to be defined first, along with a radius for the spatial query, which should not exceed 5,000m and 11,132m, respectively. This means that in order for larger regions to be covered, multiple circles need to be drawn. However, the areas at the intersection of circles contain several duplicates that first have to be filtered out, prior to being analyzed. The Foursquare API allows data to be retrieved from a predefined grid. As the API returns at most 50 records (i.e. POI venues) per grid cell, the size of latter needs to be small enough, so that less than 50 records are returned in each spatial query (therefore capturing all available POIs).

In addition to the bounding boxes, API requests based on keywords or hashtags could also carry some bias, in terms of the representativeness of the sample (Malik, Lamba, Nakos, & Pfeffer, 2015; Olteanu et al., 2015). Most importantly, the changing policies governing the amount of data that one is allowed to retrieve from a social media API, can largely affect the collected sample and, therefore, the results drawn from it. During the course of this research, Twitter, Instagram, and Foursquare API policies have undergone several changes, especially in regards to the geo-reference of posts, which plays a crucial role in the issues covered by this study.

The demographics of the collected sample may also vary from one platform to another. Members of certain social categories (e.g. white young males) may be over-represented, while others (e.g. people over 65 years of age) could resemble only a small portion of the collected sample (Hargittai, 2007; Mislove et al., 2011). Other factors that could affect the representativeness of the collected data pertain to the penetration rate of a platform (e.g. Twitter, Instagram, and Foursquare are not used in China), the usage patterns, and cultural biases (e.g. in terms of content-sharing behavior, interpretation of content semantics etc.). In Chapter 6, it was shown that temporal biases could also exist in the collected datasets (e.g. difference in activity patterns between weekdays and weekends, during public holidays/celebrations or city-scale events etc.). Overall, in using social urban data – especially those generated from social media platforms – to analyze the dynamics of human activity, it is necessary to consider that the online social activity generally represents a sample of the actual one (Miller & Goodchild, 2014). However, as reliable data about the actual activity behavior of people are scarcely available, it is often difficult to measure and evaluate the extent to which online social data are representative of reality.

Limitations of data integration and interlinkage methods

The major limitation of the data integration and interlinkage methods that were mainly presented in Chapter 3, is that they require familiarity with semantic (web) and linked data technologies, as well as with concepts and processes of ontology engineering. To a certain extent, this issue could be mitigated with the tools that were introduced in Chapter 4, yet the latter are primarily focused on visualization, rather than on ontology design and linked data generation. More specialized services and tools are needed to further simplify these processes and, hence, increase the adoption levels of ontology-based data integration and interlinkage technologies by urban planning researchers and policy makers.

Regarding the source data, a limitation that was also encountered in the process of collecting appropriate ones for the use case, pertains to the legal terms that accompany the data to be fused and linked. Although city organizations increasingly publish urban data as open data, there are currently only a few examples where the licenses clearly specify whether or not they can be further processed, republished, and reused in different applications. As regards social urban data, there also exist several aspects that could hinder their publication as LOD. In particular, anonymization techniques (e.g. generalization, aggregation, suppression etc.) need to be applied prior to publishing CDRs or social media data as linked data, so that both explicit (e.g. user names, phone numbers etc.) and quasi (e.g. place of residence, date of birth, ethnicity etc.) identifiers are excluded and privacy is preserved. In fact, certain types of social data, such as CDRs, can only be linked with other datasets at the local level, instead of being published to the LOD cloud, given the sensitive information they contain and their proprietary nature.

In regards to the LOD cloud, although there is currently an abundance of generic datasets, domain-specific linked data are still poorly represented. At present, data pertinent to human movement behavior, social activity in cities, flows between urban systems, among others that are relevant to this research have negligible presence on the LOD cloud. For this reason, in the example presented in Chapter 3 the total amount of links that were established is relatively limited (i.e. 16,080 in total), compared to the average amount of LOD cloud links (about 35% of the existing LOD datasets contain more than 20,000 links (Schmachtenberg et al., 2014)). The adoption of the methodology presented in Chapter 3 would hopefully facilitate the generation of linked data that cover the aforementioned aspects of urban dynamics, so that further links can be established between them and to the already existing linked geospatial data (e.g. GeoNames and Linked Geo Data).

Limitations of the proposed tools

The main tools proposed by this research are the *OSMoSys* framework for web-based ontology visualization and exploration of linked urban data (Chapter 4), and the *SocialGlass* web-based system for the visualization and exploratory analysis of human activity dynamics in cities (Chapter 6).

As regards the *OSMoSys* framework, the current implementation does not provide editing or graph manipulation functions (e.g. reallocation of the graph nodes, node size customization, edge adjustment etc.). Although the RDF and OWL visualizer component is based on graph visualization, it does not accommodate network metrics (e.g. node centrality, degree distribution, betweenness etc.). Instead, external software platforms – such as Gephi (Bastian, Heymann, & Jacomy, 2009) – can be used, in this regard. Moreover, the placement of the nodes (i.e. classes or instances of a class) on the graph does not correspond to an actual geo-location in a continuous Cartesian space, like in GIS – although some nodes may be accompanied by spatial coordinates – but instead follows an a-spatial approach, similar to social network representations. Usability tests with stakeholders (e.g. urban planners, researchers, and policy makers) will be necessary in the future to evaluate the comprehensibility and applicability of the proposed framework of tools.

As part of the *OSMoSys* framework, Chapter 4 also introduced an upper-level ontology of urban networks, to facilitate the semantic integration of frequently used data in urban analytics. Although the ontology has been successfully tested for semantic consistency and conciseness (see Sect. 4.3.4.5), additional tests on completeness – ideally with urban planners and policy makers – have to be performed, as part of future work. At present, the ontology supports only English terms. In future versions, descriptions in other languages could also be incorporated, so that non-English urban data instances can be mapped to it.

In regards to the *SocialGlass* system, the main limitation pertains to methodological aspects of data processing (i.e. data collection, cleaning, filtering) designed to retrieve or infer attributes (e.g. home location, age range, gender etc.) primarily from social media data. The inherent biases characterizing these data, as discussed earlier in this section, play a crucial role in the interpretation of the obtained observations. In general, it is necessary to cross-validate the inferred attributes against more reliable sources, such as census data, yet relevant information is not always available at the disaggregate level. In relation to the currently implemented modules, there is also room for improvement as regards the integration of inferred interpersonal relationships (i.e. online contacts) into the map-based GUI. Moreover, the incorporation of modules for spatial statistics would largely benefit the platform. As is the case with *OSMoSys*, usability tests with city planning and policy stakeholders are an important part of the future agenda, in order to further evaluate the comprehensibility and applicability of the system.

Ethics and privacy

In this research, large-scale datasets from a variety of sources have been used to address the main objectives and challenges. In addition, several methods and tools have been designed and implemented with the aim to facilitate researchers and practitioners to combine different data sources in order to understand aspects of human dynamics in cities. This inevitably leads to a discussion about ethics and privacy preservation, with regard to the technologies and data involved.

The entire set of tools that have been developed for the purposes of this research use solely open source software, are provided under open licenses, and can be accessed through commonly-used web browsers. To further encourage accessibility and reproducibility, special care has been taken in terms of making their documentation publicly available (see e.g. Sect. 3.3.5.1; 3.3.5.2; 4.3.1; 4.3.4.4; 6.4.1). In addition, the datasets that have been used throughout this research are publicly available and either derive from publicly accessible web-based governmental repositories (see Sect. 3.3.1) or public APIs (see Sect. 6.5.1). The published version of the *ROUTE* linked dataset on the LOD cloud, used in the example presented in Chapter 3, also complies with the legal terms and licenses of the original data sources (see Sect. 3.3.1; 3.3.5.1). The aforementioned actions aim to ensure increased transparency and accessibility with respect to the proposed tools and methods.

However, the employment and analysis of social urban data that contain several attributes about individuals and their behavior at the disaggregate level may generally put privacy at risk. The integration of these data with conventional ones could also make the preservation of sensitive personal information more vulnerable. Besides explicit identifiers, such as a person's full name, cell phone number, and home

address, among others, social urban data may also contain quasi identifiers (e.g. ethnicity, gender, date of birth, social contacts, place of work etc.) that can indirectly reveal personal information of a user. On the other hand, the perception of privacy may vary depending on the time period, culture, social category, and age (Enserink & Chin, 2015; Janssen & van den Hoven, 2015). Therefore, it is difficult to determine exactly what is considered sensitive information by each person. Yet, as social urban data – due to their nature and size – are collected without the consent of each individual involved, it is important that certain security measures are taken prior to analyzing the collected datasets.

In this research, anonymization procedures have been applied to ensure that information pertinent to explicit and quasi identifiers about individuals has been removed. Alternative techniques may include generalization (e.g. using ranges instead of exact feature values), aggregation (e.g. by social category, by weekday etc.), or anatomization (i.e. disassociation of quasi identifiers from explicit ones).

§ 7.3 Conclusions and Outlook

Having discussed the research findings and answers to the research questions, this section presents the overall conclusions by first answering the main research question, followed by a summary of the major findings (aligned with each research question). Afterwards, it presents potential applications to practice and research and, finally, concludes with pointers to future research.

§ 7.3.1 Overall Conclusions

The increasing availability of data for cities that are generated by emerging sources, such as sensor networks, mobile phones, geo-enabled social media, and LBSNs have the potential to provide new insights into urban dynamics, but also create new challenges for urban planners, researchers, and policy makers. These data are mainly characterized by heterogeneity, owing to the variety of sources and the diversity of purposes they serve, and multidimensionality, meaning that the information they contain may simultaneously address spatial, social, temporal, and topical features of people and places. In addition, they offer new perspectives on how complex socio-spatial phenomena in cities change over shorter time intervals, compared to the sparsely updated conventional urban data. On the downside, though, is the muddled

data structure, the ambiguous semantics of the contained information, and the several biases (of contextual, demographic, cultural, geographic, technological, or other nature). The integration of emerging data sources with traditional ones into urban analytics to extract richer descriptions of city dynamics remains a challenge.

This challenge subsequently led to the main question of this research, which was:

"How to integrate heterogeneous and multidimensional social urban data into the analysis of human activity dynamics in cities?"

As an answer to the above question, this research proposes a framework of methods and tools for the integration, visualization, and exploratory analysis of large-scale social urban data from multiple sources to facilitate the analysis of human activity dynamics in cities. In particular, the framework comprises a set of methods for the collection of data, mainly from geo-enabled social media and LBSNs (Chapters 5 and 6), the integration and interlinkage of spatiotemporal urban data and their transformation into multidimensional linked urban data (Chapter 3), the extraction of socio-spatial attributes (Chapter 5), and their incorporation into existing and new measurements for human movement and functional diversity (Chapter 5). Moreover, the framework introduces new web-based tools that implement these methods and facilitate the exploitation of linked urban data by city stakeholders (Chapter 4), as well as enable the interactive visualization and exploratory analysis of human activity patterns, by combining emerging with traditional urban data (Chapter 6). Additional contributions of the research in support of the answer to the main question include an ontology of urban transportation entities (Sect. 3.3.2.3), an ontology of urban networks (Sect. 4.3.4), and a publicly available linked dataset of the entire public transport network of Athens, Greece (Sect. 3.3). The ontologies and dataset can be used to inform transportation and mobility models (for the analysis and simulation of urban mobility and human movement behavior), and to semantically annotate and integrate data pertinent to the different networks comprising cities (e.g. street network, transport network, social networks etc.). The various tools and methods proposed by this research can also be used independently and can be further adapted to serve other types of social urban data (e.g. CDRs) and explore relevant aspects of urban dynamics that were not elaborated in this thesis (e.g. human movement, socio-spatial diversity of cities in space and time etc.).

In addressing the main challenge, the research was divided into five steps, each one corresponding to a sub-question, the findings of which were previously discussed in Sect. 7.2.1. Each of the findings contributed to the answer of the main question. The following paragraphs summarize the major findings of the five steps/sub-questions:

– Social urban data do not comprise a unified category of data with common characteristics. According to the source that generates them, they may be characterized

by varied levels of structural clarity, scale, timeliness, semantic expressiveness, spatiotemporal resolution, veracity, and representativeness (in terms of the population they address and in relation to actual features of the urban environment they represent). Yet, the most distinguishing characteristic that differentiates emerging social urban data from traditional ones, is the purpose guiding their generation. Unlike conventional urban data, they are not produced ad hoc and, as such, they contain contextual, technological, geographical, demographic, and cultural biases, which in turn affect the overall data quality. In using social urban data as proxies for the analysis of urban dynamics, the identification of these biases is of critical importance to the interpretation of the obtained results. Therefore, traditional data processing methods are not sufficient to extract meaningful knowledge from emerging social urban data. To leverage the intrinsic biases of social urban data and to extract unambiguous knowledge about the dynamics of cities, the integration of data from multiple sources is deemed necessary. Cross-validation against more reliable sources of urban data (e.g. census records) is also recommended (RQ1 – Chapter 2).

– In mitigating the various heterogeneities, a methodology for the transformation of heterogeneous data for cities into multidimensional linked urban data is designed. The methodology follows an ontology-based data integration approach and accommodates a variety of semantic (web) and linked data technologies. The transformation of heterogeneous data for cities into multidimensional linked urban data has potential to provide richer descriptions of urban dynamics. Moreover, their publication to the LOD cloud facilitates their discovery and exploitation by stakeholders of different (city) domains. The methodology can be replicated and adapted to serve different types of (social) urban data, irrespective of the chosen sources. As it is based on ontologies, it also enables the semi-automatic iteration of the data mapping for any future updates of the source data, provided that the latter maintain their initial schemas (RQ2 – Chapter 3).

– To further facilitate the adoption of linked data by urban planners, researchers, and policy makers, and encourage them to consume multidimensional linked urban data in urban analytics, a set of web-based tools (*OSMoSys*) for the visual representation of ontologies and linked data is designed and developed. The tools provide graphical user interfaces, in combination with navigational aids for browsing though and filtering interlinked data and knowledge models, without requiring previous experience with the technologies involved. The proposed tools can also be used by domain experts as a basis to evaluate ontologies during the various development stages (RQ3 – Chapter 4).

– Besides multidimensional linked urban data, it is also possible to derive several attributes of people and places from different geo-enabled social media content and LBSN data. To extract these attributes, a set of methods and techniques are described. In general, socio-demographic attributes of individuals, functional attributes of places, individual spatial movement patterns, and topical attributes of human activity are

possible to be derived from social media content. The incorporation of these attributes into urban analytics helps deviate from traditional approaches, in which people and places are usually perceived as aggregate uniform parameters within spatial subdivisions (RQ4 – Chapter 5).

– The aforementioned methods are combined and incorporated into a novel web-based system (*SocialGlass*) to facilitate the exploration of human activity dynamics in cities, using both traditional and emerging social urban data. The system accommodates a variety of visualization types and data filters to support the visual exploratory analysis of the spatiotemporal dynamics of human activity. A real-world case study is also analyzed and used as a demonstrator of the capacities of the proposed web-based system, and as a basis for investigating the influence of different data sources on the understanding of human activity in cities. The findings of the case study suggested that it is necessary to consider different social categories of people, rather than aggregate populations, when studying the dynamics of human activity and movement behavior. Moreover, if social urban data – especially online social media – are used as proxies for the analysis of urban dynamics, the data collection period and the data source play a crucial role, when it comes to anomalies that could be reflected in the collected data, which could in turn lead to biased interpretations (RQ5 – Chapter 6).

§ 7.3.2 Applications to Practice

The methods and tools of the proposed framework could be used in their entirety or independently in urban planning and policy making. Practitioners in these fields are expected to be increasingly encountered with a growing amount of urban data from a variety of sources in the near future, as these become available to a wider audience of stakeholders. Conventional methods of data processing, analysis, and modeling could prove insufficient in handling the heterogeneity and multidimensionality of the emerging urban data sources. Given the increasing urbanization rates worldwide, the efficient understanding and measurement of the dynamics of urban systems, in terms of inter- and intra-urban flows of people and goods, urban mobility and human movement behavior, and the distribution of social activity over space and time at the disaggregate level is essential to the management of planning of livable urban environments. The computational methods and tools presented in this research can be useful in this regard.

In particular, the proposed data integration and interlinkage methodology (Chapter 3) and linked data visualization tools (Chapter 4) can facilitate policy makers to bring together and combine data from a variety of sources or city sectors, semantically annotate ambiguous data to provide richer descriptions of urban systems and activities, understand the links between the components comprising cities, and to

transform implicit information of spatiotemporal urban data into knowledge about city dynamics. The ROUTE linked dataset (Sect. 3.3) is an example of how authoritative data sources from different public organizations can be fused and linked to external datasets from other domains. Also, the *OSMoSys* ontology of urban networks (Sect. 4.3.4) could help establish links between data that individually describe different components of urban systems (e.g. transportation, housing, economy, social networks etc.). Given that sources such as sensors, mobile phones, geo-enabled social media, and LBSNs have been scarcely used in policy making and planning, it is important that the several biases characterizing these sources are identified and addressed at the early stages of analysis, prior to selecting the data sources to be used. The discussion of the inherent characteristics of emerging social urban data sources, presented in Chapter 2, could give further insights into the various potential and limitations they bring to urban analytics.

In the initial exploration of urban dynamics, policy makers and planners could also benefit from the interactive visual exploratory analysis tools provided by the *SocialGlass* system, presented in Chapter 6. In particular, they could use the system to (a) create new experiments, (b) define the data collection period and the area of study, (c) invoke data crawling from different social media platforms, (d) upload custom datasets, (e) extract data attributes, (f) visualize and filter the collected datasets, (g) perform interactive visual exploratory analysis of human activity and mobility dynamics, and (h) export the results for integration into other specialized tools for further analysis. Finally, the methods to derive attributes from social urban data at the disaggregate level (Chapter 5), also implemented in the *SocialGlass* system (Chapter 6), could be used to inform planning support systems (PSS), decision support systems (DSS), or spatial microsimulation models.

§ 7.3.3 Applications to Research

The adaptability of the methods and tools comprising the proposed framework enables them to serve scientific fields beyond urban science and spatial analysis, such as computational social science, urban geography, GIScience, and (human) mobility studies. There is already a wealth of research on the laws governing human trajectories and their configuration over space and time, using empirical data from GPS trackers and mobile phones (i.e. CDRs). The tools and methods presented here could further benefit this area of research, especially in terms of data integration to enable correlations between e.g. (human) mobility, social networks, land uses, and housing prices. Unlike human mobility studies, there still exists little research on the dynamics of human activity in cities and the relationship between social networks and geographic space. The methods for data interlinkage (Chapters 3 and 4), as well

as for the extraction of attributes from social urban data (Chapter 5), in combination with the *SocialGlass* system (Chapter 6) could be useful in this regard. Another potential application to research is the characterization of urban districts, based on the behavioral patterns of individuals over space and time. The semantic (topic-based) and sentiment analysis of social activity data, based on the corresponding modules of the *SocialGlass* system, could offer new dimensions to the characterization of urban space.

Besides enabling correlations between various components of the urban environment, the data integration methodology (Chapter 3) could also support the addition of value to the interpretation of the collected data. Moreover, it could assist in generating and publishing new linked datasets to the LOD cloud that are specifically oriented to how urban systems function. In this way, the current misrepresentation of such data on the LOD cloud could be substantially mitigated and researchers could benefit from datasets that are of higher quality than the original ones. The availability of linked data that are site-specific – as is the case with the ROUTE dataset, presented in Sect. 3.3 – could essentially facilitate the comparative study of disparate urban systems. In the near future, it is expected that longitudinal disaggregate data, derived from the continuous collection and storage of (near) real-time social urban data, will inform urban simulation and prediction models at an unprecedented level of detail.

§ 7.3.4 Future Research

This section provides a set of pointers to future research that extend the scope, objectives, and findings of this thesis.

Comparative assessment of emerging and traditional sources of urban data

One of the objectives of this research was to explore and identify the characteristics that distinguish social urban data from traditional ones, and to understand their strengths and weaknesses as sources for the analysis of urban dynamics. Chapter 2 presented and discussed eight characteristics that are inherent to both emerging and traditional urban data in general, but argued that the degree to which each one of them typifies a certain data type is fully depended on the source in question. An important extension to this would be to operationalize the extent to which these characteristics are inherent to each data type and potentially identify new ones. This would subsequently provide quantifiable measurements of the strengths and weaknesses of social urban data as sources for city analytics. These measurements would be based on the collection of various empirical urban data from both emerging and traditional sources and on cross-check analysis of their attributes, in relation to aspects of the

urban environment (e.g. spatial and temporal distribution of human activities, inter- and intra-urban flows of people and goods, social fragmentation etc.).

Data integration

As an application of the proposed methodology for data integration and interlinkage (Chapter 3), this research demonstrated a comprehensive example of linked data generation and publication, concerning a large-scale spatiotemporal dataset (i.e. the ROUTE linked dataset). The source data of the use case derived from three different public transportation organizations. Drawing on the adaptability of the proposed methods, future research could be focused on applying the methodology to the generation of linked data from emerging sources, such as geo-enabled social media, LBSNs, and mobile phones, with the requirement that legal compliance is ensured. Links could then be established to traditional sources of urban data (e.g. census, real-estate records, travel surveys etc.) to provide richer descriptions of the urban environment and its dynamics. Of particular interest would be the creation of links between data about human social networks and spatial data, to better understand the extent to which social relationships are affected by the physical structure of cities. The *OSMoSys* ontology of urban networks, presented in Sect. 4.3.4, could be useful in this regard and can be further extended with additional concepts (i.e. classes, subclasses, and relationships) derived from empirical data.

Multidimensional models of urban dynamics

This research has introduced a set of methods and tools to extract socio-demographic attributes from disaggregate social urban data (Chapters 5 and 6). These attributes could be used to define new metrics that simultaneously address the two main components of urban systems, i.e. built infrastructure and social networks, and their evolution over time. Moreover, the analysis of multilayer networks (Boccaletti et al., 2014) in the context of complex systems such as cities, is a topic that deserves further research. This approach would allow to exploit the wealth of research in both social and spatial networks analysis; areas that are usually studied separately from one another. By combining spatial properties of social networks with social attributes of physical space, future research could focus on the development of coupled urban models that simultaneously address multiple aspects of urban dynamics. Examples could include multidimensional models of human movement and social connectivity, activity patterns and social ties, transport flows and information flows, among others. Data fusion would be of significant importance in the development of such models.

Comparative studies of urban systems using the developed computational tools

This research resulted in a framework for the study of urban dynamics which, among others, provides computational tools that facilitate the integration, visualization, and exploration of large-scale social urban data from multiple sources (i.e. *SocialGlass* and *OSMoSys*). The spatial distribution and temporal evolution of activity patterns of different social categories of people, as inferred from online social activity, have also been explored in a real-world case study, concerning the city of Amsterdam. In testing the generalization of the findings and the adaptability of the tools, additional comparative studies of urban dynamics across several urban systems, in both developed and developing countries, are deemed necessary. As previously mentioned, these studies could be focused on the coupling of spatial networks of activities and social networks of interactions, to provide fine-grained characterizations of urban space. These characterizations could help measure the diversity of urban environments and examine whether there exist common laws governing the social activity of different groups, irrespective of the spatial diversity. Attempts have already been made in this regard, as stated in Sect. 7.2.1, and additional ones are currently being undertaken. Overall, the framework proposed by this research has potential to open avenues of quantitative explorations of urban dynamics by employing a wide range of available data sources, contributing to the development of a new science of cities.

TOC

References

Abdelmoty, A. I., Smart, P., & Jones, C. B. (2007). *Building Place Ontologies for the Semantic Web: Issues and Approaches*. Paper presented at the 4th ACM workshop on Geographical Information Retrieval.

Alhasoun, F., Almaatouq, A., Greco, K., Campari, R., Alfaris, A., & Ratti, C. (2014). *The City Browser: Utilizing Massive Call Data to Infer City Mobility Dynamics*. Paper presented at the The 3rd International Workshop on Urban Computing (UrbComp 2014), New York.

Alonso, W. (1964). *Location and Land Use*. Cambridge, MA, USA: The MIT Press.

Amini, A., Kung, K., Kang, C., Sobolevsky, S., & Ratti, C. (2014). The Impact of Social Segregation on Human Mobility in Developing and Industrialized Regions. *EPJ Data Science, 3*(6), 1-20. doi:10.1140/epjds31

Andris, C. (2011). *Methods and Metrics for Social Distance*. (PhD Dissertation), Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

Andris, C. (2016). Integrating social network data into GISystems. *International Journal of Geographical Information Science*, 1–23. doi:10.1080/13658816.2016.1153103

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis, 27*(2), 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Antoniou, G., & van Harmelen, F. (2009). *A Semantic Web Primer*. Cambridge, Massachusetts: The MIT Press.

Axhausen, K. W. (2007). Activity Spaces, Biographies, Social Networks and their Welfare Gains and Externalities: Some Hypotheses and Empirical Results. *Mobilities, 2*(1), 15–36. doi:10.1080/17450100601106203

Balduini, M., Bozzon, A., Valle, E. D., Huang, Y., & Houben, G.-J. (2014). Recommending Venues Using Continuous Predictive Social Media Analytics. *IEEE Internet Computing, 18*(5), 28 - 35. doi:10.1109/MIC.2014.84

Barnaghi, P., Wang, W., Dong, L., & Wang, C. (2013). A Linked-Data Model for Semantic Sensor Streams. 468-475. doi:10.1109/GreenCom-iThings-CPSCom.2013.95

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi : An Open Source Software for Exploring and Manipulating Networks*. Paper presented at the International AAAI Conference on Weblogs and Social Media (ICWSM'09), San Jose, Menlo Park, CA, USA.

Battle, R., & Kolas, D. (2012). Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web, 3*(4), 1–17.

Batty, M. (2007). Spatial Interaction. In K. K. Kemp (Ed.), *Encyclopedia of Geographic Information Science* (pp. 416–419). Thousand Oaks, CA: Sage.

Batty, M. (2009). Urban Modeling. In R. Kitchin & N. Thrift (Eds.), *International Encyclopedia of Human Geography* (Vol. 12, pp. 51–58). Oxford, UK: Elsevier.

Batty, M. (2012). Smart Cities, Big Data. *Environment and Planning B: Planning and Design, 39*(2), 191–193. doi:10.1068/b3902ed

Batty, M. (2013a). Big data, smart cities and city planning. *Dialogues in Human Geography, 3*(3), 274–279. doi:10.1177/2043820613513390

Batty, M. (2013b). *The New Science of Cities*. Cambridge, MA: The MIT Press.

Batty, M. (2013c). *Urban Informatics and Big Data. A Report to the ESRC Cities Expert Group*. Retrieved from (Unpublished):

Batty, M. (2015). *Data About Cities: Redefining Big, Recasting Small*. Workshop Paper. National University of Ireland, Maynooth. Maynooth, Ireland.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., . . . Portugali, Y. (2012). Smart cities of the Future. *The European Physical Journal Special Topics, 214*(1), 481-518. doi:10.1140/epjst/e2012-01703-3

Batty, M., Jiang, B., & Thurstain-Goodwin, M. (1998). Local Movement: Agent-based models of pedestrian flow. Retrieved from

Batty, M., & Torrens, P. M. (2005). Modelling and prediction in a complex world. *Futures, 37*(7), 745-766. doi:10.1016/j.futures.2004.11.003

Bauer, F., & Kaltenböck, M. (2012). *Linked Open Data: The Essentials – A Quick Start Guide for Decision Makers*. Vienna, Austria: Mono/monochrom.

Bayir, M. A., Demirbas, M., & Eagle, N. (2009). *Discovering SpatioTemporal Mobility Profiles of Cellphone Users*. Paper presented at the World of Wireless, Mobile and Multimedia Networks & Workshops (WoWMoM 2009).

Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R., & Giovannini, L. (2010). Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *Journal of Statistical Mechanics: Theory and Experiment, 2010*(05), P05001. doi:10.1088/1742-5468/2010/05/p05001

Bechhofer, S., Harmelen, F. v., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2004). OWL Web Ontology Language Reference. Retrieved from https://www.w3.org/TR/owl-ref/

Becker, H., Naaman, M., & Gravano, L. (2011). *Beyond Trending Topics: Real-World Event Identification on Twitter*. Paper presented at the 5th International AAAI Conference on Web and Social Media (ICWSM '11).

Bellini, P., Benigni, M., Billero, R., Nesi, P., & Rauch, N. (2014). Km4City ontology building vs data harvesting and cleaning for smart-city services. *Journal of Visual Languages & Computing, 25*(6), 827–839. doi:10.1016/j.jvlc.2014.10.023

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Benslimane, D., Leclercq, E., Savonnet, M., Terrasse, M.-N., & Yétongnon, K. (2000). On the definition of generic multi-layered ontologies for urban applications. *Computers, Environment and Urban Systems, 24*(3), 191–214. doi:doi:10.1016/S0198-9715(99)00059-9

Berners-Lee, T. (2006). Linked Data. *Design Issues*. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5), 34-43.

Berry, B. J. L., & Garrison, W. L. (1958). Recent developments in central place theory. *Papers in Regional Science, 4*(1), 107–120.

Berry, B. J. L., & Parr, J. B. (1988). *Market Centres as Retail Locations*. Englewood Cliffs, NJ: Prentice Hall.

Berry, B. J. L., & Pred, A. (1965). *Central Place Studies: A bibliography of theory and applications*. Philadelphia, PA: Regional Science Research Institute.

Bettencourt, L. M. A. (2013). The Origins of Scaling in Cities. *Science, 340*(6139), 1438 – 1441. doi:10.1126/science.1235823

Birkin, M. (2009). Geocomputation. In R. Kitchin & N. Thrift (Eds.), *International Encyclopedia of Human Geography* (pp. 376–381). Oxford: Elsevier.

Birov, S., Robinson, S., Poveda-Villalón, M., Suárez-Figueroa, M. C., García-Castro, R., Euzenat, J., . . . Tsagkari, K. Z. (2015). *Deliverable D3.3: Ontologies and Datasets for Energy Measurement and Validation Interoperability*. Retrieved from

Bishr, Y., & Kuhn, W. (2000). *Ontology-Based Modelling of Geospatial Information*. Paper presented at the 3rd AGILE Conference on Geographic Information Science, Helsinki, Finland.

Bittner, T., Donnelly, M., & Smith, B. (2009). A spatio-temporal ontology for geographic information integration. *International Journal of Geographical Information Science, 23*(6), 765–798. doi:10.1080/13658810701776767

Bivand, R. S. (2010). Exploratory Spatial Data Analysis: Software Tools, Methods and Applications. In M. M. Fischer & A. Getis (Eds.), *Handbook of Applied Spatial Analysis* (pp. 219–254). Berlin, Heidelberg: Springer Berlin Heidelberg.

Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., . . . Ziemlicki, C. (2012). Data for Development: the D4D Challenge on Mobile Phone Data. *ARXIV, eprint arXiv:1210.0137*.

Boccaletti, S., Bianconi, G., Criado, R., del Genio, C. I., Gómez-Gardeñes, J., Romance, M., . . . Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports, 544*(1), 1–122. doi:10.1016/j.physrep.2014.07.001

Bocconi, S., Bozzon, A., Psyllidis, A., Bolivar, C. T., & Houben, G.-J. (2015). *Social Glass: A Platform for Urban Analytics and Decision-making Through Heterogeneous Social Data*. Paper presented at the 24th International World Wide Web Conference (WWW 2015), Florence, Italy.

Bolivar, C. T. (2014). *City Usage Analysis using Social Media*. (MSc Thesis), Delft University of Technology (TU Delft), Delft.

Botts, M., Percivall, G., Reed, C., & Davidson, J. (2007). *OGC Sensor Web Enablement: Overview and High Level Architecture (OGC 07-165)*. Open Geospatial Consortium white paper.

Boulos, M. N. K., Resch, B., Crowley, d. N., Breslin, J. G., Sohn, G., Burtner, R., . . . Chuang, K.-Y. S. (2011). Crowdsourcing, Citizen Sensing and Sensor Web Technologies for Public and Environmental Health Surveillance and Crisis Management: Trends, OGC Standards and Application Examples. *International Journal of Health Geographies, 10*(67).

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society, 15*(5), 662–679. doi:10.1080/1369118x.2012.678878

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature, 439*(7075), 462–465. doi:10.1038/nature04292

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating Gender on Twitter*. Paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP '11).

Burger, M., & Meijers, E. (2011). Form Follows Function? Linking Morphological and Functional Polycentricity. *Urban Studies, 49*(5), 1127–1149. doi:10.1177/0042098011407095

Burke, J. A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., & Srivastava, M. B. (2006). *Participatory Sensing*. Paper presented at the Workshop on World-Sensor-Web (WSW'06): Mobile Device Centric Sensor Networks and Applications.

Button, K. J., Haynes, K. E., Stopher, P., & Hensher, D. A. E. (2004). *Handbook of Transport Geography and Spatial Systems* (Vol. 5 (Handbooks in Transport)). New York: Elsevier Science.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems, 12*(1), 141-151.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies, 26*, 301–313. doi:10.1016/j.trc.2012.09.009

Calabrese, F., Smoreda, Z., Blondel, V. D., & Ratti, C. (2011). Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data. *PLoS One, 6*(7), e20814. doi:10.1371/journal.pone.0020814

Camarda, D. V., Mazzini, S., & Antonuccio, A. (2012). *LodLive, exploring the web of data*. Paper presented at the 8th International Conference on Semantic Systems.

Cambria, E., Schuller, B. r., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *Intelligent Systems, IEEE, 28*(2), 15–21. doi:10.1109/MIS.2013.30

Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS One, 5*(7), e11596. doi:10.1371/journal.pone.0011596

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment, 2*(3), 199–219. doi:10.1016/S1361-9209(97)00009-6

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). *Exploring Millions of Footprints in Location Sharing Services*. Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011).

Christakos, G., Bogaert, P., & Serre, M. (2001). *Temporal GIS: Advanced Functions for Field-Based Applications*. Heidelberg: Springer-Verlag Berlin Heidelberg.

Christaller, W. (1933). *Die Zentralen Orte in Süd-Deutschland*. Jena: Gustav Fischer Verlag.

Ciuccarelli, P., Lupi, G., & Simeone, L. (2014). *Visualizing the Data City: Social Media as a Source of Knowledge for Urban Planning and Management*. Heidelberg: Springer.

Clark, C. (1951). Urban Population Densities. *Journal of the Royal Statistical Society. Series A (General), 114*(4), 490–496.

Cliff, A. D., & Ord, J. K. (1973). *Spatial Autocorrelation*. London: Pion.

Cliff, A. D., & Ord, J. K. (1981). *Spatial Processes: Models and applications*. London: Pion.

Communities, E. I. P. o. S. C. a. (2014). *Operational Implementation Plan*. Retrieved from Brussels:

Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., . . . Taylor, K. (2012). The SSN Ontology of the W3C Semantic Sensor Network Incubator Group. *Web Semantics: Science, Services and Agents on the World Wide Web, 17*, 25-32.

Council, S. C. (2014). Smart Cities Readiness Guide: The planning manual for building tomorrow's cities today. Redmond, Washington: Smart Cities Council.

Cramer, H., Rost, M., & Holmquist, L. E. (2011). *Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare*. Paper presented at the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11).

Crane, R. (2000). The Influence of Urban Form on Travel: An Interpretive Review. *Journal of Planning Literature, 15*(1), 3–23. doi:doi: 10.1177/08854120022092890

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). *The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City*. Paper presented at the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland.

Cruz, I. F., & Xiao, H. (2009). Ontology-driven Data Integration in Heterogeneous Networks. In A. Tolk & L. Jain (Eds.), *Complex Systems in Knowledge-based Environments: Theory, Models and Applications* (pp. 75-97): Springer.

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*(9), 1149–1170. doi:10.1080/00343400903365102

Del Bimbo, A., Ferracani, A., Pezzatini, D., D'Amato, F., & Sereni, M. (2014). *LiveCities: Revealing the Pulse of Cities by Location-Based Social Networks Venues and Users Analysis.* Paper presented at the 23rd international conference on World Wide Web companion (WWW).

Derczynski, L., & Bontcheva, K. (2014). *Pheme: Veracity in Digital Social Networks*. Paper presented at the 22nd Conference on User Modelling, Adaptation and Personalization (UMAP 2014) – Project Synergy Workshop.

Devlin, B. (2013). *Business Unintelligence: Insight and Innovation beyond Analytics and Big Data*. New Jersey, NJ, USA: Technics Publications, LLC.

Diao, M., Zhu, Y., Ferreira, J., & Ratti, C. (2015). Inferring individual daily activities from mobile phone traces: A Boston example. *Environment and Planning B: Planning and Design*, 1–21. doi:10.1177/0265813515600896

Domingue, J., Fensel, D., & Hendler, J. A. (2011). Introduction to the Semantic Web Technologies *Handbook of Semantic Web Technologies* (pp. 1-41). Berlin, Heidelberg: Springer.

Dutcher, J. (2014). What Is Big Data? Retrieved from https://datascience.berkeley.edu/what-is-big-data

Ekman, P. (1972). Universals and Cultural Differences in Facial Expressions of Emotion. In J. Cole (Ed.), *Nebraska Symporium of Motivation* (Vol. 19, pp. 207–282). Lincoln: University of Nebraska Press.

Enserink, M., & Chin, G. (2015). The end of privacy. *Science, 347*(6221), 490–491. doi:10.1126/science.347.6221.490

ESPON. (2007). *Study on Urban Functions: Final Report (1.4.3)*. Retrieved from Luxembourg:

Euzenat, J., & Shvaiko, P. (2013). *Ontology Matching* (2nd ed.). Berlin Heidelberg: Springer.

Falconer, S. (2010). OntoGraf. Retrieved from http://protegewiki.stanford.edu/wiki/OntoGraf

Falconer, S. M., Callendar, C., & Storey, M.-A. (2010). A Visualization Service for the Semantic Web. In P. Cimiano & H. S. Pinto (Eds.), *Knowledge Engineering and Management by the Masses* (Vol. 6317). Berlin Heidelberg: Springer Berlin Heidelberg.

Falquet, G., Métral, C., Teller, J., & Tweed, C. E. (2011). *Ontologies in Urban Development Projects*. London: Springer London.

Fernandez, S., Marsa-Maestre, I., Velasco, J. R., & Alarcos, B. (2013). Ontology Alignment Architecture for Semantic Sensor Web Integration. *Sensors (Basel), 13*(9), 12581-12604.

Fielding, R., Gettys, J., Gettys, J., Frystyk, H., Masinter, L., Leach, P., & Berners-Lee, T. (1999). Hypertext Transfer Protocol – HTTP/1.1. Retrieved from https://www.w3.org/Protocols/rfc2616/rfc2616.html

Fischer, M. M. (2006). *Spatial Analysis and GeoComputation: Selected Essays*. Berlin – Heidelberg: Springer Berlin–Heidelberg.

Fischer, M. M., & Wang, J. (2011). *Spatial Data Analysis: Models, Methods and Techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Forrester, J. W. (1969). *Urban Dynamics*. Cambridge, MA, USA: The MIT Press.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: SAGE.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley and Sons.

Fotheringham, A. S., & O'Kelly, M. E. (1989). *Spatial Interaction Models: Formulations and Applications*. Dordrecht: Kluwer Academic Publishers.

Fox, A., Eichelberger, C., Hughes, J., & Lyon, S. (2013). *Spatio-temporal Indexing in Non-relational Distributed Databases*. Paper presented at the IEEE International Conference on Big Data, Silicon Valley, CA.

Freeman, L. C. (1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks, 1*(3), 215–239. doi:10.1016/0378-8733(78)90021-7

Frith, J. (2014). Communicating Through Location: The Understood Meaning of the Foursquare Check-In. *Journal of Computer-Mediated Communication, 19*(4), 890–905. doi:10.1111/jcc4.12087

Gandon, F. L., Krummenacher, R., Han, S.-K., & Toma, I. (2011). Semantic Annotation and Retrieval: RDF *Handbook of Semantic Web Technologies* (pp. 117-155). Berlin, Heidelberg: Springer.

Gangemi, A., & Presutti, V. (2009). Ontology Design Patterns. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (2nd ed., pp. 221–243). Heidelberg: Springer-Verlag Berlin Heidelberg.

Gao, Q. (2013). *User Modeling and Personalization in the Microblogging Sphere.* (PhD Dissertation), Delft University of Technology (TU Delft), Delft. (2013-33)

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding Urban Traffic-Flow Characteristics: A Rethinking of Betweenness Centrality. *Environment and Planning B: Planning and Design, 40*(1), 135–153. doi:10.1068/b38141

Gehl, J. (1996). *Life Between Buildings: Using Public Space* (J. Koch, Trans.). Copenhagen: Arkitektens Forlag.

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis, 24*(3), 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., & Trasarti, R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal, 20*(5), 695–719. doi:10.1007/s00778-011-0244-8

Golab, L., & Özsu, T. M. (2003). Issues in Data Stream Management. *ACM SIGMOD Record, 32*(2), 5-14. doi:10.1145/776985.776986

Gómez-Pérez, A. (2004). Ontology Evaluation. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 251–273). Berlin – Heidelberg: Springer Berlin Heidelberg.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological Engineering*. London, UK: Springer London.

Gong, V., Bozzon, A., Psyllidis, A., & Yang, J. (2016). *Exploring Human Activity Patterns Across Cities through Social Media Data.* (MSc Thesis), Delft University of Technology, Delft, the Netherlands.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782. doi:10.1038/nature06958

Goodchild, M. F. (1986). *Spatial Autocorrelation*. Norwich: Geobooks.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211–221. doi:10.1007/s10708-007-9111-y

Goodchild, M. F. (2011). Formalizing Place in Geographic Information Systems. In L. M. Burton, S. A. Matthews, M. Leung, S. P. Kemp, & D. T. Takeuchi (Eds.), *Communities, Neighborhoods, and Health: Expanding the Boundaries of Place* (pp. 21–33). New York, NY, USA: Springer New York.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*, 110–120. doi:10.1016/j.spasta.2012.03.002

Goodspeed, R. (2013). *The Limited Usefulness of Social Media and Digital Trace Data for Urban Social Research*. Paper presented at the Seventh International AAAI Conference on Weblogs and Social Media.

Grabowicz, P. A., Ramasco, J. J., Gonçalves, B., & Eguíluz, V. M. (2014). Entangling Mobility and Interactions in Social Media. *PLoS One, 9*(3), e92196. doi:10.1371/journal.pone.0092196.t001

Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., & Ratti, C. (2015). Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong. In M. Helbich, J. J. Arsanjani, & M. Leitner (Eds.), *Computational Approaches for Urban Environments* (pp. 363–387): Springer International Publishing.

Grimm, S., Abecker, D. A., Völker, J., & Studer, R. (2011). Ontologies and the Semantic Web *Handbook of Semantic Web Technologies* (pp. 507-579). Berlin, Heidelberg: Springer.

Gröger, G., Kolbe, T. H., Czerwinski, A., & Nagel, C. (2008). *OpenGIS City Geography Markup Language (CityGML) Encoding Standard* (OGC 08-007r1). Retrieved from http://www.opengeospatial.org/standards/citygml

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199–220. doi:10.1006/knac.1993.1008

Guarino, N., & Giaretta, P. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. In N. J. I. Mars (Ed.), *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing* (pp. 25–32). Amsterdam, the Netherlands: IOS Press.

Guo, S. S., & Chan, C. W. (2010). *A tool for ontology visualizaiton in 3D graphics: Onto3DViz*. Paper presented at the 23rd Canadian Conference on Electrical and Computer Engineering (CCECE '10), Calgary, AB.

Haining, R. P. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design, 37*(4), 682–703. doi:doi: 10.1068/b35097

Hall, P. (1988). *Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century*. Oxford, UK: Basil Blackwell.

Hall, P., & Pain, K. (2006). *The Polycentric Metropolis: Learning from Mega-city Regions in Europe*. London: Routledge.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to Social Network Methods*. Riverside, CA, USA: University of California, Riverside.

Hargittai, E. (2007). Whose Space? Differences Among Users and Non-Users of Social Network Sites. *Journal of Computer-Mediated Communication, 13*(1), 276–297. doi:10.1111/j.1083-6101.2007.00396.x
Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2012). Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics, 151*(1-2), 304–318. doi:10.1007/s10955-012-0645-0
Heath, T., & Bizer, C. (2011a). *Linked Data: Evolving the Web into a Global Data Space* (1st ed.): Morgan & Claypool.
Heath, T., & Bizer, C. (2011b). Semantic Annotation and Retrieval: Web of Data *Handbook of Semantic Web Technologies* (pp. 191–229). Berlin, Heidelberg: Handbook of Semantic Web Technologies.
Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., & Stegemann, T. (2009). RelFinder: Revealing Relationships in RDF Knowledge Bases. In T.-S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, & W. Bailer (Eds.), *Semantic Multimedia* (Vol. 5887, pp. 182–187). Berlin Heidelberg: Springer Berlin Heidelberg.
Helbich, M., Arsanjani, J. J., & Leitner, M. E. (2015). *Computational Approaches for Urban Environments*. Switzerland: Springer International Publishing.
Herrera-Yague, C., Schneider, C. M., Couronne, T., Smoreda, Z., Benito, R. M., Zufiria, P. J., & Gonzalez, M. C. (2015). The anatomy of urban social networks and its implications in the searchability problem. *Sci Rep, 5*, 10265. doi:10.1038/srep10265
Hess, P. M., Moudon, A. V., & Logsdon, M. G. (2001). Measuring Land Use Patterns for Transportation Research. *Transportation Research Record, 1780*, 17–24.
Hillier, B. (1996). *Space is the Machins: A Configurational Theory of Architecture*. Cambridge, UK: Cambridge University Press.
Hope, A. C. A. (1968). A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society. Series B (Methodological), 30*(3), 582–598.
Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., & Mascolo, C. (2016). *Measuring Urban Social Diversity Using Interconnected Geo-Social Networks*. Paper presented at the 25th International World Wide Web Conference (WWW 2016), Montréal, Québec, Canada.
Hu, Y. (2005). Efficient and High Quality Force-Directed Graph Drawing. *Mathematica Journal, 10*(1), 37-71.
Hunt, J. D., Kriger, D. S., & Miller, E. J. (2005). Current operational urban land-use–transport modelling frameworks: A review. *Transport Reviews, 25*(3), 329-376. doi:10.1080/0144164052000336470
Institution, B. S. (2014a). Smart Cities – Vocabulary. London: BSI Standards.
Institution, B. S. (2014b). Smart City Concept Model – Guide to establishing a model for data interoperability. London: BSI Standards.
Jackson, M. O. (2010). *Social and Economic Networks*. Princeton, NJ, USA: Princeton University Press.
Jacobs, J. (1961). *The Death and Life of Great American Cities*. New York, NY, USA: Random House.
Janowicz, K., Scheider, S., Pehle, T., & Hart, G. (2012). Geospatial Semantics and Linked Spatiotemporal Data – Past, Present, and Future. *Semantic Web Journal, 3*(4), 321–332.
Janssen, M., & van den Hoven, J. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly, 32*(4), 363–368. doi:10.1016/j.giq.2015.11.007
Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems, 53*, 36–46. doi:10.1016/j.compenvurbsys.2014.12.001
Jones, C. B., Alani, H., & Tudhope, D. (2001). *Geographical Information Retrieval with Ontologies of Place*. Paper presented at the Foundations of Geographic Information Science International Conference (COSIT 2001), Morro Bay, CA, USA.
Kang, C., Ma, X., Tong, D., & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1702–1717. doi:10.1016/j.physa.2011.11.005
Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods – a survey. *ACM Computing Surveys, 39*(4), 10:11–43. doi:10.1145/1287620.1287621
Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography, 3*(3), 262–267. doi:10.1177/2043820613513388
Kitchin, R. (2014a). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE Publications Ltd.
Kitchin, R. (2014b). The real-time city? Big data and smart urbanism. *GeoJournal, 79*(1), 1–14. doi:10.1007/s10708-013-9516-8
Kitchin, R. (2015). *Data-driven, Networked Urbanism*. Maynooth University. Maynooth, Ireland. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2641802

Kloosterman, R., & Musterd, S. (2001). The Polycentric Urban Region: Towards a Research Agenda. *Urban Studies, 38*(4), 623–633. doi:10.1080/00420980120035259

Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A. (2004). The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In S. A. McIlraith, D. Plexousakis, & F. v. Harmelen (Eds.), *The Semantic Web – ISWC 2004: Proceedings of the Third International Semantic Web Conference* (pp. 229–243). Berlin - Heidelberg: Springer Berlin Heidelberg.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS One, 9*(6), e96180. doi:10.1371/journal.pone.0096180

Kwan, M.-P., & Schwanen, T. (2009). Quantitative Revolution 2: The Critical (Re)Turn. *The Professional Geographer, 61*(3), 283–291. doi:10.1080/00330120902931903

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Alstyne, M. V. (2009). Computational Social Science. *Science, 323*(5915), 721–723. doi:10.1126/science.1167742

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the Global Twitter Heartbeat: The Geography of Twitter. *First Monday, 18*(5). doi:10.5210/fm.v18i5.4366

Lenormand, M., Picornell, M., Cantu-Ros, O. G., Tugores, A., Louail, T., Herranz, R., . . . Ramasco, J. J. (2014). Cross-checking different sources of mobility information. *PLoS One, 9*(8), e105184. doi:10.1371/journal.pone.0105184

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS One, 9*(1), e86026. doi:10.1371/journal.pone.0086026

Llorente, A., Garcia-Herranz, M., Cebrian, M., & Moro, E. (2015). Social media fingerprints of unemployment. *PLoS One, 10*(5), e0128692. doi:10.1371/journal.pone.0128692

Lohmann, S., Negru, S., Haag, F., & Ertl, T. (2016). Visualizing Ontologies with VOWL. *Semantic Web, 7*(4), to appear.

Lösch, A. (1944). *The Economics of Location*. New Haven, CT, USA: Yale University Press.

Lowry, I. S. (1965). A Short Course in Model Design. *Journal of the American Institute of Planners, 31*, 53 – 64.

Lutz, M., & Klien, E. (2006). Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science, 20*(3), 233–260. doi:10.1080/13658810500287107

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). *Population Bias in Geotagged Tweets*. Paper presented at the 9th International AAAI Conference on Web and Social Media 2015 (ICWSM'15).

Mars, N. J. I. (1995). *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing* (1st ed.). Amsterdam, the Netherlands: IOS Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

McDonald, J. F. (1987). The identification of urban employment subcenters. *Journal of Urban Economics, 21*(2), 242–258.

McDonald, J. F., & Prather, P. J. (1994). Suburban Employment Centres: The Case of Chicago. *Urban Studies, 31*(2), 201–218. doi:10.1080/00420989420080201

McGovern, T. e. (2015). *Big Data Now*. Sebastopol, CA, USA: O'Reilly.

McKenzie, G., Janowicz, K., Gao, S., & Gong, L. (2015). How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*. doi:10.1016/j.compenvurbsys.2015.10.002

Métral, C., Falquet, G., & Cutting-Decelle, A.-F. (2009). *Towards Semantically Enriched 3D City Models: An Ontology-based Approach*. Paper presented at the GeoWeb Conference Academic Track - Cityscapes, Vancouver, Canada. http://www.isprs.org/proceedings/XXXVIII/3_4-C3/Paper_GeoW09/

Miller, H. J. (2010). The Data Avalanche is Here: Shouldn't We Be Digging? *Journal of Regional Science, 50*(1), 181-201.

Miller, H. J., & Goodchild, M. F. (2014). Data-driven geography. *GeoJournal, 80*(4), 449–461. doi:10.1007/s10708-014-9602-6

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). *Understanding the Demographics of Twitter Users*. Paper presented at the 5th International AAAI Conference on Web and Social Media (ICWSM'11).

Montenegro, N., Gomes, J. C., Urbano, P., & Duarte, J. P. (2012). A Land Use Planning Ontology: LBCS. *Future Internet, 4*(4), 65-82. doi:10.3390/fi4010065

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1-2), 17–23.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose*. Paper presented at the 7th International Conference on Web and Social Media (ICWSM '13).

Nikolov, A., Ferrara, A., & Scharffe, F. (2011). Data Linking for the Semantic Web. *International Journal on Semantic Web & Information Systems, 7*(3), 46–76. doi:10.4018/jswis.2011070103

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS One, 7*(5), e37027. doi:10.1371/journal.pone.0037027

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). *Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks*. Paper presented at the Workshop on the Social Mobile Web at ICWSM 2011.

Olteanu, A., Castillo, C., Diakopoulos, N., & Aberer, K. (2015). *Comparing Events Coverage in Online News and Social Media: The Case of Climate Change*. Paper presented at the 9th International AAAI Conference on Web and Social Media 2015 (ICWSM'15).

Openshaw, S. (1984). *The Modifiable Areal Unit Problem: Concepts and Techniques*. Norwich: Geo Books.

Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional issues and an application. *Geographical Analysis, 27*(4), 286–306.

Parr, J. B. (1987). Interaction in an Urban System: Aspects of Trade and Commuting. *Economic Geography, 63*(3), 223–240. doi:10.2307/143951

Pennacchiotti, M., & Popescu, A.-M. (2011). *A Machine Learning Approach to Twitter User Classification*. Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11).

Portugali, J. (2011). Complexity Theories of Cities: Implications to Urban Planning. In H. M. Juval Portugali, Egbert Stolk, Ekim Tan (Ed.), *Complexity Theories of Cities Have Come of Age* (pp. 221 - 244). Heidelberg: Springer.

Poveda-Villalón, M., García-Castro, R., & Gómez-Pérez, A. (2015). *Building an Ontology Catalogue for Smart Cities*. Paper presented at the eWork and eBusiness in Architecture, Engineering and Construction (ECPPM 2014).

Poveda-Villalón, M., Gómez-Pérez, A., & Suárez-Figueroa, M. C. (2014). OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. *International Journal on Semantic Web & Information Systems, 10*(2), 7–34. doi:10.4018/ijswis.2014040102

Prell, C. (2012). *Social Network Analysis: History, Theory and Methodology*. London: Sage Publications.

Psyllidis, A. (2015). *Ontology-Based Data Integration from Heterogeneous Urban Systems: A Knowledge Representation Framework for Smart Cities*. Paper presented at the 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015), Cambridge, MA, USA. http://web.mit.edu/cron/project/CUPUM2015/proceedings/Content/infrastructure/240_psyllidis_h.pdf

Psyllidis, A., & Biloria, N. (2014). *OntoPolis: A semantic participatory platform for performance assessment and augmentation of urban environments*. Paper presented at the 10th IEEE International Conference on Intelligent Environments 2014 (IE '14), Shanghai, China. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6910439&isnumber=6910409

Psyllidis, A., Bozzon, A., Bocconi, S., & Bolivar, C. T. (2015a). *Harnessing Heterogeneous Social Data to Explore, Monitor, and Visualize Urban Dynamics*. Paper presented at the 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015), Cambridge, MA, USA. http://web.mit.edu/cron/project/CUPUM2015/proceedings/Content/analytics/239_psyllidis_h.pdf

Psyllidis, A., Bozzon, A., Bocconi, S., & Bolivar, C. T. (2015b). A Platform for Urban Analytics and Semantic Data Integration in City Planning. In G. Celani, D. M. Sperling, & J. M. S. Franco (Eds.), *Computer-Aided Architectural Design Futures – New Technologies and the Future of the Built Environment: 16th International Conference, CAAD Futures 2015, São Paulo, Brazil, July 8-10, 2015. Selected Papers* (pp. 21-36). Berlin Heidelberg: Springer Berlin Heidelberg.

Psyllidis, A., Bozzon, A., Yang, J., & Mesbah, S. (2016 (in press)). *The spatiotemporal footprint of online social networks: Coupling human activity and social interactions in cities*.

Purver, M., & Battersby, S. (2012). *Experimenting with Distant Supervision for Emotion Classification*. Paper presented at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12).

Purves, R., & Hollenstein, L. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*(1), 21–48. doi:10.5311/josis.2010.1.3

Quercia, D., & Sáez-Trumper, D. (2014). Mining Urban Deprivation from Foursquare: Implicit Crowdsourcing of City Land Use. *IEEE Pervasive Computing, 13*(2), 30-36. doi:10.1109/MPRV.2014.31

R, D. C. T. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Radulovic, F., Poveda-Villalón, M., Vila-Suero, D., Rodríguez-Doncel, V., García-Castro, R., & Gómez-Pérez, A. (2015). Guidelines for Linked Data generation and publication: An example in building energy consumption. *Automation in Construction, 57*, 178–187. doi:10.1016/j.autcon.2015.04.002
Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). *Classifying Latent User Attributes in Twitter*. Paper presented at the 2nd international workshop on Search and mining user-generated contents (SMUC '10).
Ratti, C. (2004). Space syntax: some inconsistencies. *Environment and Planning B: Planning and Design, 31*(4), 487–499. doi:DOI:10.1068/b3019
Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design, 33*, 727 – 748. doi:-DOI:10.1068/b32047
Riguelle, F., Thomas, I., & Verhetsel, A. (2007). Measuring urban polycentrism: a European case study and its implications. *Journal of Economic Geography, 7*(2), 193–215. doi:10.1093/jeg/lbl025
Rogerson, P. A. (2010). *Statistical Methods for Geography* (3rd ed.). London: Sage.
Rosenbaum, M. S. (2006). Exploring the Social Supportive Role of Third Places in Consumers' Lives. *Journal of Service Research, 9*(1), 59–72. doi:10.1177/1094670506289530
Roth, C., Kang, S. M., Batty, M., & Barthelemy, M. (2011). Structure of urban movements: polycentric activity and entangled hierarchical flows. *PLoS One, 6*(1), e15923. doi:10.1371/journal.pone.0015923
Sagarra, O., Szell, M., Santi, P., Diaz-Guilera, A., & Ratti, C. (2015). Supersampling and Network Reconstruction of Urban Mobility. *PLoS One, 10*(8), e0134508. doi:10.1371/journal.pone.0134508
Sagl, G., Resch, B., Hawelka, B., & Beinat, E. (2012). *From Social Sensor Data to Collective Human Behaviour Patterns: Analysing and Visualising Spatio-Temporal Dynamics in Urban Environments.* Paper presented at the GI Forum 2012.
Sauermann, L., & Cyganiak, R. (2008). Cool URIs for the Semantic Web. Retrieved from http://www.w3.org/TR/2008/NOTE-cooluris-20081203/
Sayer, A. (1992). *Method in Social Science*. London: Routledge.
Scellato, S., Lambiotte, R., Noulas, A., & Mascolo, C. (2011). *Socio-spatial properties of online location-based social networks*. Paper presented at the International Conference on Weblogs and Social Media (ICWSM 2011), Barcelona.
Scharffe, F., & Euzenat, J. (2011). *Linked data meets ontology matching: Enhancing data linking through ontology alignments*. Paper presented at the 3rd international conference on Knowledge engineering and ontology development (KEOD '11), Paris, France.
Schlapfer, M., Bettencourt, L. M., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., . . . Ratti, C. (2014). The scaling of human interactions with city size. *J R Soc Interface, 11*(98), 20130789. doi:10.1098/rsif.2013.0789
Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). *State of the LOD Cloud 2014*. Retrieved from http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/
Schreiber, G., & Raimond, Y. e. (2014). *RDF 1.1 Primer* Retrieved from https://www.w3.org/TR/rdf11-primer/
Sengstock, C., & Gertz, M. (2012). *Latent Geographic Feature Extraction from Social Media*. Paper presented at the 20th International Conference on Advances in Geographic Information Systems — SIGSPATIAL/GIS '12.
Sevtsuk, A., & Mekonnen, M. (2012). *Urban network analysis: a new toolbox for measuring city form in ArcGIS*. Paper presented at the Symposium on Simulation for Architecture and Urban Design (SimAUD '12), San Diego, CA, USA.
Shelton, T., Poorthuis, A., & Zook, M. (2015). Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information. *Landscape and Urban Planning (Forthcoming)*. doi:10.2139/ssrn.2571757
Silberschatz, A., Galvin, P. B., & Gagne, G. (2009). *Operating System Concepts* (8th ed.). Jefferson City, MO, USA: Wiley.
Silva, T. H., Melo, P. O. S. V. d., Almeida, J. M., Salles, J., & Loureiro, A. A. F. (2013). *A Comparison of Foursquare and Instagram to the Study of City Dynamics and Urban Social Behavior*. Paper presented at the 2nd ACM SIGKDD International Workshop on Urban Computing (UrbComp '13), Chicago, IL.
Simini, F., Gonzalez, M. C., Maritan, A., & Barabasi, A. L. (2012). A universal model for mobility and migration patterns. *Nature, 484*(7392), 96–100. doi:10.1038/nature10856
Sobolevsky, S., Bojic, I., Belyi, A., Sitko, I., Hawelka, B., Arias, J. M., & Ratti, C. (2015). *Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity*. Paper presented at the IEEE BigData Congress 2015, New York, USA.

Sobolevsky, S., Szell, M., Campari, R., Couronne, T., Smoreda, Z., & Ratti, C. (2013). Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS One, 8*(12), e81707. doi:10.1371/journal.pone.0081707

Solecki, W., Seto, K. C., & Marcotullio, P. J. (2013). It's Time for an Urbanization Science. *Environment: Science and Policy for Sustainable Development, 55*(1), 12–17. doi:10.1080/00139157.2013.748387

Song, Y., & Knaap, G.-J. (2004). Measuring Urban Form: Is Portland Winning the War on Sprawl? *Journal of the American Planning Association, 70*(2), 210–225. doi:10.1080/01944360408976371

Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). LinkedGeoData: A Core for a Web of Spatial Open Data. *Semantic Web, 3*(4), 333–354.

Stanford University, C. Protégé open-source ontology editor. Retrieved from http://protege.stanford.edu/

Statistiek, C. B. v. d. (2011). *Wijk- en Buurtkaart 2011*.

Statistiek, C. B. v. d. (2014). *Toelichting Wijk- en Buurtkaart 2011*. The Hague, the Netherlands: Centraal Bureau voor de Statistiek (CBS) Retrieved from http://www.cbs.nl/nl-NL/menu/themas/dossiers/nederland-regionaal/publicaties/geografische-data/archief/2012/2012-wijk-en-buurtkaart-2011-art.htm.

Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems, 54*, 255–265. doi:10.1016/j.compenvurbsys.2015.09.007

Studer, R., Benjamins, R., & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering, 25*(1–2), 161–197. doi:10.1016/S0169-023X(97)00056-6

Stuhr, M., Roman, D., & Norheim, D. (2011). *LODWheel: JavaScript-based Visualization of RDF Data*. Paper presented at the 2nd International Workshop on Consuming Linked Data (COLD 2011) - In Conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011), Bonn, Germany.

Thiemann, C., Theis, F., Grady, D., Brune, R., & Brockmann, D. (2010). The Structure of Borders in a Small World. *PLoS One, 5*(11), e15422. doi:10.1371/journal.pone.0015422.g001

Thünen, J. H. v. (1966). *Von Thünen's Isolated State* (C. M. Wartenberg, Trans. P. Hall Ed.). Oxford, UK: Pergamon Press.

Thurstain-Goodwin, M., & Unwin, D. (2000). Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations. *Transactions in GIS, 4*(4), 305–317.

Toole, J. L., Herrera-Yaque, C., Schneider, C. M., & Gonzalez, M. C. (2015). Coupling human mobility and social ties. *J R Soc Interface, 12*(105), 20141128. doi:10.1098/rsif.2014.1128

Townsend, A. M. (2013). *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. New York: W. W. Norton & Company.

UNFPA. (2007). State of World Population 2007: Unleasing the Potential of Urban Growth. New York: United Nations Population Fund.

Vaccari, A., Calabrese, F., Liu, B., & Ratti, C. (2009). *Towards the SocioScope: An Information System for the Study of Social Dynamics through Digital Traces*. Paper presented at the SIGSPATIAL International Conferenceon Advances in Geographic Information Systems

van Oort, F., Burger, M., & Raspe, O. (2010). On the Economic Foundation of the Urban Network Paradigm: Spatial Integration, Functional Integration and Economic Complementarities within the Dutch Randstad. *Urban Studies, 47*(4), 725–748. doi:10.1177/0042098009352362

Verdone, R., Dardari, D., Mazzini, G., & Conti, A. (2008). *Wireless Sensor and Actuator Networks: Technologies, Analysis and Design*. London, UK: Elsevier.

Villazón-Terrazas, B., Vila-Suero, D., Garijo, D., Vilches-Blázquez, L. M., Poveda-Villalón, M., Mora, J., . . . Gómez-Pérez, A. (2012). *Publishing Linked Data - There is no One-Size-Fits-All Formula*. Paper presented at the European Data Forum, Copenhagen, Denmark.

Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological Guidelines for Publishing Government Linked Data *Linking Government Data* (pp. 27–49). New York, NY, USA: Springer New York.

Vinyals, M., Rodriguez-Aguilar, J. A., & Cerquides, J. (2008). *A Survey on Sensor Networks from a Multi-Agent Perspective.* Paper presented at the 2nd International Workshop on Agent Technology for Sensor Networks (ATSN-08).

Wachsmann, L. (2008). OWLPropViz. Retrieved from http://protegewiki.stanford.edu/wiki/OWLPropViz

Wang, Y., Kang, C., Bettencourt, L. M. A., Liu, Y., & Andris, C. (2015). Linked Activity Spaces: Embedding Social Networks in Urban Space. In M. Helbich, J. J. Arsanjani, & M. Leitner (Eds.), *Computational Approaches for Urban Environments* (pp. 313–336): Springer International Publishing.

Weber, A. (1909). *Theory of the Location of Industries*. Chicago: University of Chicago Press.

Weigend, A. (2009). The Social Data Revolution(s). Retrieved from https://hbr.org/2009/05/the-social-data-revolution.html

Wilson, A. G. (1967). A Statistical Theory of Spatial Distribution Models. *Transportation Research, 1*(3), 253–269.

Wilson, A. G. (1970). *Entropy in Urban and Regional Modeling*. London: Pion Press.

Wilson, A. G. (1975). Some New Forms of Spatial Interaction Model: A Review. *Transportation Research, 9*(2-3), 167–179. doi:10.1016/0041-1647(75)90054-4

Yang, J., Hauff, C., Houben, G.-J., & Bolivar, C. T. (2016). Diversity in Urban Social Media Analytics. In A. Bozzon, P. Cudré-Mauroux, & C. Pautasso (Eds.), *Proceedings of the 16th International Conferrence on Web Engineering (ICWE 2016)* (pp. 335–353). Lugano, Switzerland: Springer International Publishing.

Yuan, J., Zheng, Y., & Xie, X. (2012). *Discovering regions of different functions in a city using human mobility and POIs*. Paper presented at the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China.

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., & Schmitt, G. (2016). Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data. *PLoS One, 11*(2), e0149222. doi:10.1371/journal.pone.0149222

Zhong, C., Schlapfer, M., Muller Arisona, S., Batty, M., Ratti, C., & Schmitt, G. (2015). Revealing centrality in the spatial structure of cities from human activity patterns. *Urban Studies*, 1–19. doi:10.1177/0042098015601599

Zhu, Y. (2014). *Spatiotemporal Learning and Geo-visualization Methods for Constructing Activity-Travel Patterns from Transit Card Transaction Data*. (PhD), Massachusetts Institute of Technology, Boston, MA. Retrieved from http://hdl.handle.net/1721.1/93807

Zikopoulos, P., Eaton, C., Roos, D. d., Deutsch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw-Hill Osborne Media.

TOC

Appendix A ROUTE Ontology (Chapter 3)

The following code represents the ROUTE Ontology (Chapter 3), expressed in RDF/XML syntax.

```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF [
  <!ENTITY schema "http://schema.org/#" >
  <!ENTITY foaf "http://xmlns.com/foaf/0.1" >
  <!ENTITY terms "http://purl.org/dc/terms/" >
  <!ENTITY dct "http://purl.org/dc/terms/#" >
  <!ENTITY vann "http://purl.org/vocab/vann/" >
  <!ENTITY time "http://www.w3.org/2006/time#" >
  <!ENTITY dbpedia-owl "http://dbpedia.org/ontology#" >
  <!ENTITY ns "http://creativecommons.org/ns#" >
  <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
  <!ENTITY dc "http://purl.org/dc/elements/1.1" >
  <!ENTITY dc "http://purl.org/dc/elements/1.1/" >
  <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
  <!ENTITY owl2xml "http://www.w3.org/2006/12/owl2-xml#" >
  <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
  <!ENTITY wgs84_pos "http://www.w3.org/2003/01/geo/wgs84_pos#" >
  <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <!ENTITY otn "http://www.pms.ifi.lmu.de/rewerse-wga1/otn/OTN.owl" >
]>

<rdf:RDF xmlns="http://www.w3.org/2002/07/owl#"
   xml:base="http://www.w3.org/2002/07/owl"
   xmlns:dc="&dc;/"
   xmlns:ns="http://creativecommons.org/ns#"
   xmlns:dbpedia-owl="http://dbpedia.org/ontology#"
   xmlns:wgs84_pos="http://www.w3.org/2003/01/geo/wgs84_pos#"
   xmlns:foaf="http://xmlns.com/foaf/0.1"
   xmlns:terms="http://purl.org/dc/terms/"
   xmlns:vann="http://purl.org/vocab/vann/"
   xmlns:schema="http://schema.org/#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:time="http://www.w3.org/2006/time#"
   xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
   xmlns:dct="&terms;#"
```

```
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:otn="http://www.pms.ifi.lmu.de/rewerse-wga1/otn/OTN.owl">
  <Ontology rdf:about="https://route-owl.github.io/ontology">
    <rdfs:comment></rdfs:comment>
    <terms:creator>Achilleas Psyllidis</terms:creator>
    <terms:title>ROUTE - Route Ontology of Urban Transportation Entities</
terms:title>
    <dct:subject xml:lang="en">This ROUTE ontology describes public urban
transportation routes. It also describes concepts pertinent to trip services, pickup and
drop-off types, time intervals, frequency, geographical information about stops, among
other related concepts.</dct:subject>
    <ns:license>https://creativecommons.org/licenses/by/4.0/</ns:license>
    <vann:preferredNamespacePrefix>route</vann:preferredNamespacePrefix>
    <dc:identifier>https://route-owl.github.io/ontology</dc:identifier>
    <dct:subject xml:lang="gr">Η οντολογία ROUTE περιγράφει τις διαδρομές των
αστικών συγκοινωνιών. Παράλληλα, περιγράφει έννοιες σχετικές με τις υπηρεσίες
διαδρομών, τους τύπους επιβίβασης και αποβίβασης, τα χρονικά διαστήματα,
τη συχνότητα, γεωγραφικές πληροφορίες σχετικές με τις στάσεις, καθώς και
παραπλήσιες έννοιες.</dct:subject>
    <imports rdf:resource="http://vocab.gtfs.org/terms#"/>
    <imports rdf:resource="http://www.w3.org/2006/time"/>
  </Ontology>

  <!-- / Annotation properties  // -->


  <!-- http://purl.org/dc/elements/1.1/identifier -->

  <AnnotationProperty rdf:about="&dc;/identifier"/>

  <!-- http://purl.org/dc/terms/description -->

  <AnnotationProperty rdf:about="&terms;description"/>


  <!-- http://purl.org/dc/terms/#subject -->

  <AnnotationProperty rdf:about="&terms;#subject"/>
```

```
<!-- // Object Properties // -->

  <!-- https://route-owl.github.io/ontology#belongsTo -->

  <ObjectProperty rdf:about="https://route-owl.github.io/ontology#belongsTo">
    <inverseOf rdf:resource="https://route-owl.github.io/ontology#stopsAt"/>
  </ObjectProperty>

  <!-- https://route-owl.github.io/ontology#endsOn -->

  <ObjectProperty rdf:about="https://route-owl.github.io/ontology#endsOn"/>


  <!-- https://route-owl.github.io/ontology#hasDescription -->

  <ObjectProperty rdf:about="https://route-owl.github.io/
ontology#hasDescription"/>


  <!-- https://route-owl.github.io/ontology#hasEndPoint -->

  <ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasEndPoint"/>


  <!-- https://route-owl.github.io/ontology#hasHeadSign -->

  <ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasHeadSign"/>


  <!-- https://route-owl.github.io/ontology#hasID -->

  <ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasID"/>


  <!-- https://route-owl.github.io/ontology#hasLocation -->

  <ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasLocation">
    <rdfs:label>hasLocation</rdfs:label>
    <rdfs:comment xml:lang="en">The relation between something and the point, or
other geometrical thing in space, where it is.  For example, the relationship between a
radio tower and a Point with a given lat and long. Or a relationship between a park and
its outline as a closed arc of points, or a road and its location as a arc (a sequence of
```

points). Clearly in practice there will be limit to the accuracy of any such statement, but one would expect an accuracy appropriate for the size of the object and uses such as mapping .
</rdfs:comment>
</ObjectProperty>

<!-- https://route-owl.github.io/ontology#hasOrder -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasOrder"/>

<!-- https://route-owl.github.io/ontology#hasPoint -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasPoint"/>

<!-- https://route-owl.github.io/ontology#hasSchedule -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasSchedule"/>

<!-- https://route-owl.github.io/ontology#hasStartPoint -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasStartPoint"/>

<!-- https://route-owl.github.io/ontology#hasType -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasType"/>

<!-- https://route-owl.github.io/ontology#hasURL -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#hasURL"/>

<!-- https://route-owl.github.io/ontology#inDirection -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#inDirection"/>

<!-- https://route-owl.github.io/ontology#isLocatedIn -->

```
<ObjectProperty rdf:about="https://route-owl.github.io/ontology#isLocatedIn"/>

<!-- https://route-owl.github.io/ontology#runBy -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#runBy"/>

<!-- https://route-owl.github.io/ontology#startsOn -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#startsOn"/>

<!-- https://route-owl.github.io/ontology#stopsAt -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#stopsAt"/>

<!-- https://route-owl.github.io/ontology#worksOn -->

<ObjectProperty rdf:about="https://route-owl.github.io/ontology#worksOn"/>

<!-- // Data properties // -->

 <!-- http://vocab.gtfs.org/terms#headsign -->

<rdf:Description rdf:about="http://vocab.gtfs.org/terms#headsign">
  <rdfs:domain rdf:resource="http://vocab.gtfs.org/terms#Trip"/>
</rdf:Description>

<!-- http://www.w3.org/2003/01/geo/wgs84_pos#lat -->

<DatatypeProperty rdf:about="&wgs84_pos;lat">
  <rdfs:range rdf:resource="&xsd;double"/>
</DatatypeProperty>

<!-- http://www.w3.org/2003/01/geo/wgs84_pos#long -->

<DatatypeProperty rdf:about="&wgs84_pos;long">
  <rdfs:range rdf:resource="&xsd;double"/>
```

```
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#ID -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#ID">
  <rdfs:domain rdf:resource="http://vocab.gtfs.org/terms#Trip"/>
  <rdfs:range rdf:resource="&xsd;string"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#URL -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#URL">
  <rdfs:range rdf:resource="&xsd;string"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#code -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#code">
  <rdfs:range rdf:resource="&xsd;string"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#description -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#description"/>

<!-- https://route-owl.github.io/ontology#language -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#language">
  <rdfs:range rdf:resource="&xsd;language"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#longName -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#longName">
  <rdfs:range rdf:resource="&xsd;string"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#name -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#name">
  <rdfs:range rdf:resource="&xsd;string"/>
</DatatypeProperty>
```

```
<!-- https://route-owl.github.io/ontology#order -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#order">
  <rdfs:range rdf:resource="&xsd;integer"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#phone -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#phone">
  <rdfs:range rdf:resource="&xsd;string"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#shortName -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#shortName">
  <rdfs:range rdf:resource="&xsd;string"/>
</DatatypeProperty>

<!-- https://route-owl.github.io/ontology#timezone -->

<DatatypeProperty rdf:about="https://route-owl.github.io/ontology#timezone">
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger"/>
</DatatypeProperty>


<!-- // Classes // -->

<!-- http://dbpedia.org/ontology/City -->

<Class rdf:about="http://dbpedia.org/ontology/City">
  <rdfs:subClassOf rdf:resource="http://schema.org/AdministrativeArea"/>
  <rdfs:comment xml:lang="en">A relatively large and permanent settlement,
particularly a large urban settlement</rdfs:comment>
  <rdfs:comment xml:lang="fr">Ville</rdfs:comment>
  <rdfs:comment xml:lang="ga">Cathair</rdfs:comment>
  <rdfs:comment xml:lang="gr">Πόλη</rdfs:comment>
</Class>

<!-- http://schema.org/AdministrativeArea -->

<Class rdf:about="http://schema.org/AdministrativeArea">
  <rdfs:subClassOf rdf:resource="http://schema.org/Place"/>
```

```
    <rdfs:comment xml:lang="en">A geographical region under the jurisdiction of a
particular government.</rdfs:comment>
    <rdfs:comment xml:lang="fr">Zone administrative</rdfs:comment>
    <rdfs:comment xml:lang="ga">Ceantair Riaracháin</rdfs:comment>
    <rdfs:comment xml:lang="gr">Διοικητική Περιοχή</rdfs:comment>
  </Class>

  <!-- http://schema.org/Place -->

  <Class rdf:about="http://schema.org/Place">
    <terms:description xml:lang="en">A Place might have definite or indefinite
boundaries. Geographic spaces can be a position, line, area, or volume.</
terms:description>
    <rdfs:comment xml:lang="en">A geographic or virtual part of space.</
rdfs:comment>
    <rdfs:comment xml:lang="fr">Lieu</rdfs:comment>
    <rdfs:comment xml:lang="ga">Áit</rdfs:comment>
    <rdfs:comment xml:lang="gr">Τόπος</rdfs:comment>
  </Class>

  <!-- http://vocab.gtfs.org/terms#Agency -->

  <rdf:Description rdf:about="http://vocab.gtfs.org/terms#Agency">
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/
ontology#timezone"/>
        <someValuesFrom rdf:resource="&xsd;nonNegativeInteger"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#URL"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#phone"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
```

```
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#name"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#language"/>
        <someValuesFrom rdf:resource="&xsd;language"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="http://vocab.gtfs.org/terms#hasService"/>
        <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#Service"/>
      </Restriction>
    </rdfs:subClassOf>
  </rdf:Description>

  <!-- http://vocab.gtfs.org/terms#Route -->

  <rdf:Description rdf:about="http://vocab.gtfs.org/terms#Route">
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#hasType"/>
        <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#RouteType"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/
ontology#longName"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/
ontology#shortName"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
```

```
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/ontology#ID"/>
          <someValuesFrom rdf:resource="&xsd;string"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="http://vocab.gtfs.org/terms#color"/>
          <someValuesFrom rdf:resource="&xsd;string"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/
ontology#description"/>
          <someValuesFrom rdf:resource="&xsd;string"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:comment xml:lang="fr">Route</rdfs:comment>
      <rdfs:comment xml:lang="ga">Chúrsa</rdfs:comment>
      <rdfs:comment xml:lang="gr">Πορεία</rdfs:comment>
    </rdf:Description>


    <!-- http://vocab.gtfs.org/terms#Service -->

    <rdf:Description rdf:about="http://vocab.gtfs.org/terms#Service">
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="http://vocab.gtfs.org/terms#monday"/>
          <someValuesFrom rdf:resource="&xsd;boolean"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/ontology#endsOn"/>
          <someValuesFrom rdf:resource="&time;DateTimeDescription"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/ontology#runBy"/>
          <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#Agency"/>
```

```
    </Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <Restriction>
      <onProperty rdf:resource="http://vocab.gtfs.org/terms#sunday"/>
      <someValuesFrom rdf:resource="&xsd;boolean"/>
    </Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <Restriction>
      <onProperty rdf:resource="http://vocab.gtfs.org/terms#tuesday"/>
      <someValuesFrom rdf:resource="&xsd;boolean"/>
    </Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <Restriction>
      <onProperty rdf:resource="http://vocab.gtfs.org/terms#wednesday"/>
      <someValuesFrom rdf:resource="&xsd;boolean"/>
    </Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <Restriction>
      <onProperty rdf:resource="http://vocab.gtfs.org/terms#friday"/>
      <someValuesFrom rdf:resource="&xsd;boolean"/>
    </Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <Restriction>
      <onProperty rdf:resource="http://vocab.gtfs.org/terms#thursday"/>
      <someValuesFrom rdf:resource="&xsd;boolean"/>
    </Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <Restriction>
      <onProperty rdf:resource="http://vocab.gtfs.org/terms#saturday"/>
      <someValuesFrom rdf:resource="&xsd;boolean"/>
    </Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <Restriction>
      <onProperty rdf:resource="https://route-owl.github.io/ontology#worksOn"/>
      <someValuesFrom rdf:resource="&time;DayOfWeek"/>
    </Restriction>
```

```
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#startsOn"/>
        <someValuesFrom rdf:resource="&time;DateTimeDescription"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#ID"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:comment xml:lang="fr">Service</rdfs:comment>
    <rdfs:comment xml:lang="ga">Seirbhís</rdfs:comment>
    <rdfs:comment xml:lang="gr">Υπηρεσία</rdfs:comment>
  </rdf:Description>


  <!-- http://vocab.gtfs.org/terms#Stop -->

  <rdf:Description rdf:about="http://vocab.gtfs.org/terms#Stop">
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#ID"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="&wgs84_pos;lat"/>
        <someValuesFrom rdf:resource="&xsd;double"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#code"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
```

```
        <onProperty rdf:resource="https://route-owl.github.io/
ontology#description"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="&wgs84_pos;long"/>
        <someValuesFrom rdf:resource="&xsd;double"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/ontology#name"/>
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:comment xml:lang="fr">Arrêt</rdfs:comment>
    <rdfs:comment xml:lang="ga">Stop</rdfs:comment>
    <rdfs:comment xml:lang="gr">Στάση</rdfs:comment>
  </rdf:Description>


  <!-- http://vocab.gtfs.org/terms#StopTime -->

  <rdf:Description rdf:about="http://vocab.gtfs.org/terms#StopTime">
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="http://vocab.gtfs.org/terms#hasPickupType"/>
        <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#PickupType"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="http://vocab.gtfs.org/terms#isStop"/>
        <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#Stop"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="http://vocab.gtfs.org/terms#hasDropOffType"/>
        <someValuesFrom rdf:resource="http://vocab.gtfs.org/
terms#DropOffType"/>
```

```
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/ontology#ID"/>
          <someValuesFrom rdf:resource="&xsd;string"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/ontology#order"/>
          <someValuesFrom rdf:resource="&xsd;integer"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/
ontology#belongsTo"/>
          <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#Trip"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:comment xml:lang="fr">Horaire</rdfs:comment>
      <rdfs:comment xml:lang="ga">Am Stad</rdfs:comment>
      <rdfs:comment xml:lang="gr">Ωράριο Στάσεων</rdfs:comment>
    </rdf:Description>


    <!-- http://vocab.gtfs.org/terms#Trip -->

    <rdf:Description rdf:about="http://vocab.gtfs.org/terms#Trip">
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/ontology#stopsAt"/>
          <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#StopTime"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="http://vocab.gtfs.org/terms#hasPickupType"/>
          <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#PickupType"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
```

```
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/
ontology#hasEndPoint"/>
          <someValuesFrom rdf:resource="https://route-owl.github.io/
ontology#EndPoint"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/ontology#ID"/>
          <someValuesFrom rdf:resource="&xsd;string"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="http://vocab.gtfs.org/terms#hasRoute"/>
          <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#Route"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="http://vocab.gtfs.org/terms#direction"/>
          <someValuesFrom rdf:resource="&xsd;boolean"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/
ontology#hasStartPoint"/>
          <someValuesFrom rdf:resource="https://route-owl.github.io/
ontology#StartPoint"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="http://vocab.gtfs.org/terms#hasDropOffType"/>
          <someValuesFrom rdf:resource="http://vocab.gtfs.org/
terms#DropOffType"/>
        </Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf>
        <Restriction>
          <onProperty rdf:resource="http://vocab.gtfs.org/terms#headsign"/>
```

```
        <someValuesFrom rdf:resource="&xsd;string"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="http://vocab.gtfs.org/terms#hasService"/>
        <someValuesFrom rdf:resource="http://vocab.gtfs.org/terms#Service"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:comment xml:lang="fr">Voyage</rdfs:comment>
    <rdfs:comment xml:lang="ga">Turas</rdfs:comment>
    <rdfs:comment xml:lang="gr">Διαδρομή</rdfs:comment>
  </rdf:Description>


  <!-- http://www.w3.org/2003/01/geo/wgs84_pos#Point -->

  <Class rdf:about="&wgs84_pos;Point">
    <rdfs:label xml:lang="en">Point</rdfs:label>
    <rdfs:subClassOf rdf:resource="&wgs84_pos;SpatialThing"/>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="&wgs84_pos;lat"/>
        <someValuesFrom rdf:resource="&xsd;double"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="&wgs84_pos;long"/>
        <someValuesFrom rdf:resource="&xsd;double"/>
      </Restriction>
    </rdfs:subClassOf>
    <rdfs:comment xml:lang="en"> Uniquely identified by lat/long/alt. i.e.

spaciallyIntersects(P1, P2) :- lat(P1, LAT), long(P1, LONG), alt(P1, ALT),
 lat(P2, LAT), long(P2, LONG), alt(P2, ALT).

sameThing(P1, P2) :- type(P1, Point), type(P2, Point), spaciallyIntersects(P1, P2).
 </rdfs:comment>
    <rdfs:comment xml:lang="en">A point, typically described using a coordinate
system relative to Earth, such as WGS84.</rdfs:comment>
    <rdfs:comment xml:lang="fr">Position</rdfs:comment>
    <rdfs:comment xml:lang="ga">Paointe</rdfs:comment>
```

```
    <rdfs:comment xml:lang="gr">Σημείο / Θέση</rdfs:comment>
  </Class>


  <!-- http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing -->

  <Class rdf:about="&wgs84_pos;SpatialThing">
    <rdfs:label xml:lang="en">SpatialThing</rdfs:label>
    <rdfs:comment xml:lang="en">Anything with spatial extent, i.e. size, shape, or
position.
 e.g. people, places, bowling balls, as well as abstract areas like cubes.</rdfs:comment>
  </Class>


  <!-- https://route-owl.github.io/ontology#EndPoint -->

  <Class rdf:about="https://route-owl.github.io/ontology#EndPoint">
    <rdfs:subClassOf>
      <Restriction>
        <onProperty rdf:resource="https://route-owl.github.io/
ontology#isLocatedIn"/>
        <someValuesFrom rdf:resource="http://schema.org/AdministrativeArea"/>
      </Restriction>
    </rdfs:subClassOf>
  </Class>


  <!-- https://route-owl.github.io/ontology#OASAFeed -->

  <Class rdf:about="https://route-owl.github.io/ontology#OASAFeed">
    <rdfs:label>OASA Feed</rdfs:label>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/ns/dcat#Dataset"/>
    <rdfs:comment xml:lang="en">Athens Urban Transport Organisation data. The
routes of the public urban transportation system for the city of Athens are contained.
The data include the stops and routes for bus, trolley, subway, tram and commuter
rail.</rdfs:comment>
  </Class>


  <!-- https://route-owl.github.io/ontology#StartPoint -->

  <Class rdf:about="https://route-owl.github.io/ontology#StartPoint">
    <rdfs:subClassOf>
```

```
        <Restriction>
          <onProperty rdf:resource="https://route-owl.github.io/
ontology#isLocatedIn"/>
          <someValuesFrom rdf:resource="http://schema.org/AdministrativeArea"/>
        </Restriction>
      </rdfs:subClassOf>
    </Class>
</rdf:RDF>
```

Appendix B DCAT & VoID Documentation (Chapter 3)

The following respectively comprise the DCAT and VoID documentation of the ROUTE Ontology (Chapter 3) and the resulting linked dataset that is mapped to it.

DCAT Documentation

```
@prefix os:   <http://a9.com/-/spec/opensearch/1.1/> .
@prefix dct:  <http://purl.org/dc/terms/> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix api:  <http://purl.org/linked-data/api/vocab#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xhv:  <http://www.w3.org/1999/xhtml/vocab#> .

<http://route-owl.github.io/>
    a               dct:Dataset;
    dct:license    <https://creativecommons.org/licenses/by/4.0/>;
        dct:source     "This ROUTE ontology describes public urban transportation
routes. It also describes concepts pertinent to trip services, pickup and drop-off types,
time intervals, frequency, geographical information about stops, among other related
concepts.";
    dct:publisher           "Achilleas Psyllidis";
        dct:language    <http://id.loc.gov/vocabulary/iso639--1/en>,
    dct:accrualPeriodicity  <http://purl.org/linked--data/sdmx/2009/code#freq-A>;
```

VoID Documentation

```
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>.
@prefix foaf:<http://xmlns.com/foaf/0.1>.
@prefix dcterms:<http://purl.org/dc/terms/#>.
@prefix void:<http://rdfs.org/ns/void#>.
@prefix xsd:<http://www.w3.org/2001/XMLSchema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#> .
```

```
##dataset

<http://route-owl.github.io/>
    rdf:type void:Dataset;
    foaf:homepage<http://route-owl.github.io/>
    dcterms:title "ROUTE - Route Ontology of Urban Transportation Entities"
    dcterms:description "This ROUTE ontology describes public urban transportation routes. It also describes concepts pertinent to trip services, pickup and drop-off types, time intervals, frequency, geographical information about stops, among other related concepts."
    void:sparqlEndpoint <http://route-owl.github.io/sparql>;
    void:uriSpace "http://route-owl.github.io/resource/";
    dcterms:source "This ROUTE ontology describes public urban transportation routes. It also describes concepts pertinent to trip services, pickup and drop-off types, time intervals, frequency, geographical information about stops, among other related concepts.";
    dcterms:created "2015-06-11"^^xsd:date;
    dcterms:licence <https://creativecommons.org/licenses/by/4.0/>;
    dcterms:subject <http://route-owl.github.io/resource/agency/agency>;
    void:triples 4593531;
    void:entities 271;
    void:classes 51;
    void:properties 166;
    void:distinctSubjects ?;
    void:distinctObjects ?;

    :DBpedia rdf:type void:Dataset;
     foaf:homepage <http://dbpedia.org/>;
     dcterms:title "Athen Mass Transit System";
     dcterms:description "DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. We hope that this work will make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.";
     void:exampleResource <http://dbpedia.org/resource/Athens_Mass_Transit_System>.

:DBpedia rdf:type void:Dataset;
     foaf:homepage <http://dbpedia.org/>;
     dcterms:title "AdministrativeArea";
```

dcterms:description “DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. We hope that this work will make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.”;
void:exampleResource <http://dbpedia.org/resource/Glyfada>.

:DBpedia_ROUTE rdf:type void:Linkset;
void:linkPredicate owl:sameAs;
void:target <http://route-owl.github.io/>;
void:target :DBpedia

:GeoNames rdf:type void:Dataset;
foaf:homepage <http://www.geonames.org>;
dcterms:title “lat”;
dcterms:description “The GeoNames geographical database covers all countries and contains over eight million placenames that are available for download free of charge.”;
void:exampleResource <http://sws.geonames.org/379838/>.

:GeoNames rdf:type void:Dataset;
foaf:homepage <http://www.geonames.org>;
dcterms:title “long”;
dcterms:description “The GeoNames geographical database covers all countries and contains over eight million placenames that are available for download free of charge.”;
void:exampleResource <http://sws.geonames.org/237275/>.

:GeoNames_ROUTE rdf:type void:Linkset;
void:linkPredicate owl:sameAs;
void:target <http://route-owl.github.io/>;
void:target :GeoNames

TOC

Appendix C Data exploration and visualization – *SocialGlass* frontend (Chapter 6)

The following images illustrate the various data visualization and filtering options of the *SocialGlass* frontend for data coming from different geo-enabled social media. Overall, they comprise the *data exploration* and *visualization* component of the web-based system (see Sect. 6.4.1).

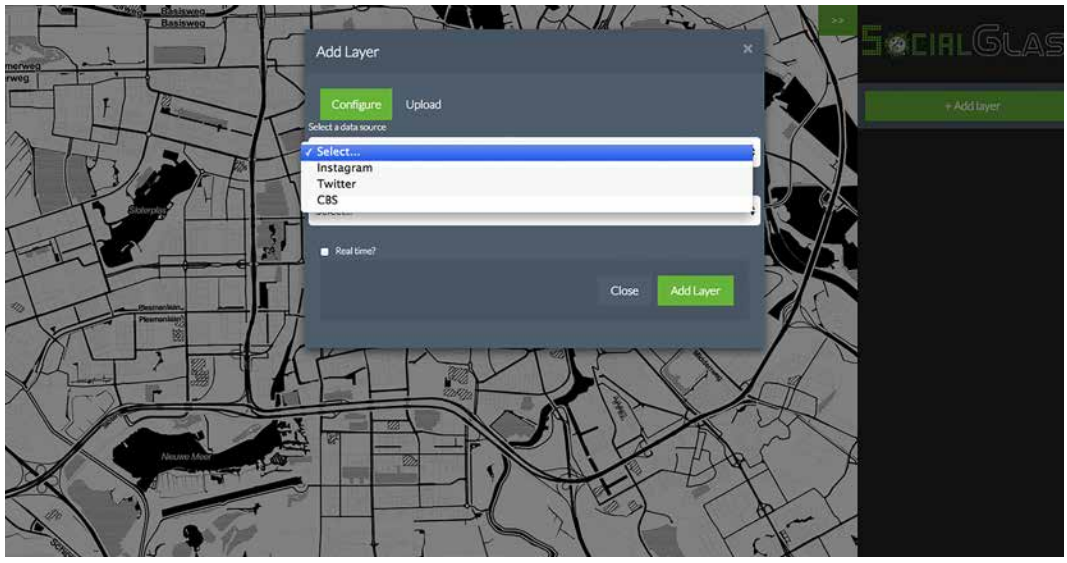


FIGURE 32 Selection of data sources. Sina Weibo is an additional source, in the case of Chinese cities.

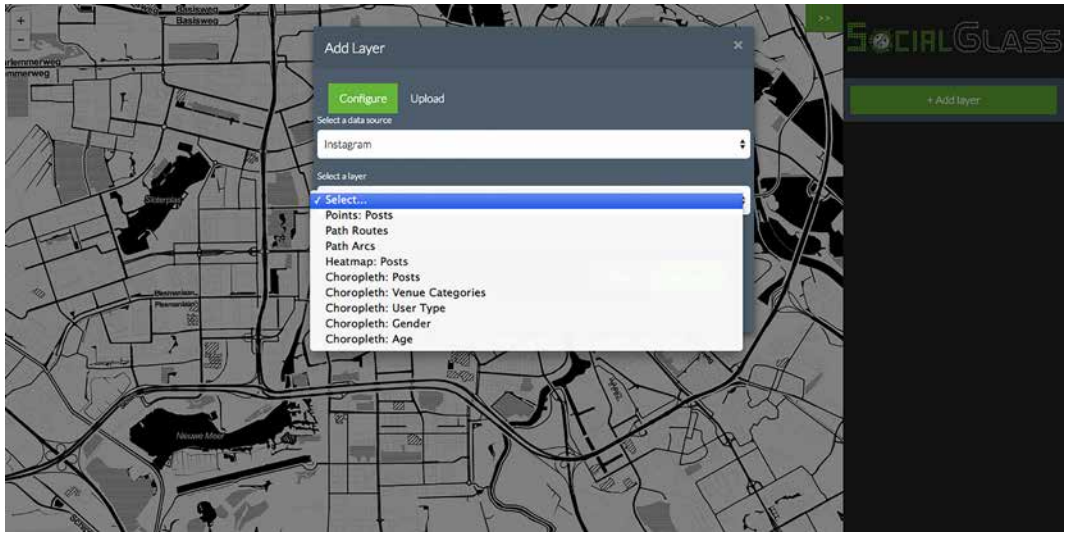


FIGURE 33 Types of data visualization. Each type represents a separate layers, on top of the map-based user interface.

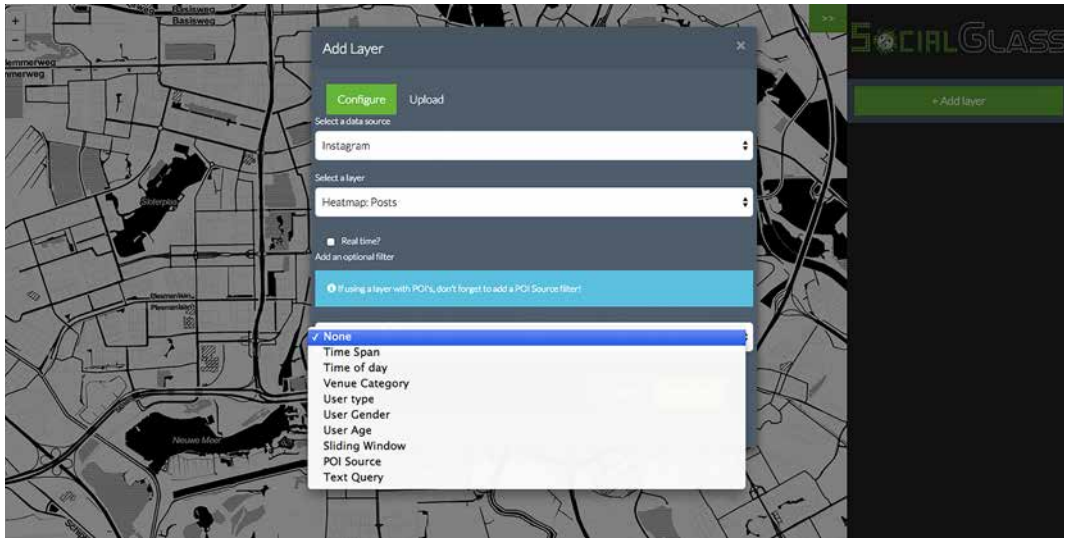


FIGURE 34 Data filters.

FIGURE 35 Dynamic point clusters.

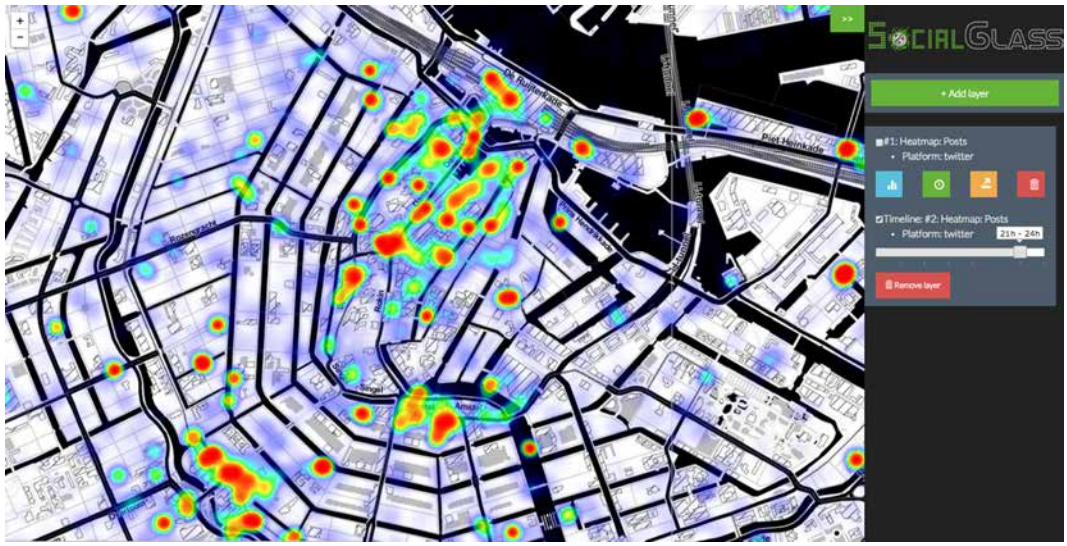


FIGURE 36 Activity heat maps. Time sliders (right pane) enable the exploration of changes in the activity patterns in the course of a day.

FIGURE 37 Origin-Destination (OD) paths. Larger edge thickness and color density illustrate larger flow volumes.

TOC

FIGURE 38 Individual trajectories (path routes).

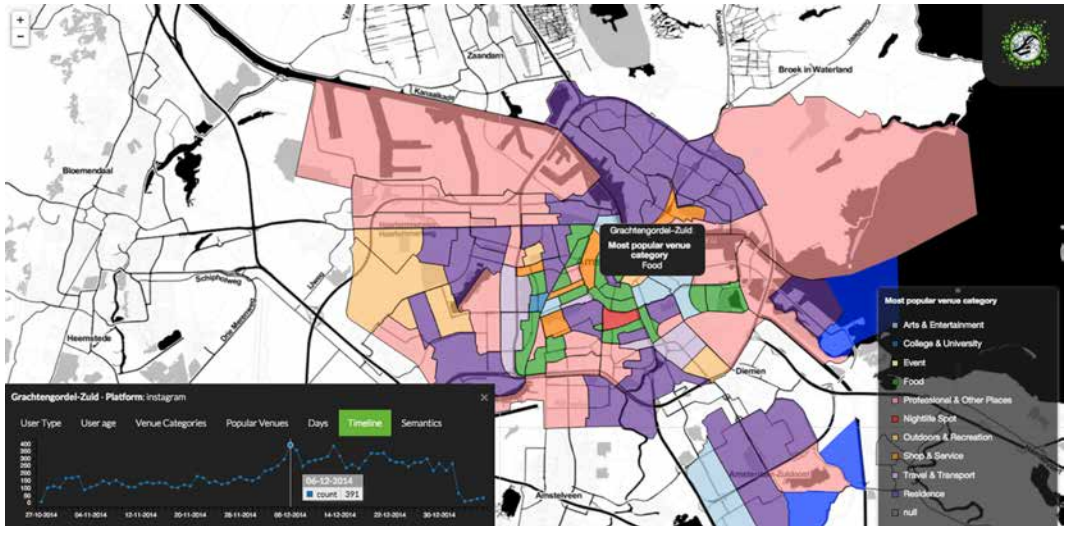


FIGURE 39 Choropleth maps with additional information on the daily distribution of social activity.

TOC

Appendix D Visual exploratory analysis of spatiotemporal activity using *SocialGlass* (Chapter 6)

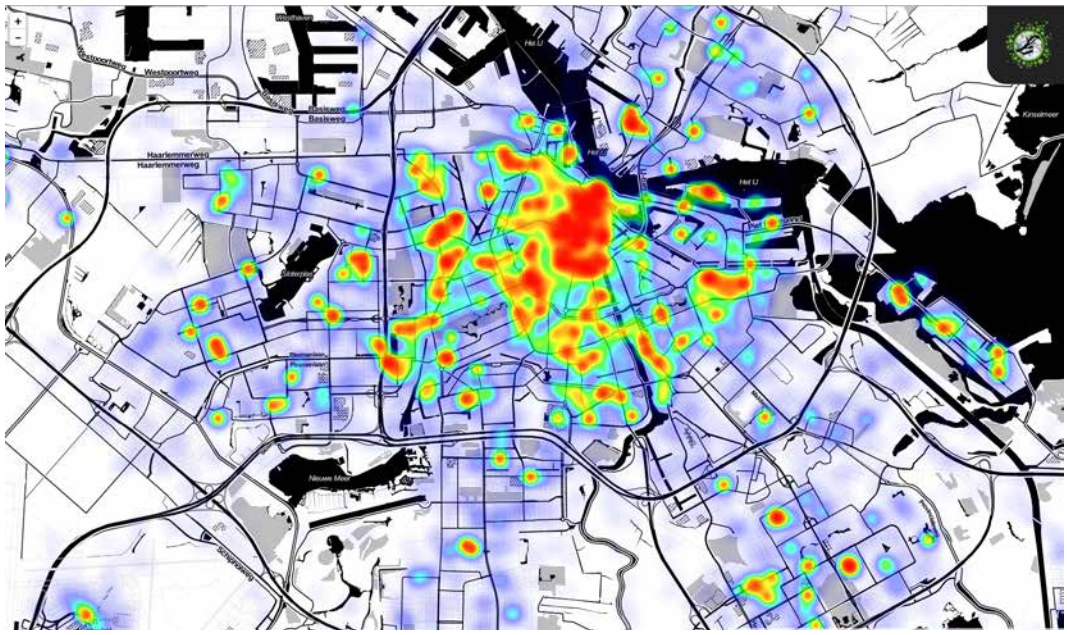


FIGURE 40 Heat map of residents' activity during the ALF event (27/11/2014 — 18/11/2015), as inferred from Twitter.

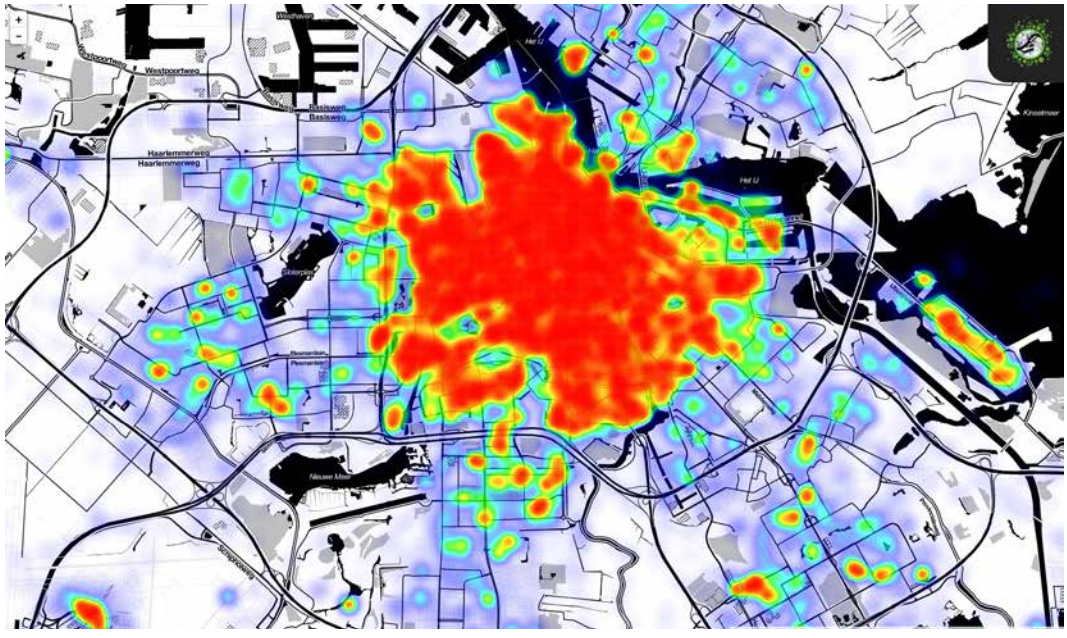


FIGURE 41 Heat map of residents' activity during the ALF event (27/11/2014 — 18/11/2014), as inferred from Instagram.

FIGURE 42 Heat map of residents' activity before the ALF event (13/11/2014 — 26/11/2014), as inferred from Twitter.

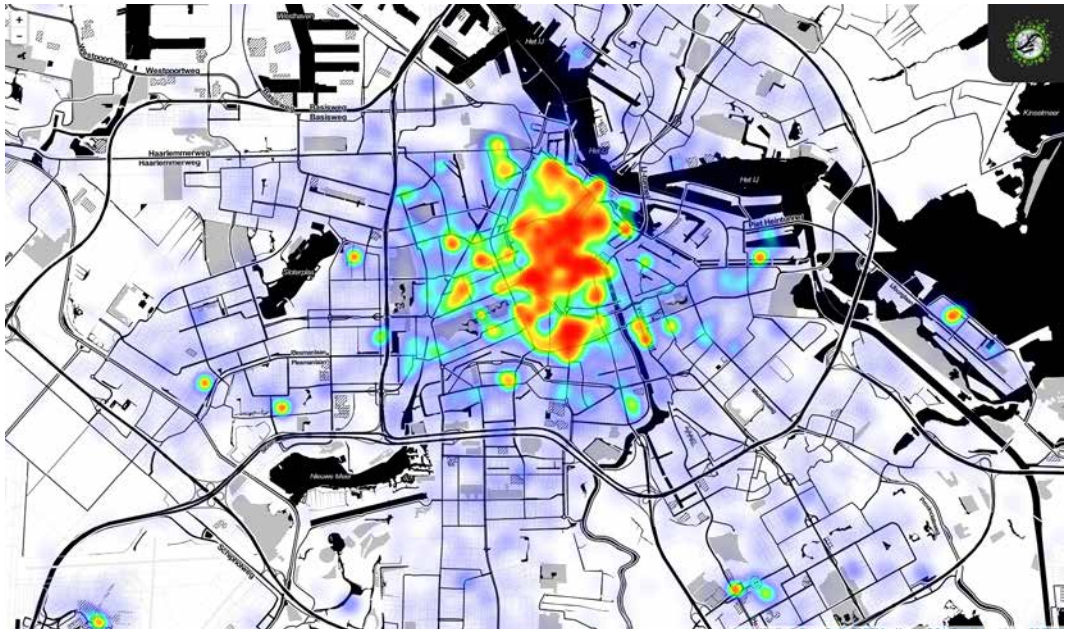


FIGURE 43 Heat map of residents' activity before the ALF event (13/11/2014 — 26/11/2014), as inferred from Instagram.

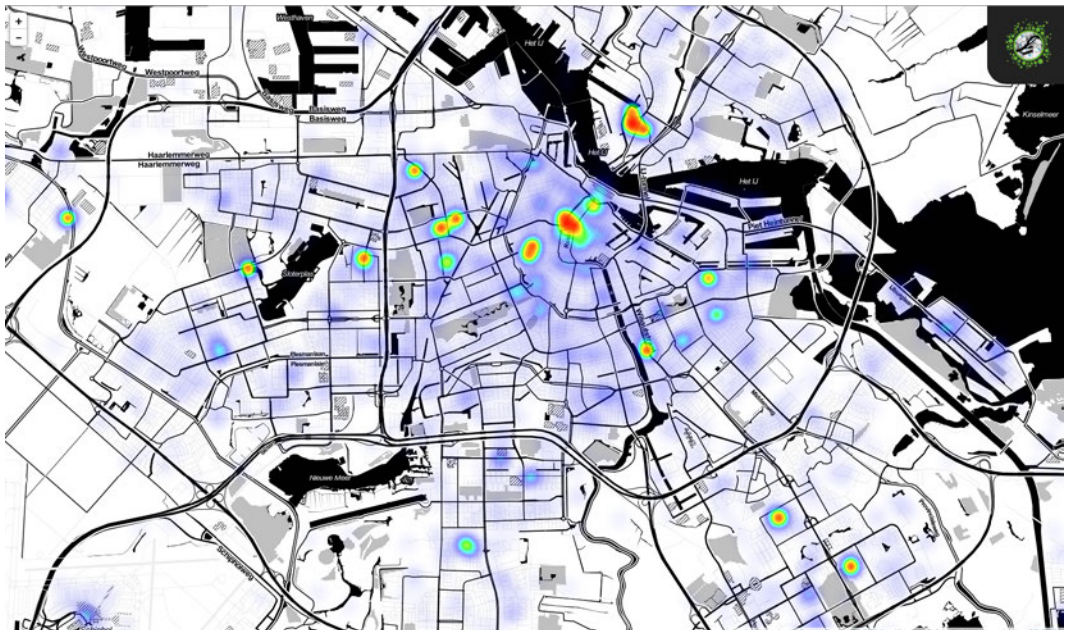


FIGURE 44 Heat map of residents' activity after the ALF event (19/01/2015 — 31/01/2015), as inferred from Twitter.

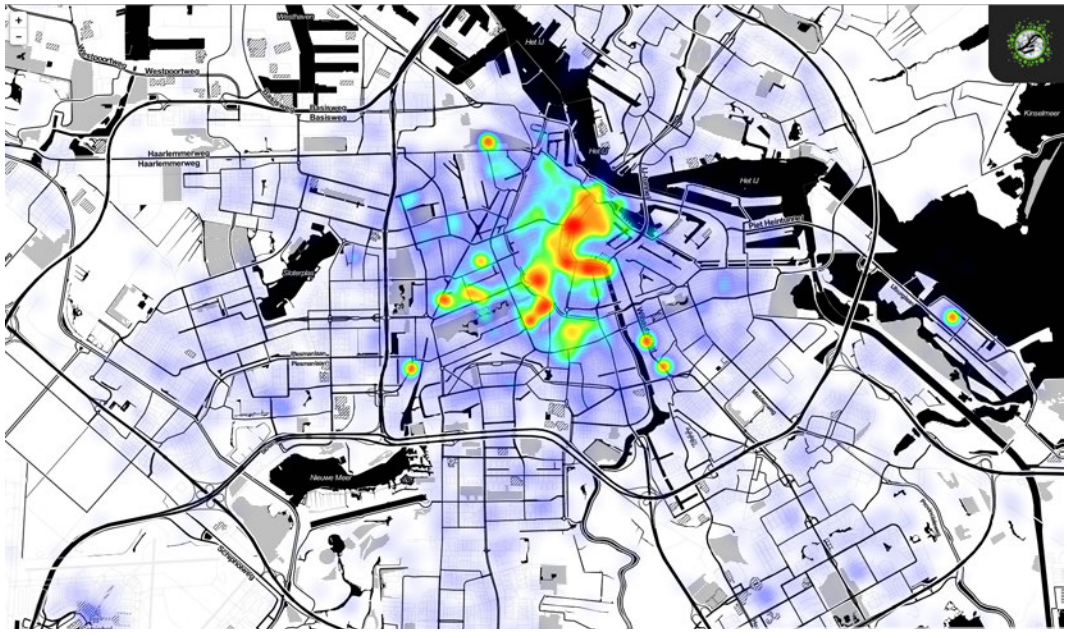


FIGURE 45 Heat map of residents' activity after the ALF event (19/01/2015 — 31/01/2015), as inferred from Instagram.

TOC

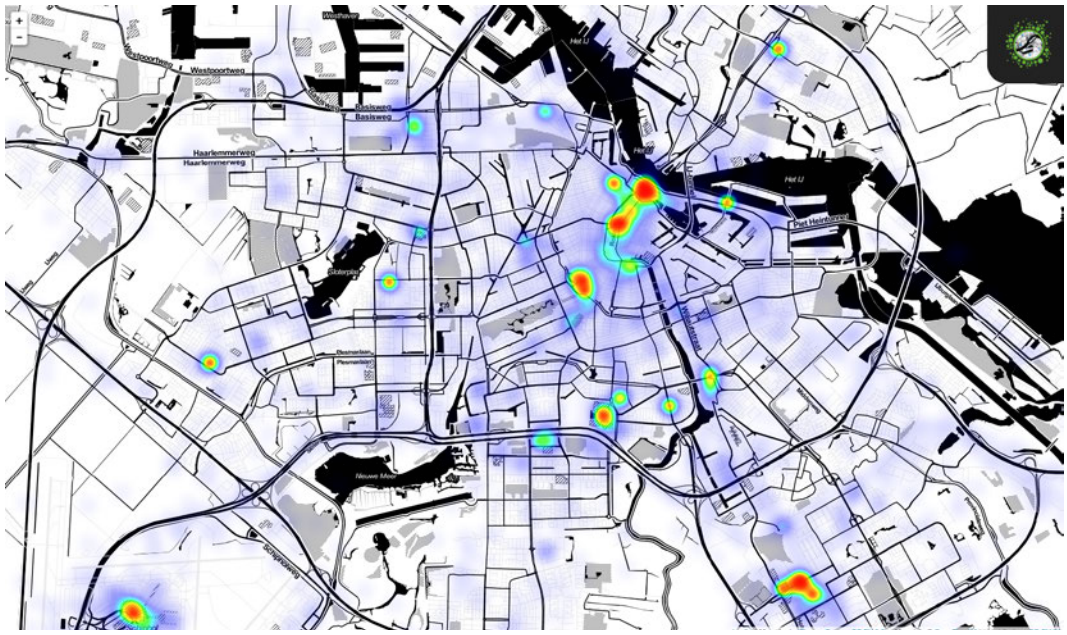


FIGURE 46 Heat map of non-residents' activity during the ALF event (27/11/2014 — 18/11/2015), as inferred from Twitter.

FIGURE 47 Heat map of non-residents' activity during the ALF event (27/11/2014 — 18/11/2015), as inferred from Instagram.

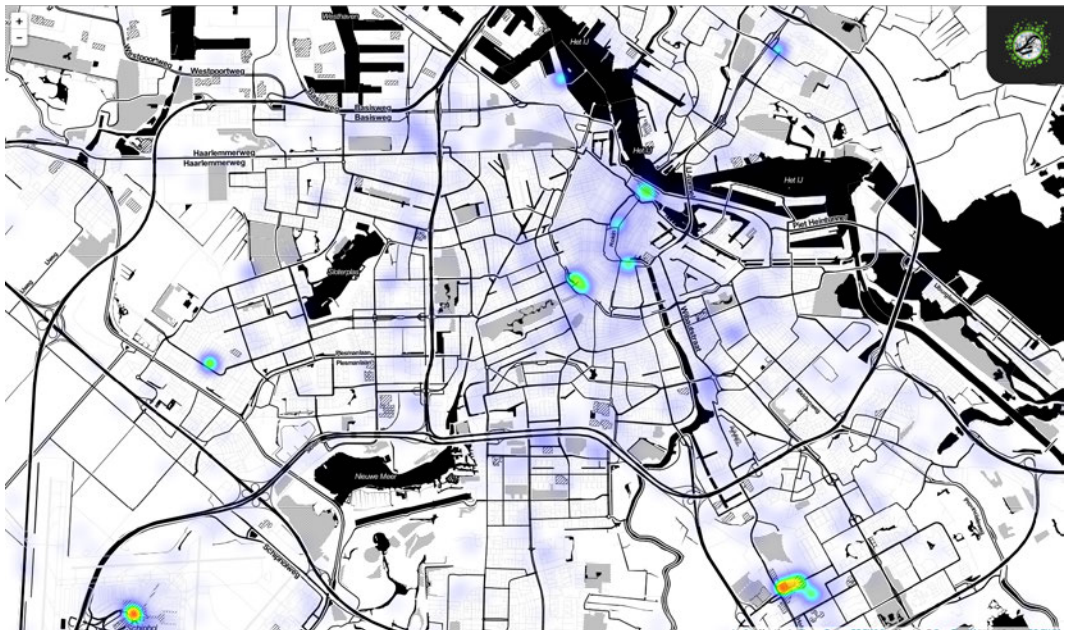


FIGURE 48 Heat map of non-residents' activity before the ALF event (13/11/2014 — 26/11/2014), as inferred from Twitter.

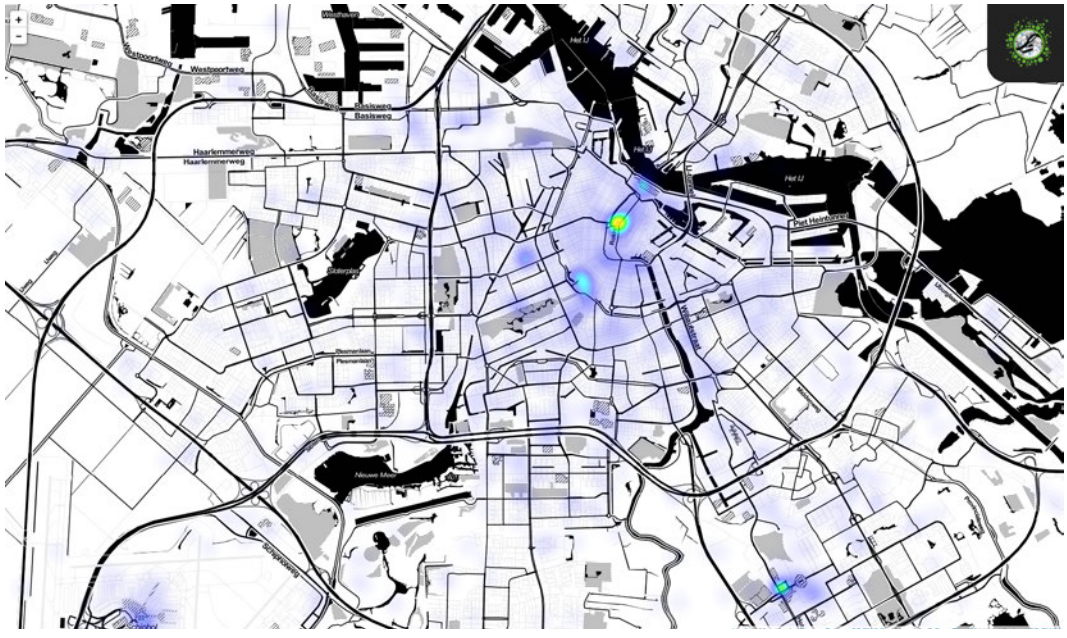


FIGURE 49 Heat map of non-residents' activity before the ALF event (13/11/2014 — 26/11/2014), as inferred from Instagram.

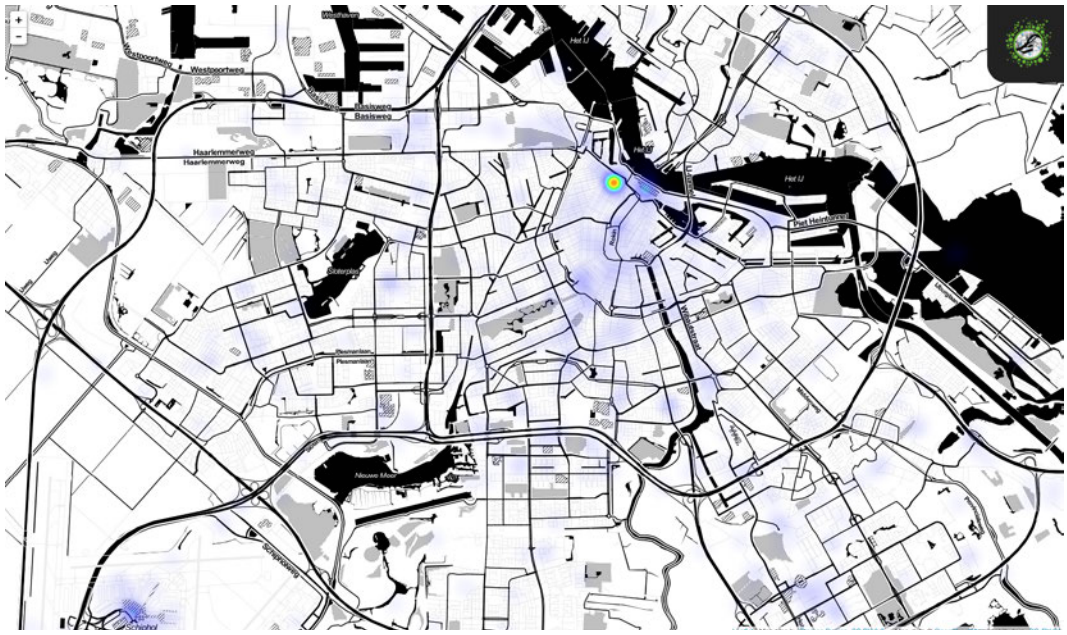


FIGURE 50 Heat map of non-residents' activity after the ALF event (19/01/2015 — 31/01/2015), as inferred from Twitter.

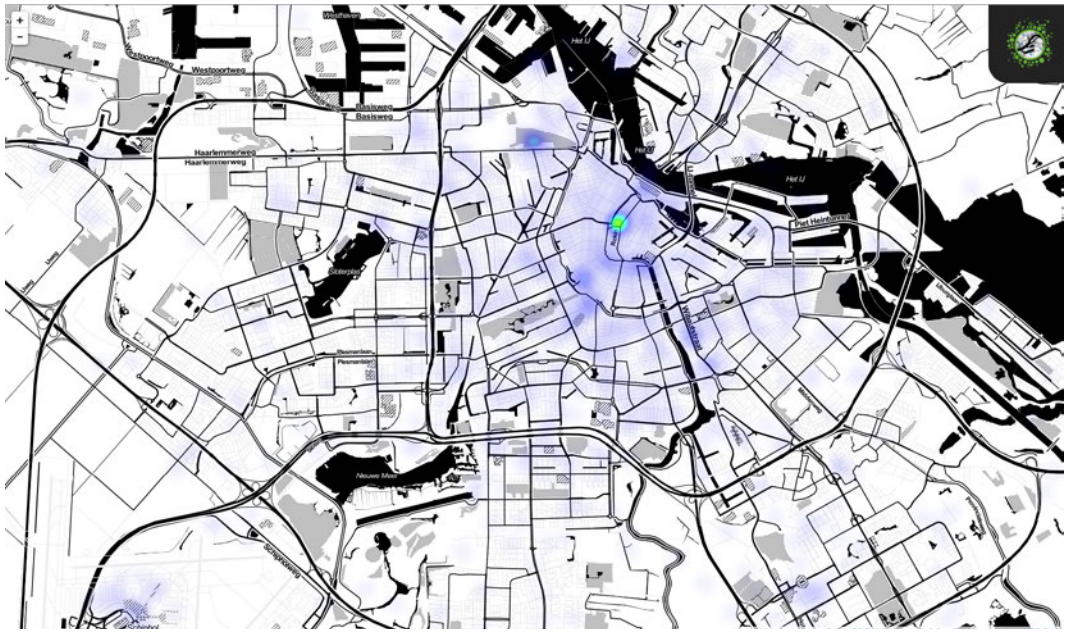


FIGURE 51 Heat map of non-residents' activity after the ALF event (19/01/2015 — 31/01/2015), as inferred from Instagram.

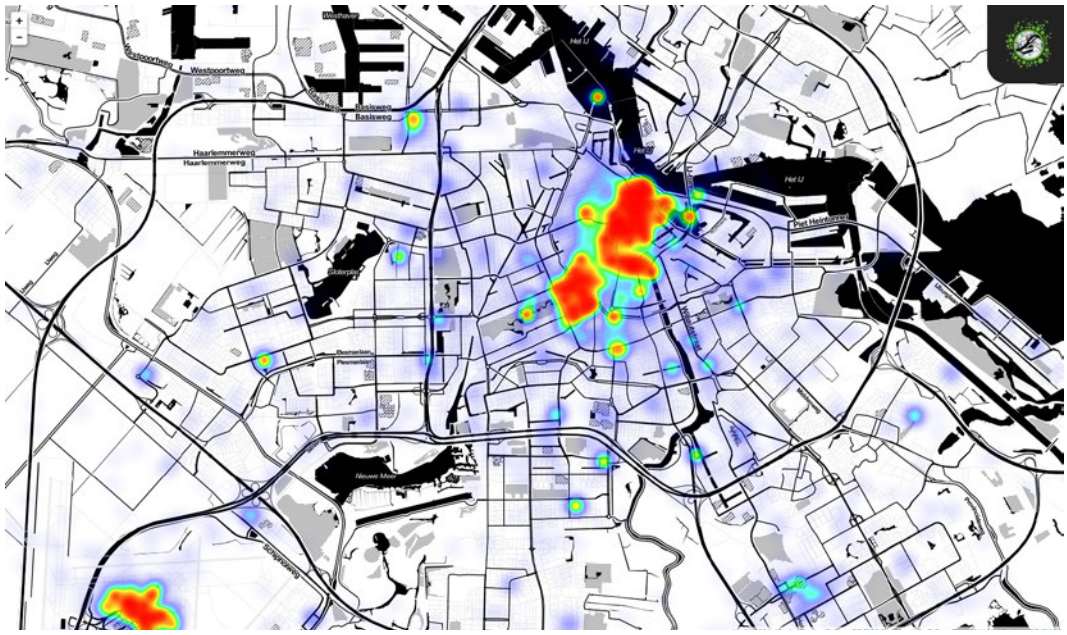


FIGURE 52 Heat map of foreign tourists' activity during the ALF event (27/11/2014 — 18/11/2015), as inferred from Twitter.

FIGURE 53 Heat map of foreign tourists' activity during the ALF event (27/11/2014 — 18/11/2015), as inferred from Instagram.

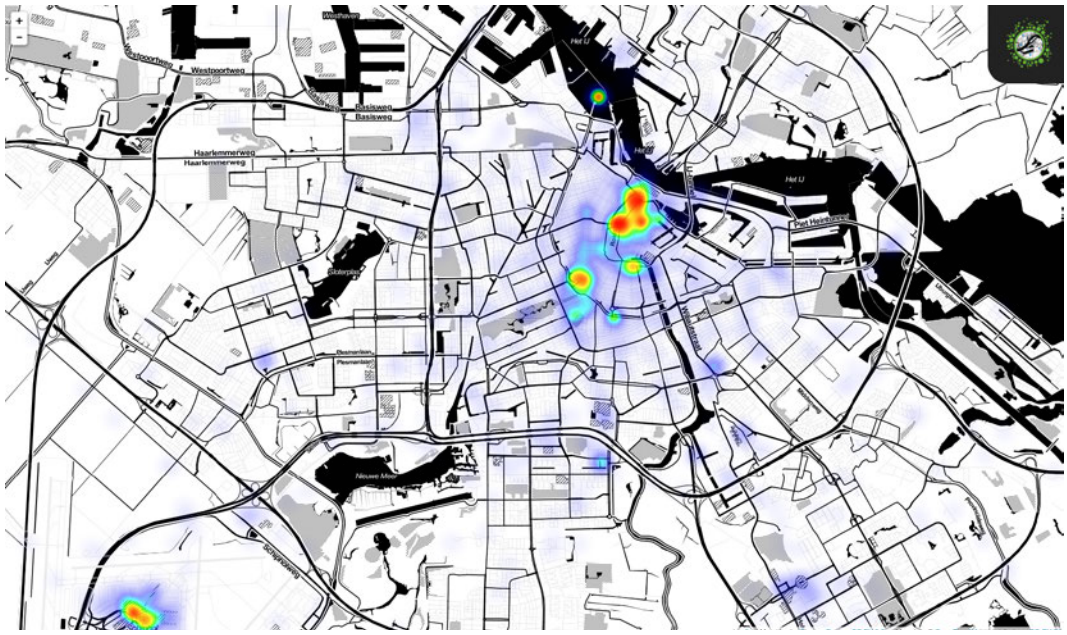


FIGURE 54 Heat map of foreign tourists' activity before the ALF event (13/11/2014 — 26/11/2014), as inferred from Twitter.

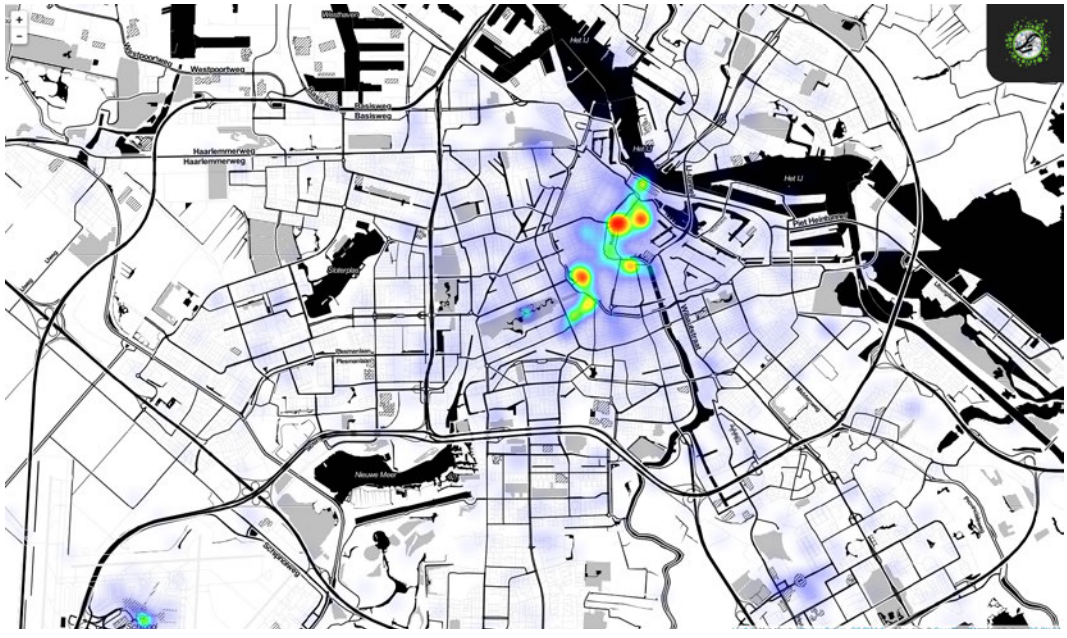


FIGURE 55 Heat map of foreign tourists' activity before the ALF event (13/11/2014 — 26/11/2014), as inferred from Instagram.

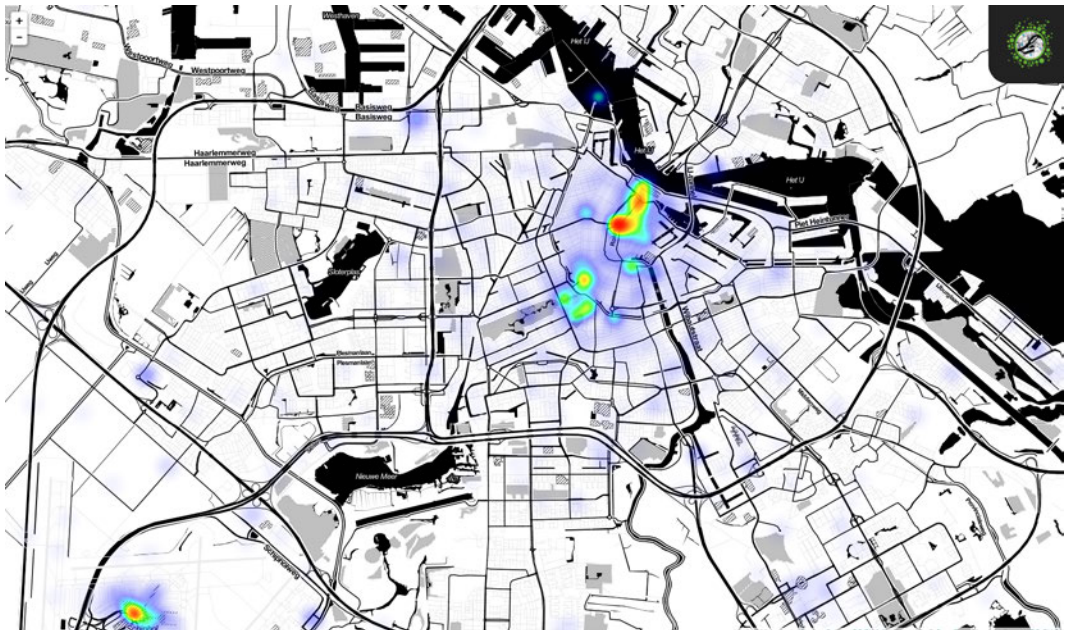


FIGURE 56 Heat map of foreign tourists' activity after the ALF event (19/01/2015 — 31/01/2015), as inferred from Twitter.

FIGURE 57 Heat map of foreign tourists' activity after the ALF event (19/01/2015 — 31/01/2015), as inferred from Instagram.

Appendix E Statistical significance tests for global and local spatial autocorrelation: Expected mean and variance of *z*-scores (Chapter 6)

In Chapter 6, a set of statistical hypothesis tests are presented to assess the significance of the outcomes obtained from the global and local spatial autocorrelation analysis (i.e. Global Moran's I, Local Moran's I_i, and Getis-Ord G_i^*) against the null hypothesis of no spatial autocorrelation. This Appendix provides the set of formulas that describe the expected mean and variance of the global and local z-scores for the resampling and randomization hypotheses.

Global tests for spatial autocorrelation

In the *resampling hypothesis*, the score of statistical significance $z_N(I)$ is given by:

$$z_N(I) = \frac{I - E_N(I)}{\sqrt{V_N(I)}} \tag{1}$$

Where the expected mean $E_N(I)$ is given by:

$$E_N(I) = -\frac{1}{(n-1)} \tag{2}$$

And the corresponding variance $V_N(I)$ is:

$$V_N(I) = \frac{n^2 W_1 - nW_2 + 3W_0^2}{(n^2-1)W_0^2} - [E_N(I)]^2 \tag{3}$$

Where:

TOC

$$W_0 = \sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{ij} \tag{4}$$

$$W_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(w_{ij} + w_{ji}\right)^2 \tag{5}$$

$$W_2 = \sum_{k=1}^{n}\left(\sum_{j=1}^{n} w_{kj} + \sum_{i=1}^{n} w_{ik}\right)^2 \tag{6}$$

Accordingly, for the *randomization hypothesis* the score of statistical significance $z_R(I)$ is given by:

$$z_R(I) = \frac{I - E_R(I)}{\sqrt{V_R(I)}} \tag{7}$$

Where the expected mean $E_R(I)$ is given given by:

$$E_R(I) = -\frac{1}{(n-1)} \tag{8}$$

And the corresponding variance $V_R(I)$ is:

$$V_R(I) = \frac{n[(n^2-3n+3)W_1 - nW_2 + 3W_0^2]}{(n-1)(n-2)(n-3)W_0^2} - \frac{b_2[(n^2-n)W_1 - 2nW_2 + 6W_0^2]}{(n-1)(n-2)(n-3)W_0^2} - [E_R(I)]^2 \tag{9}$$

Where:

$$b_2 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})^4 / n}{\left[\sum_{i=1}^{n}(z_i - \bar{z})^2 / n\right]^2} \tag{10}$$

The weights W_0, W_1, W_2 are given by Eq. (4), (5), (6) respectively.

The corresponding $z(I_i)$-scores of a random permutation test, under a null hypothesis of no spatial association, given by the following (Anselin, 1995):

$$z_R(I_i) = \frac{I_i - E_R(I_i)}{\sqrt{V_R(I_i)}} \tag{11}$$

Where the expected mean $E_R(I_i)$ is given by:

$$E_R(I_i) = -\frac{\sum_{j=1}^{n} W_{ij}}{(n-1)} \tag{12}$$

And the variance $V_R(I_i)$ is:

$$V_R(I_i) = \sum_{j \neq i}^{n} W_{ij}^2 \frac{(n-b_2)}{(n-1)} + \sum_{k \neq i}^{n} \sum_{h \neq i}^{n} W_{ik} W_{ih} \frac{(2b_2 - n)}{(n-1)(n-2)} - [E_R(I_i)]^2 \tag{13}$$

Where b_2 is given by Eq. (10).

TOC

Appendix F Local spatial autocorrelation analysis of human activity (scatterplots, choropleths, cluster maps) (Chapter 6)

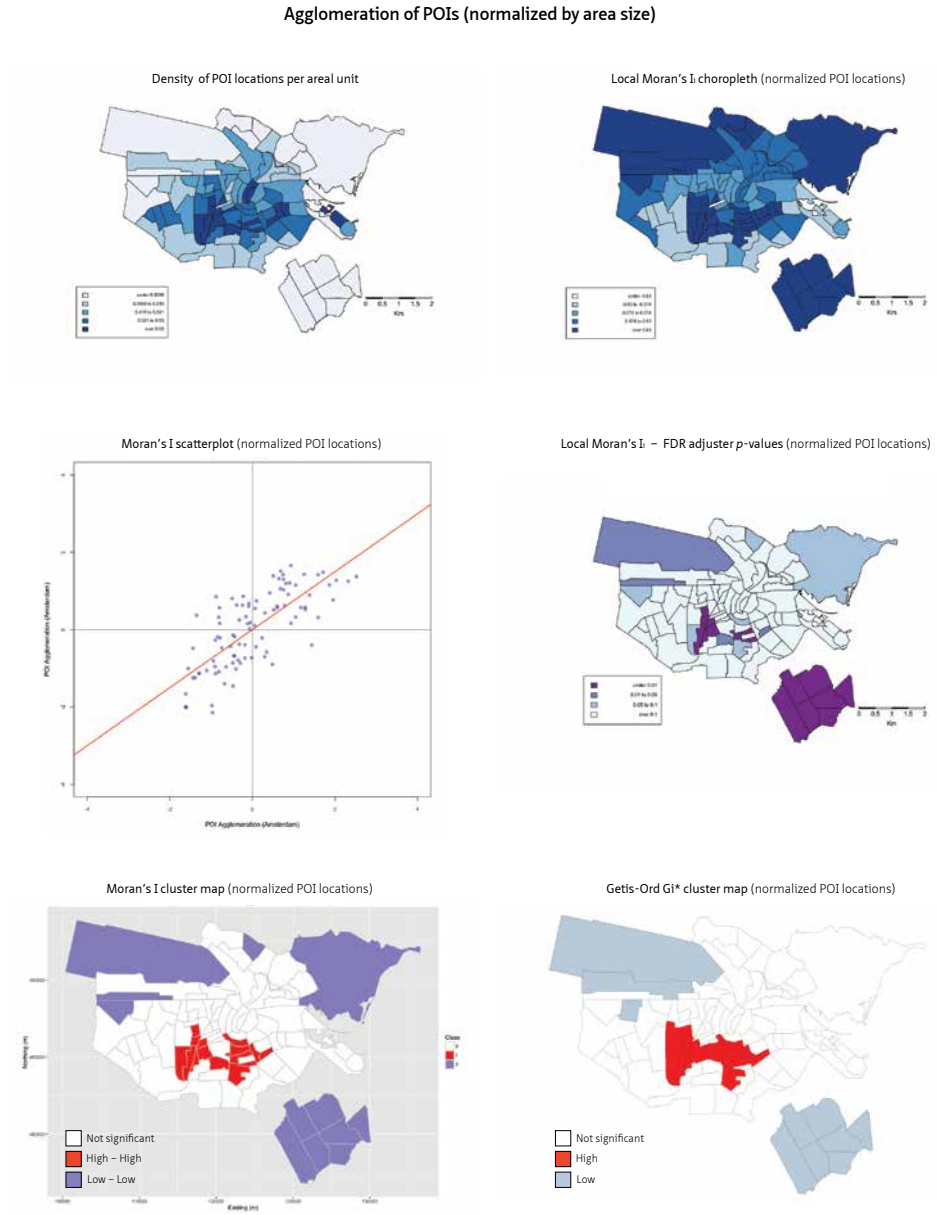


FIGURE 58 Spatial autocorrelation analysis of the density of POI locations (normalized by area size).

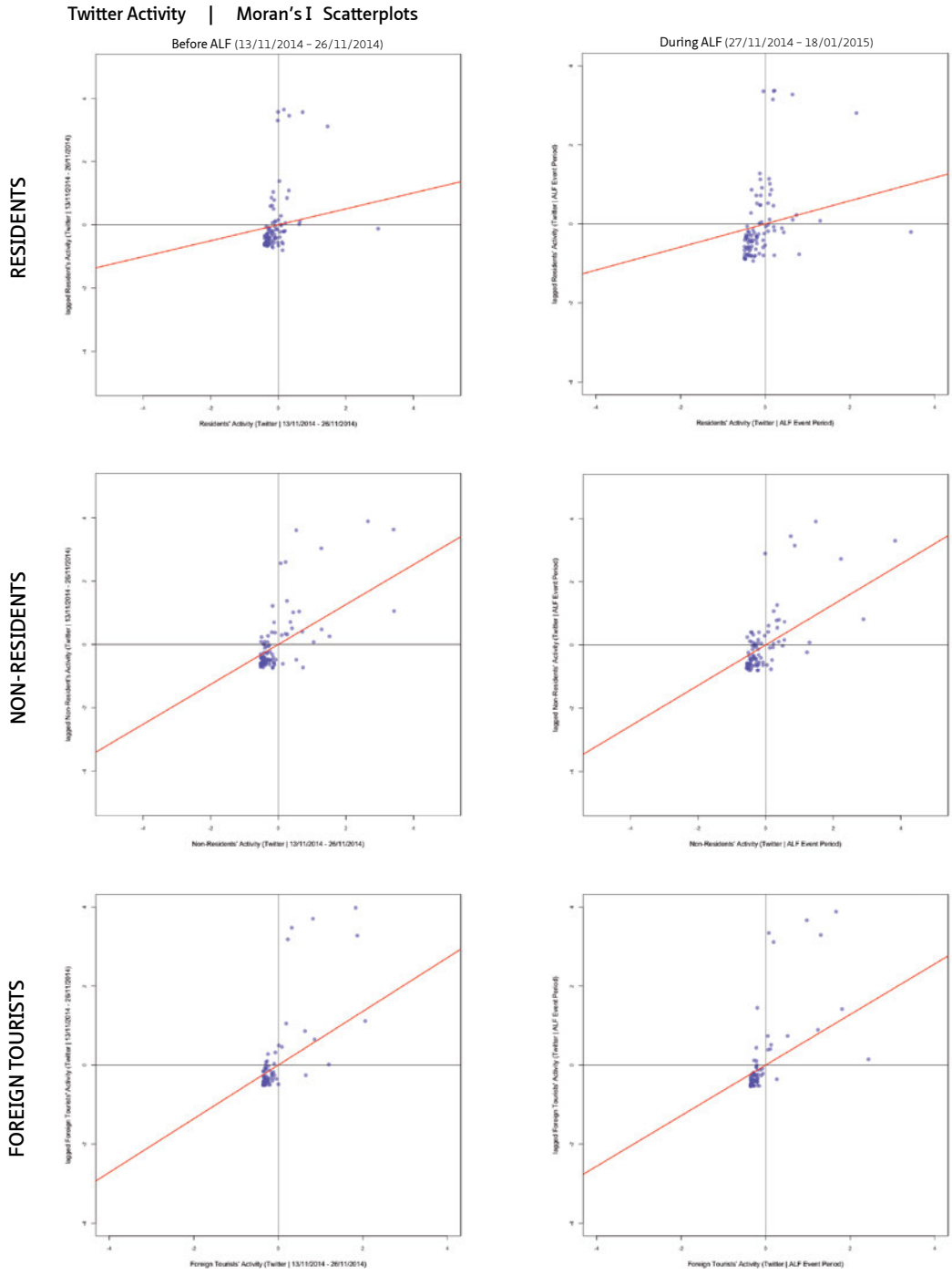
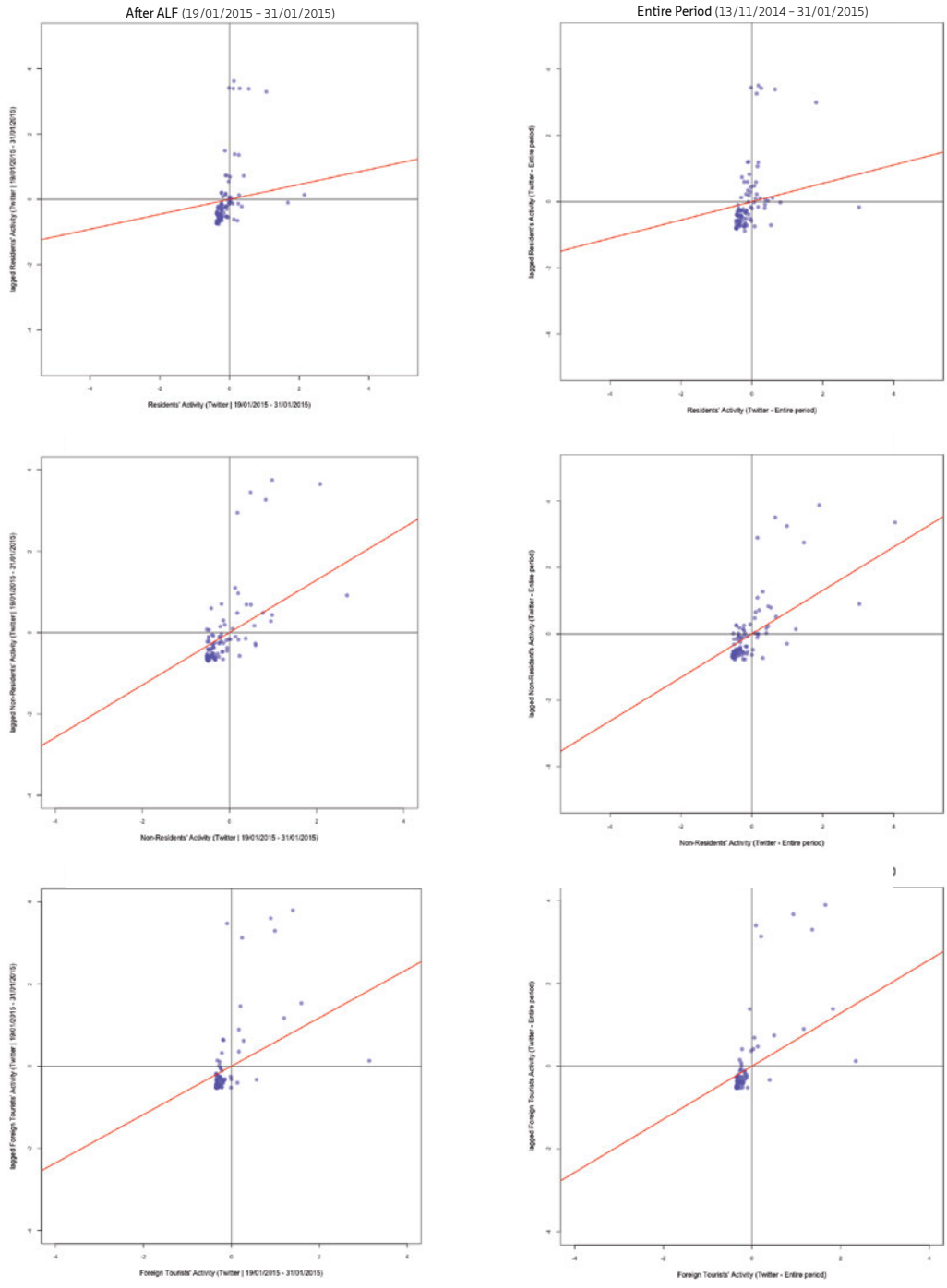


FIGURE 59 Moran's I scatterplots of Twitter activity (different social categories, different time periods). Each dot represents an areal unit (i.e. postcode area). Areas in the upper right and lower left quadrants indicate positive spatial autocorrelation (i.e. high I_i-values neighboring with other high I_i-value areas, or low values with low values), thus contributing more to the overall result

TOC



of global autocorrelation. Conversely, the areas in the upper left and lower right quadrants indicate negative spatial autocorrelation (i.e. high I_i-values with low values).

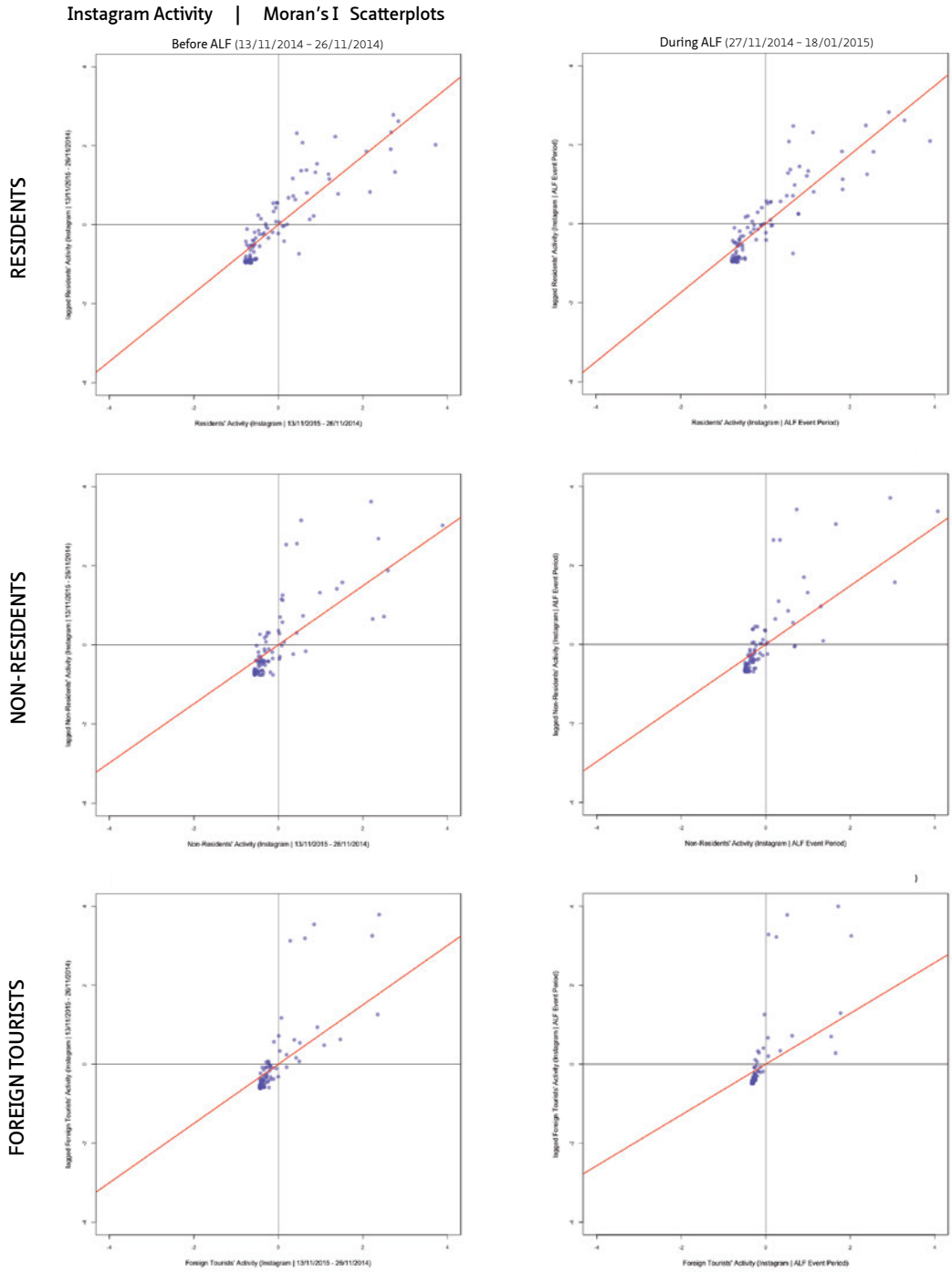
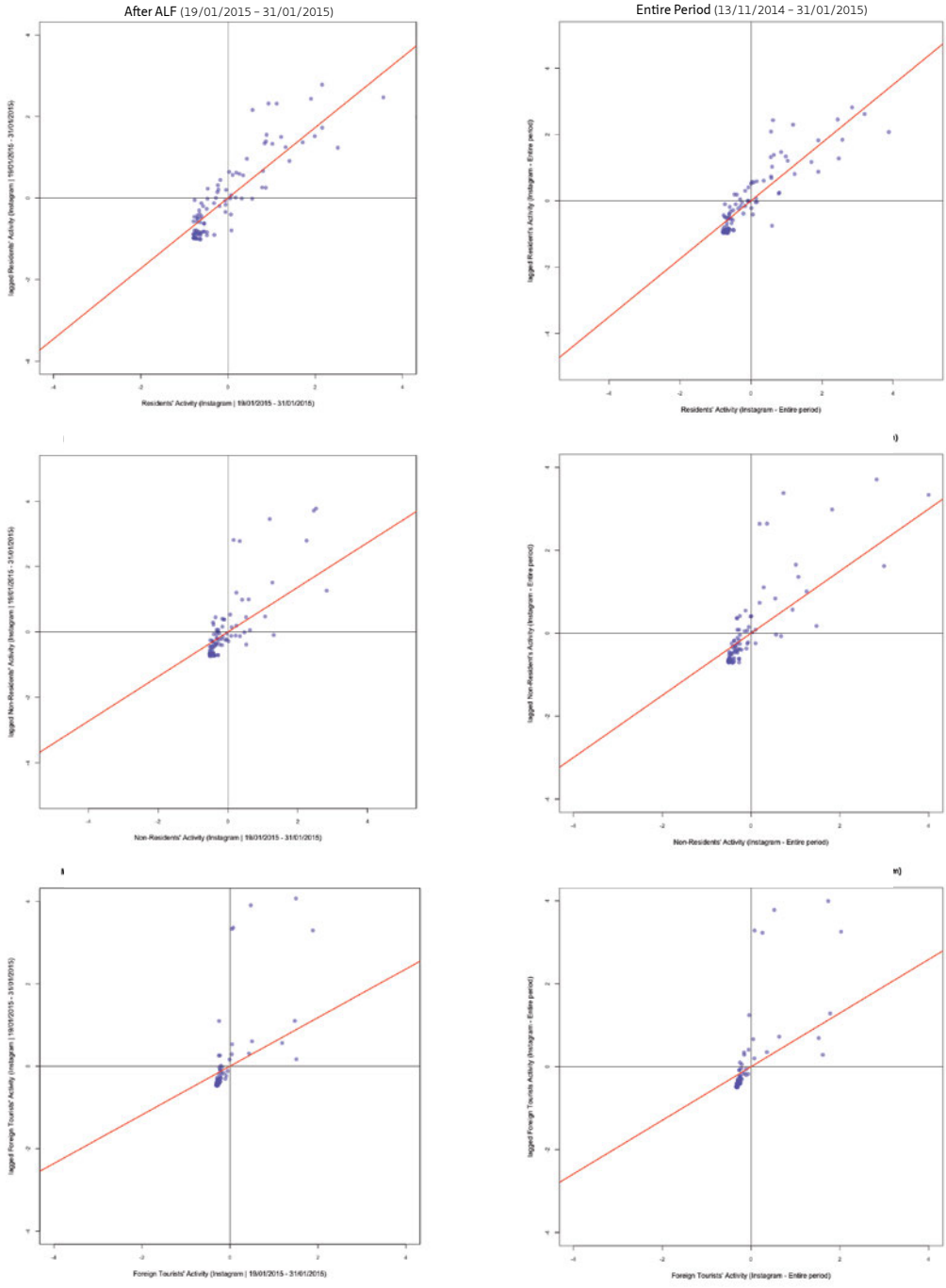


FIGURE 60 Moran's *I* scatterplots of Instagram activity (different social categories, different time periods). Each dot represents an areal unit (i.e. postcode area).

TOC


After ALF (19/01/2015 – 31/01/2015)
Entire Period (13/11/2014 – 31/01/2015)
Residents' Activity (Instagram | 19/01/2015 - 31/01/2015)
Residents' Activity (Instagram - Entire period)
Non-Residents' Activity (Instagram | 19/01/2015 - 31/01/2015)
Non-Residents' Activity (Instagram - Entire period)
Foreign Tourists' Activity (Instagram | 19/01/2015 - 31/01/2015)
Foreign Tourists' Activity (Instagram - Entire period)

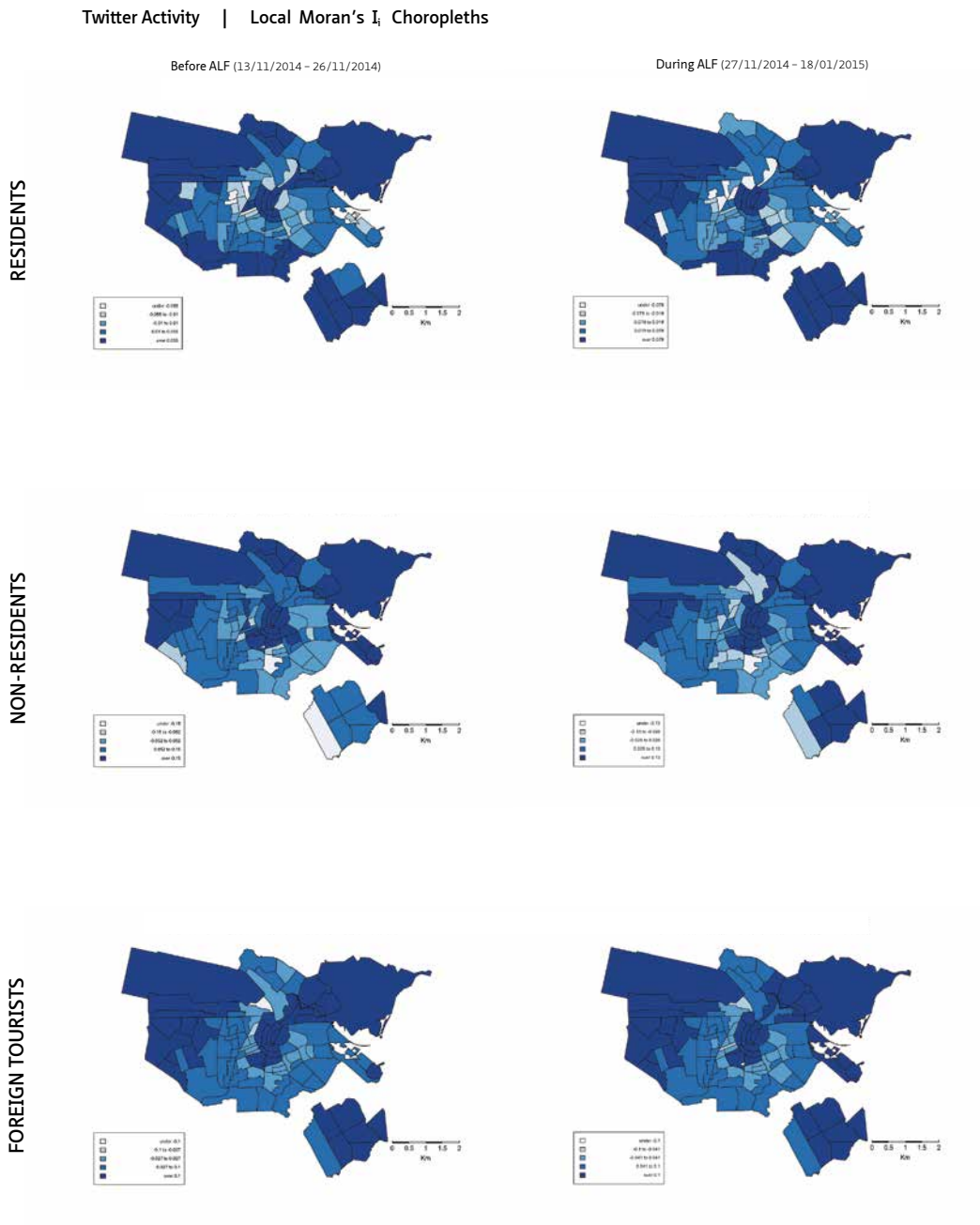
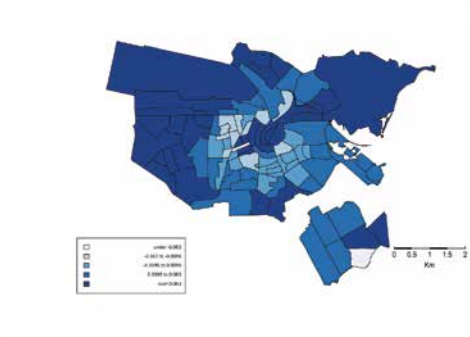


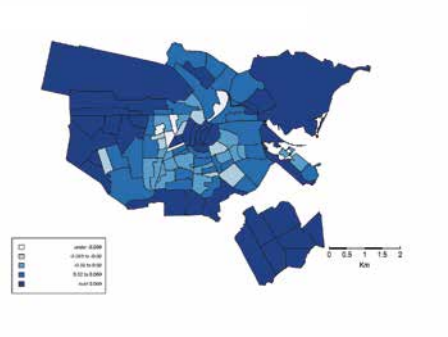
FIGURE 61 Choropleths of local Moran's I_i values of Twitter activity (different social categories, different time periods. Areas are shaded in proportion to their respective I_i-values (also illustrated in the Moran's scatterplots – Fig. 59-60).

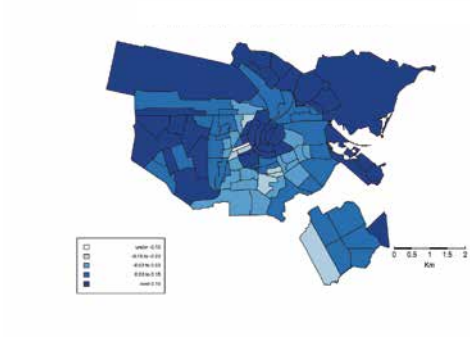
TOC

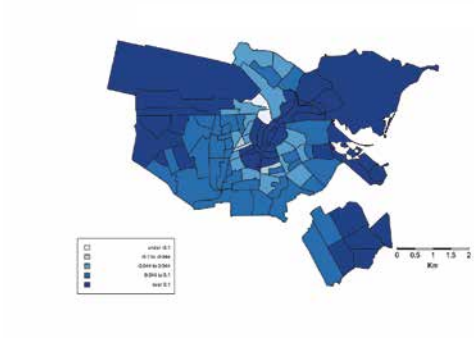
After ALF (19/01/2015 – 31/01/2015)



Entire Period (13/11/2014 – 31/01/2015)







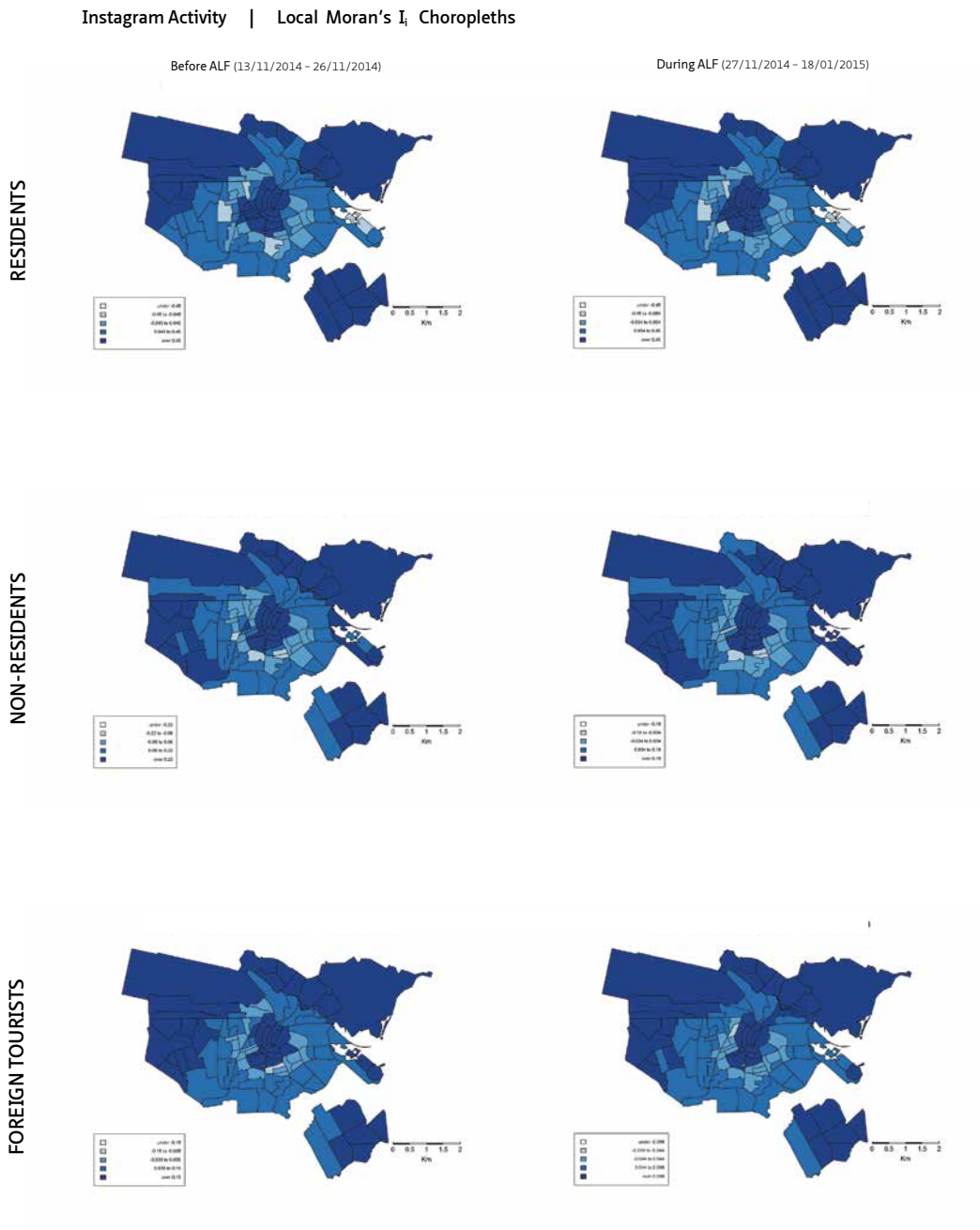
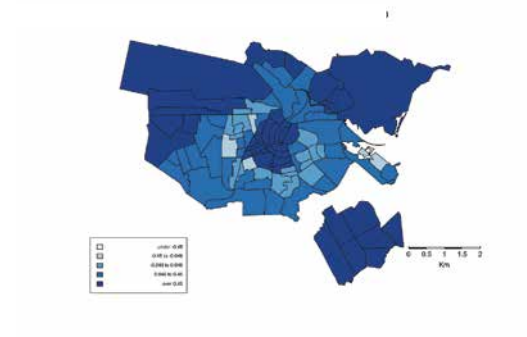


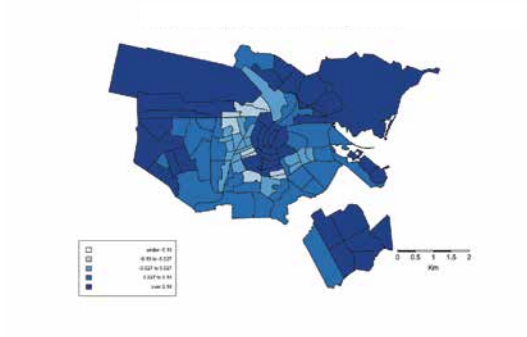
FIGURE 62 Choropleths of local Moran's I_i values of Instagram activity (different social categories, different time periods. Areas are shaded in proportion to their respective I_i-values (also illustrated in the Moran's scatterplots – Fig. 59-60).

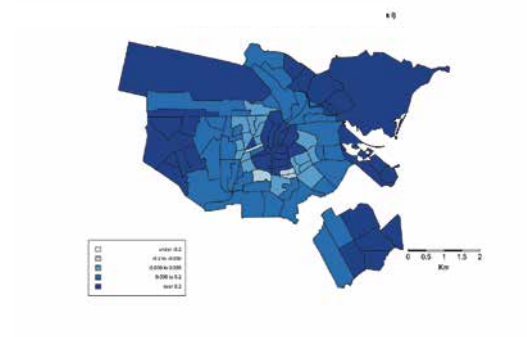
TOC

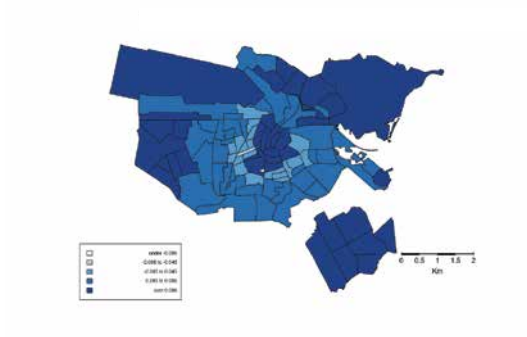
After ALF (19/01/2015 – 31/01/2015)

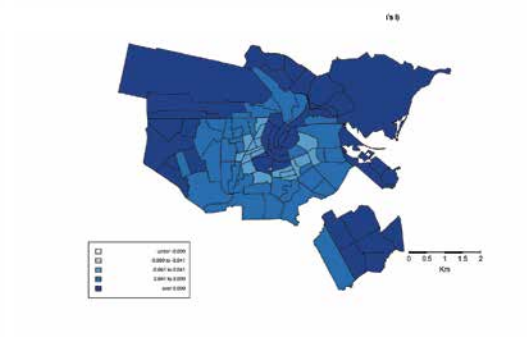
Entire Period (13/11/2014 – 31/01/2015)











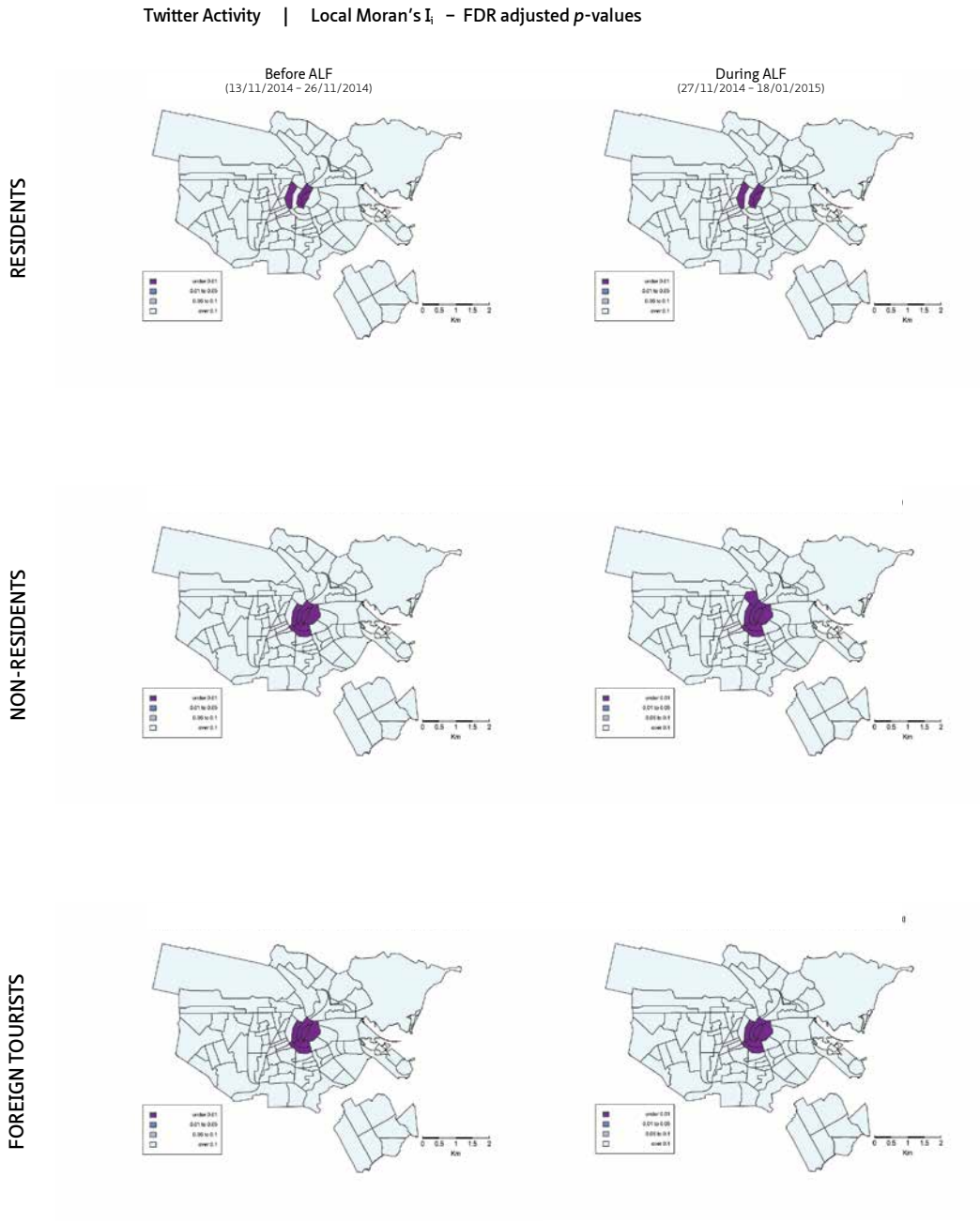
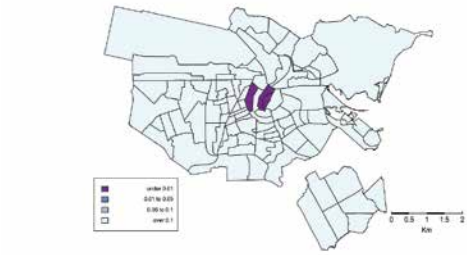


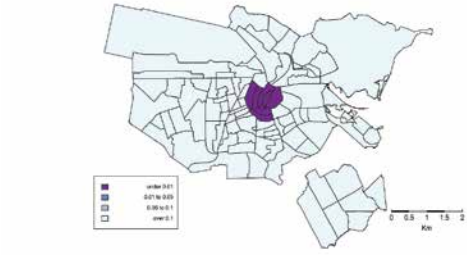
FIGURE 63 False Discovery Rate (FDR) adjustments of p-values for Twitter activity, to determine the probability of falsely detecting significant clusters of I_i-values. Dark purple areas suggest that the identified HH clusters are indeed statistically significant and, therefore, the null hypothesis of zero spatial autocorrelation can be rejected.

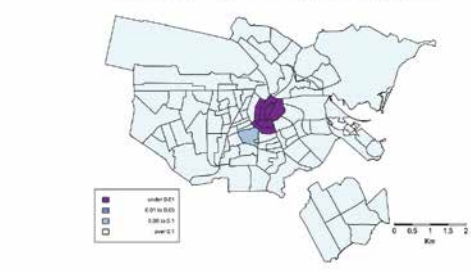
TOC

After ALF
(19/01/2015 – 31/01/2015)



Entire Period
(13/11/2014 – 31/01/2015)





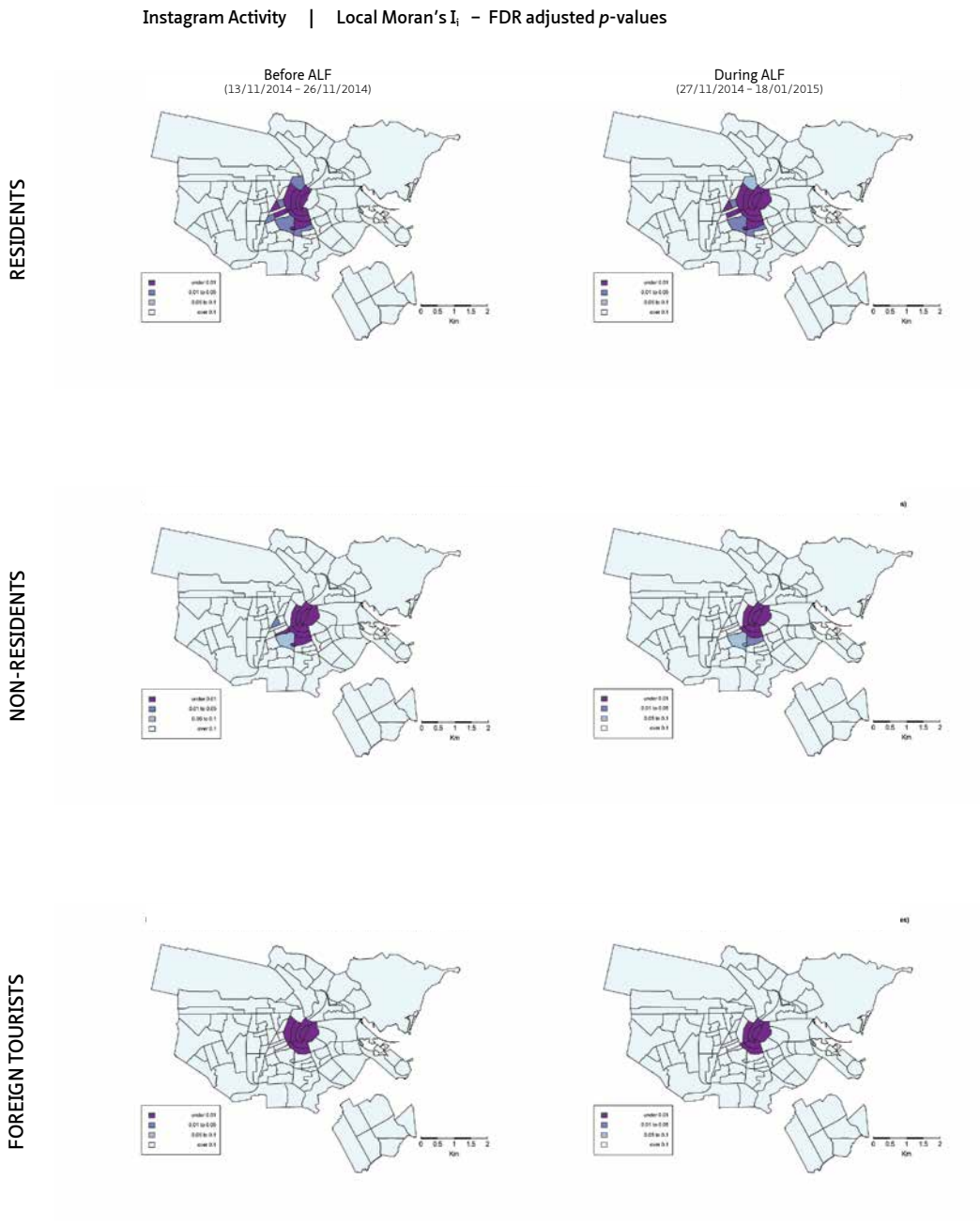
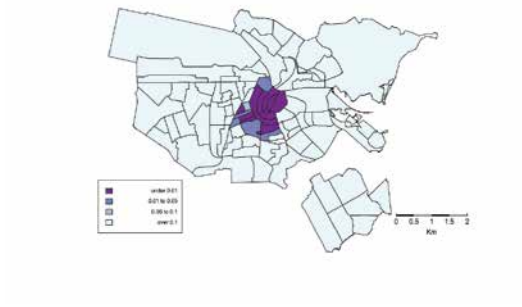


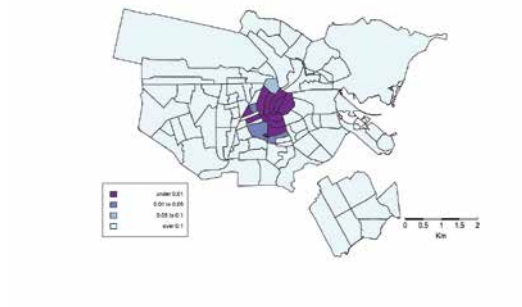
FIGURE 64 False Discovery Rate (FDR) adjustments of *p*-values for Instagram activity, to determine the probability of falsely detecting significant clusters of I_i-values. Dark purple areas suggest that the identified HH clusters are indeed statistically significant and, therefore, the null hypothesis of zero spatial autocorrelation can be rejected.

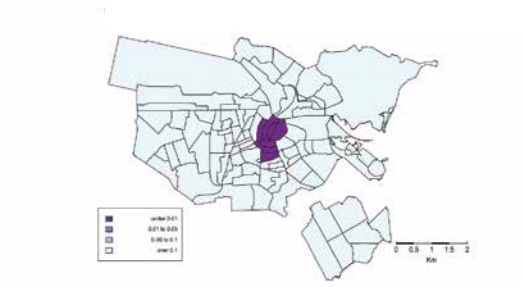
TOC

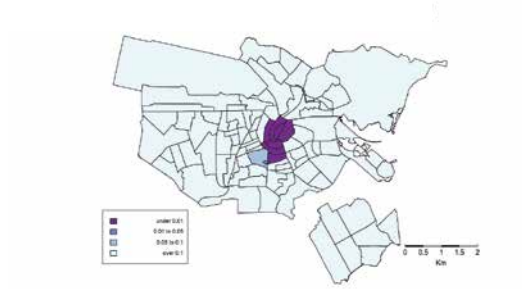
After ALF
(19/01/2015 – 31/01/2015)

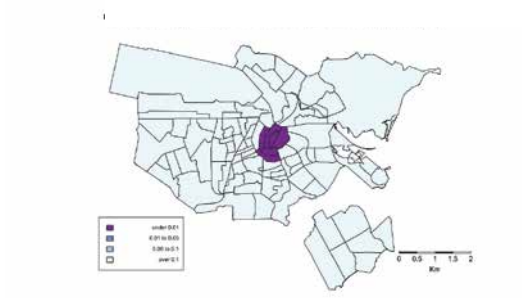


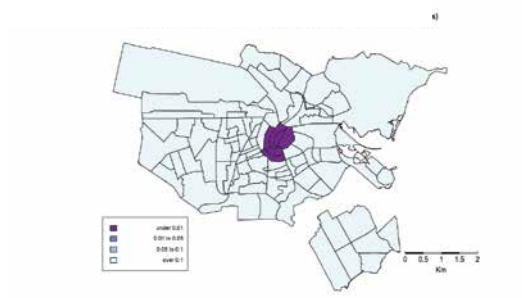
Entire Period
(13/11/2014 – 31/01/2015)











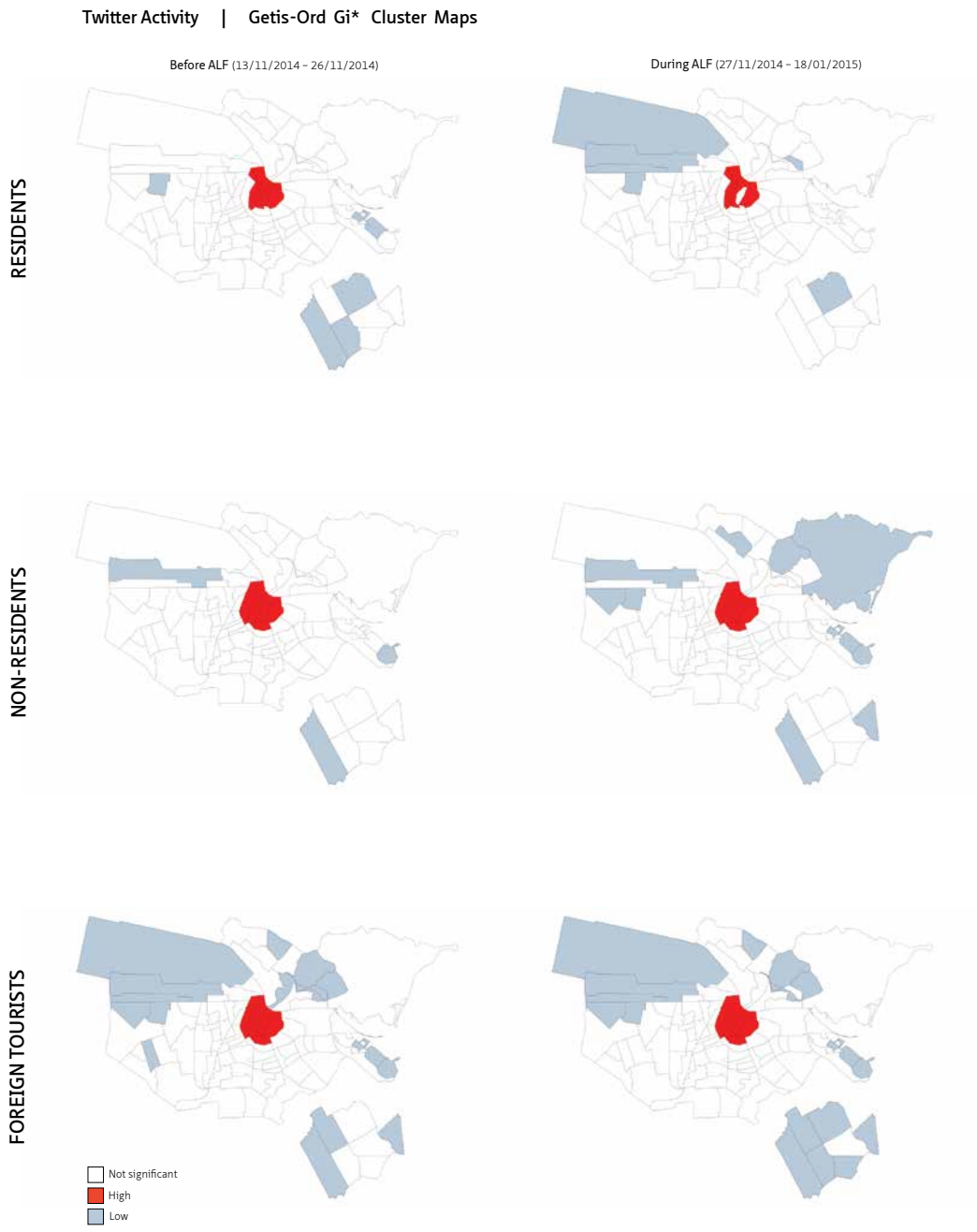
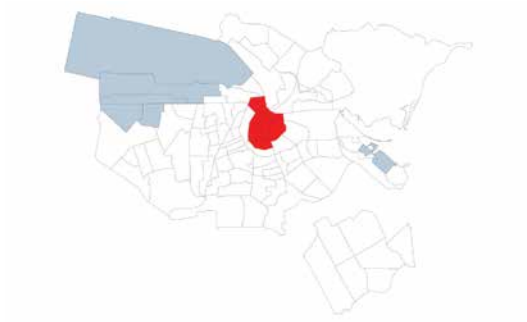


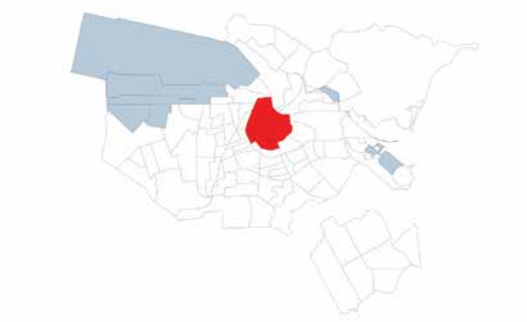
FIGURE 65 Getis-Ord G_i^*-cluster maps of Twitter activity (different social categories, different time periods). Red areas indicate clusters of high G_i^*-values (hotspots), whereas the light blue/green areas indicate clusters of low G_i^*-values (coldspots).

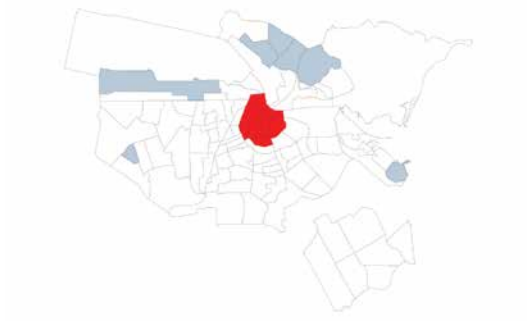
TOC

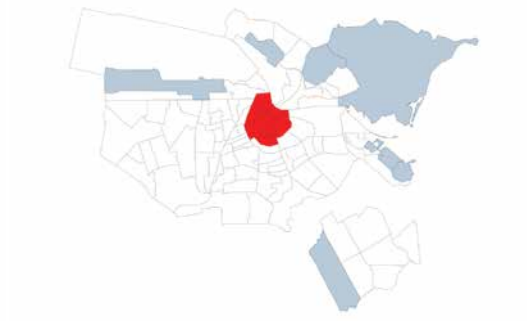
After ALF (19/01/2015 – 31/01/2015)

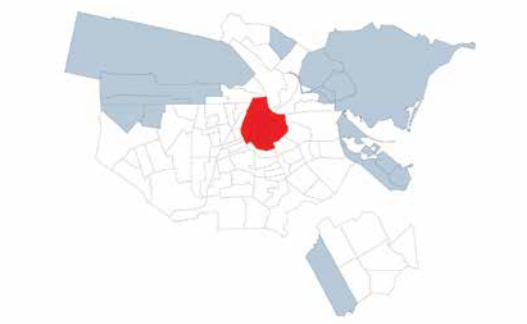
Entire Period (13/11/2014 – 31/01/2015)

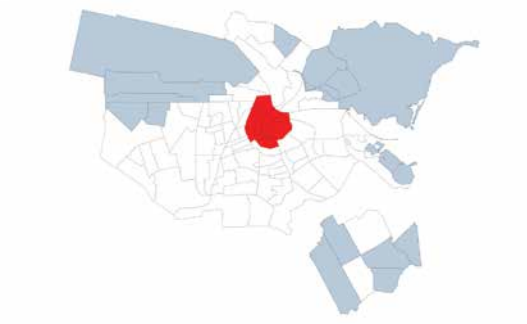












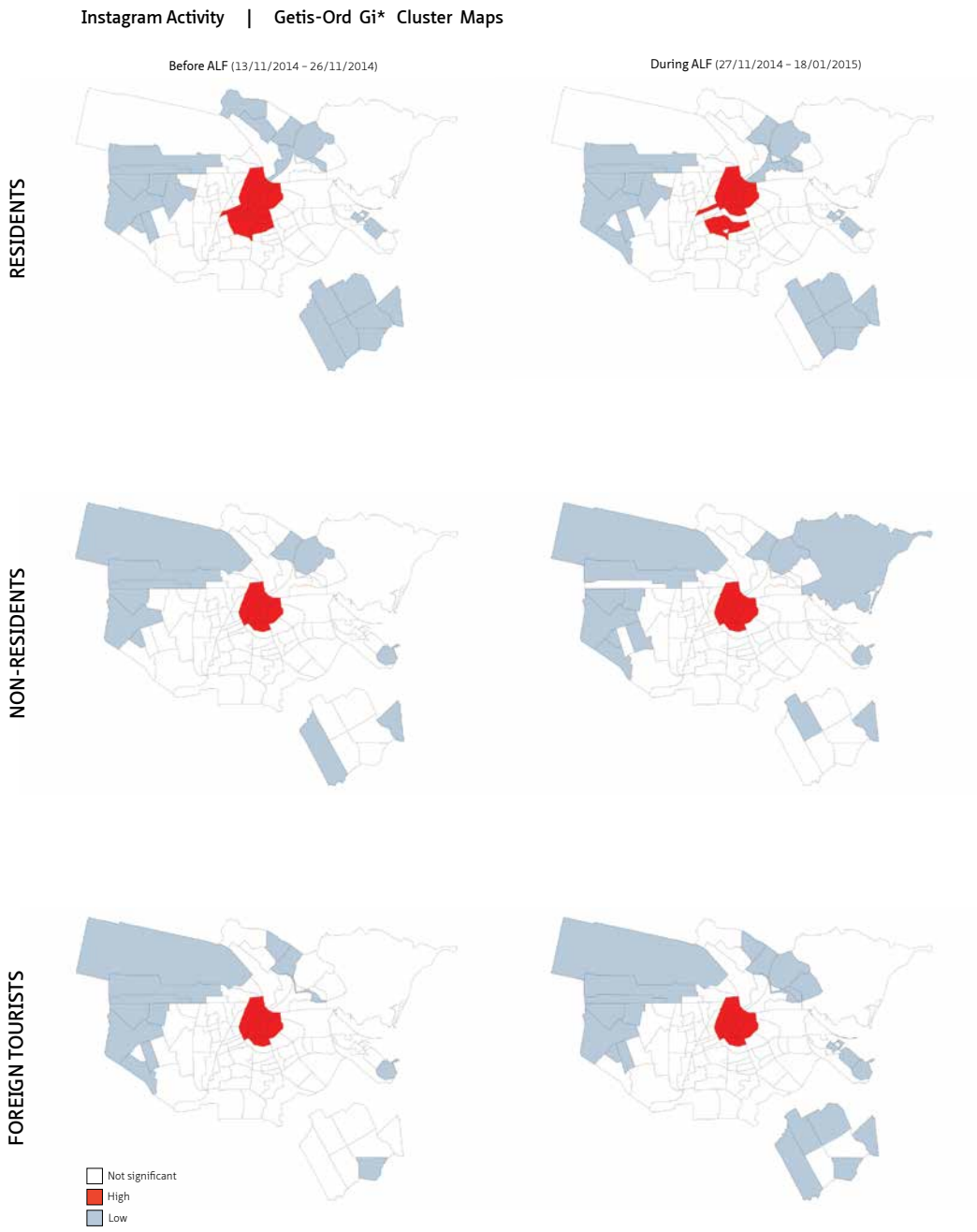
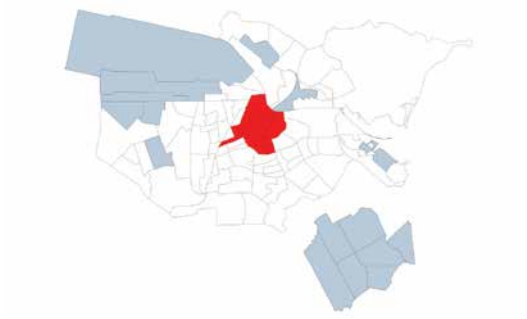


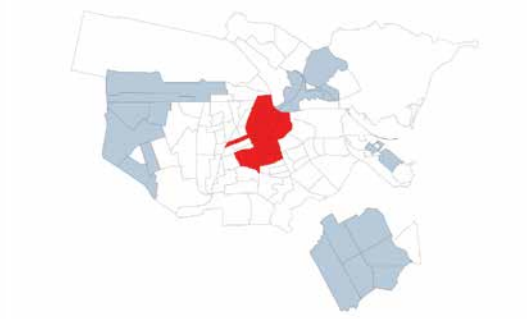
FIGURE 66 Getis-Ord G_i^*-cluster maps of Instagram activity (different social categories, different time periods). Red areas indicate clusters of high G_i^*-values (hotspots), whereas the light blue/green areas indicate clusters of low G_i^*-values (coldspots).

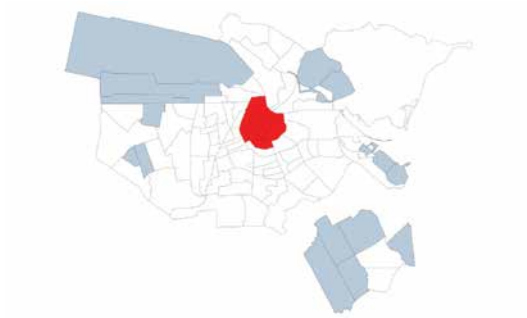
TOC

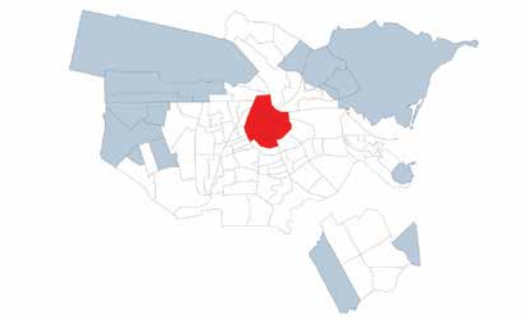
After ALF (19/01/2015 – 31/01/2015)

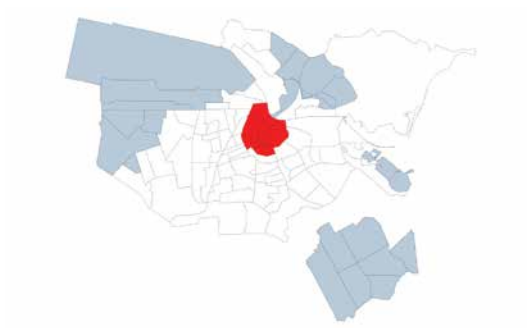


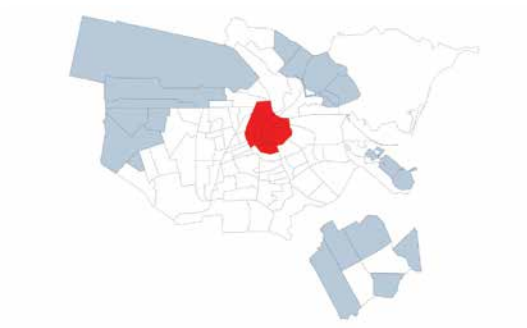
Entire Period (13/11/2014 – 31/01/2015)











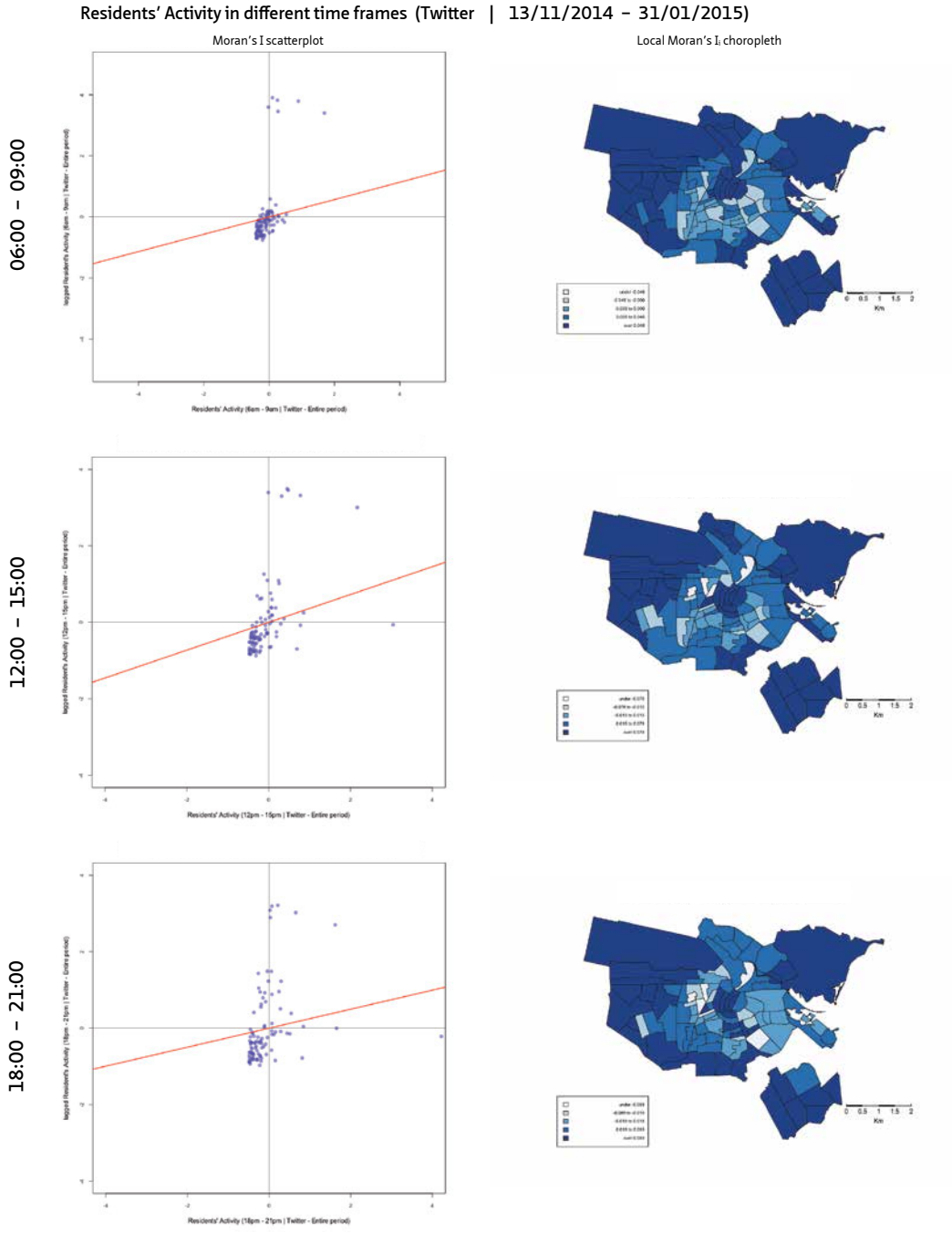


FIGURE 67 Spatial autocorrelation analysis of residents' activity in different time frames within a day for the entire period (Moran's *I* scatterplots, local Moran's *I* choropleths, FDR adjustments of *p*-vaues, and local Moran's *I* cluster maps).

TOC

Local Moran's I_i – FDR adjusted *p*-values

Local Moran's I_i cluster map

TOC

Appendix G Source code – Spatial autocorrelation analysis (Chapter 6)

The spatial autocorrelation analysis (global and local indicators, along with statistical tests) of different variables of human activity, described in Chapter 6, has been carried out by means of the open-source R statistical language. The following piece of code is an indicative example of the scripts that have been written in this regard, and refers specifically to the analysis of residents' activity, as inferred from Twitter, covering the entire period in question (i.e. 13/11/2014 – 31/01/2015). For the spatial autocorrelation analysis and the respective tests of statistical significance of the remaining variables listed in Table 17, similar scripts have been written.

R Code (Residents' activity – Twitter | Entire period)

```
#Libraries
require(GISTools)
require(lctools)
require(ggplot2)
require(rgeos)
require(spdep)

#Load shapefile
alf <- readShapePoly("Amsterdam_ALF.shp")

#Combine centroid coordinates
Coords <- cbind(alf$Xi, alf$Yi)

#Define k for weights
bw <- c(2, 3, 4, 5, 6, 9, 12, 18, 24)

#Global Moran's I
moran <- matrix(data = NA, nrow = 9, ncol = 7)

#For-Loop, calculation of Global Moran's I for multiple k-neighbors
counter <- 1
for (b in bw) {
    moranI <- moransI(Coords, b, alf$RES_T_TOTn)
    moran[counter,1] <- counter
    moran[counter,2] <- b
```

```
    moran[counter,3] <- moranI$Morans.I
    moran[counter,4] <- moranI$z.resampling
    moran[counter,5] <- moranI$p.value.resampling
    moran[counter,6] <- moranI$z.randomization
    moran[counter,7] <- moranI$p.value.randomization
    counter <- counter+1
}

colnames(moran) <- c("ID", "k", "Moran's I", "Z resampling", "p-value resampling", "Z
randomization", "p-value randomization")

moran

#Save csv file of Global Moran's I matrix
write.csv(moran, file = "MoransI_RES_T_TOT.csv", row.names = F)

#Local Moran's I calculation
l.moran <- l.moransI(Coords, 6, alf$RES_T_TOTn)

#False Discovery Rate (FDR) test
pval.shade <- shading(c(0.01, 0.05, 0.1), cols = rev(brewer.pal(4, 'BuPu')))
choropleth(alf, p.adjust(l.moran$p.value, method = 'fdr'), shading = pval.shade)

choro.legend(109450.6, 481826.2, pval.shade, cex = 0.8)
title("Residents' Activity - Twitter / Entire period (Local Moran's I / FDR adjusted
p-values)", cex.main=1)
map.scale(132909.3, 479326.5, 5000, "Km", 4, 0.5)

#Local Moran's I choropleth
shade_M <- auto.shading(c(l.moran$Ii, -l.moran$Ii), cols = brewer.pal(5, "Blues"))
choropleth(alf, l.moran$Ii, shade_M)

choro.legend(108104.6, 480095.7, shade_M, cex = 0.7)
title("Residents' Activity | Twitter - Entire period (Local Moran's I)", cex.main=1)
map.scale(132909.3, 479326.5, 5000, "Km", 4, 0.5)

#Local Moran's I scatterplot preparation
xmin <- round(ifelse(abs(min(l.moran[,7])) > abs(min(l.moran[,8])), abs(min(l.
moran[,7])), abs(min(l.moran[,8]))))
xmax <- round(ifelse(abs(max(l.moran[,7])) > abs(max(l.moran[,8])), abs(max(l.
moran[,7])), abs(max(l.moran[,8]))))
xmax <- ifelse(xmin > xmax, xmin, xmax)+1
ymax <- xmax
```

```
xmin <- -xmax
ymin <- -ymax

reg1 <- lm(l.moran[,8]~l.moran[,7])

#Local Moran's I scatterplot
plot(l.moran[,7], l.moran[,8], xlim=c(xmin+5, xmax-5), ylim=c(ymin+5, ymax-5),
main="Moran's I scatterplot of Residents' Activity (Entire period - Twitter)",
xlab="Residents' Activity (Twitter - Entire period)", ylab="lagged Resident's Activity
(Twitter - Entire period)",
pch=16, col=rgb(0.4,0.4,0.8,0.6), cex=1.3, cex.axis=0.8, col.axis="gray25")

abline(h=0, lwd=0.7)

abline(v=0, lwd=0.7)

abline(reg1, col=2, lwd=2)

#Getis-Ord Gi* statistic
res.tot.G <- localG(alf$RES_T_TOT, bw)

#Merge data with the map
alf@data$Idx <- seq_len(nrow(alf))
alf_temp <- merge(alf@data, l.moran, by.x="ObjectID", by.y="ID", sort=F, all=T)
alf@data <- alf_temp[order(alf_temp$Idx), ]
alf@data$Idx <- seq_len(nrow(alf))
alf_temp1 <- merge(alf@data, res.tot.G, by.x="ObjectID", by.y="ID", sort=F, all=T)
alf@data1 <- alf_temp1[order(alf_temp1$Idx), ]

#Data preparation for map visualization
map.f <- fortify(alf, region = "ObjectID")
map.f <- merge(map.f, alf@data, by.x="id", by.y="ObjectID")
map.f1 <- fortify(alf, region = "ObjectID")
map.f1 <- merge(map.f1, alf@data1, by.x="id", by.y="ObjectID")

#Map visualization
map <- ggplot(map.f, aes(long, lat, group=group)) +
geom_polygon(colour="gray80", aes(fill=as.factor(Cluster))) +
scale_fill_manual(values = c("white", "red", "gray50", "turquoise", "pink")) +
coord_equal() +
labs(x="Easting (m)", y="Northing (m)", fill="Class") +
ggtitle("Moran's I Cluster Map (Residents' Activity - Twitter/Entire period)")
```

```
map

mapl <- ggplot(map.fl, aes(long, lat, group=group)) +
geom_polygon(colour="gray80", aes(fill=as.factor(Cluster))) +
scale_fill_manual(values = c("white", "red", "#B8CADB")) +
coord_equal() +
labs(x="Easting (m)", y="Northing (m)", fill="Class") +
ggtitle("Getis-Ord Gi* Cluster Map (Residents' Activity - Twitter/Entire period)")

mapl
```

Curriculum Vitae

Achilleas Psyllidis was born on September 12, 1983 in Athens, Greece. In 2007 he received a professional diploma (Dipl. Ing.) *summa cum laude* in Architectural Engineering from the National Technical University of Athens (NTUA). He also received the Andreas Ploumistos Award for achieving the highest ranking among his graduating class, as well as the Technical Chamber of Greece (TEE) Award for his graduation project, concerning large-scale urban interventions at the Athenian seafront. Throughout his studies, he has received several awards for his performance as a student, namely, the Stamos Stournas Award (2004), the Kalliopi Venizelou-Sfaellou Award (2005), the Lysandros Kaftantzoglou Award (2006), the Thomaides Award (2006), and yearly scholarships from the Greek State Scholarships Foundation (2001 – 2006). In 2005, his project on experimental sustainable school buildings was selected and exhibited at the 22nd UIA (International Union of Architects) World Congress, in Istanbul, Turkey.

After completing his military service in 2008, he worked as an architect from 2008 to 2012, collaborating with architectural offices in Athens on large-scale building and urban design projects in Greece and other European countries. His work has been exhibited in several international exhibitions. In 2009, he began pursuing a Postgraduate Master's degree (MPhil) in Architecture & Spatial Planning at NTUA. He graduated *summa cum laude* in 2011, ranking first among his graduating class, after defending his MPhil thesis on generative methods in urban design and planning. In 2011, he also received the Thomaides Award for Science and Arts for his work on computational approaches to designing urban public spaces. Between 2010 and 2012, Achilleas was an Assistant Lecturer at the School of Architectural Engineering, NTUA. In this capacity, he tutored advanced architectural and urban design studios, gave lectures on computational design, and coordinated several international workshops on urban systems and planning, with students from Parsons School of Design (USA), Université Paris Malaquais (France), NTUA (Greece), TU Dortmund (Germany), among others.

In 2012, Achilleas was awarded a prestigious scholarship from the A. S. Onassis Foundation, as well as a scholarship from the Greek State Scholarships Foundation (IKY) and the European Social Fund of the EU, to conduct PhD research at Delft University of Technology (TU Delft) in the Netherlands. He also received individual research grants from the Foundation for Education and European Culture (2012 – 2016) and the A. G. Leventis Foundation (2014 – 2016). The main goal of his PhD thesis was to design and develop a framework of novel methods and tools for data integration, visualization, and exploratory analysis to understand the spatiotemporal dynamics of human activity in cities. During his doctoral studies, he tutored several Master's courses, coordinated a number of workshops on urban science and informatics, gave lectures on similar subjects, presented his research results in various conferences worldwide, and supervised the thesis of a graduate student in Computer Science. His research on data integration and interlinkage (presented in Chapters 3 & 4 of this thesis) has been awarded the 1st Prize for Linked Open Data for Smart Cities (2015). Since 2014, Achilleas has been collaborating with the Web Information Systems group and the Delft Social Data Science Lab of TU Delft on the development of the *SocialGlass* platform (presented in Chapter 6 of this thesis) and a number of projects at the intersection of data science and urban analytics. Results of this work have been extensively presented in conferences and research exhibitions. Achilleas was also a researcher at The Why Factory (2012 – 2013), the urban think tank and research institute for future cities between MVRDV and TU Delft, and a member of the Joint Research Center for Urban Systems and Environment (USE) between TU Delft and the South China University of Technology.

Achilleas completed his PhD research in July 2016. From September 2016, he is a Postdoctoral Researcher at the Web Information Systems group, Delft University of Technology (TU Delft), and at the Amsterdam Institute for Advanced Metropolitan Solutions (AMS), where he is also the project leader of the Social Urban Data Lab (SUDL). His research interests include urban data science, spatial analysis, geographic information science (GIScience), machine learning, multilayered geo-social networks, linked urban data, the semantic web, urban systems and social flows.

Contact: A.Psyllidis[at]tudelft.nl | apsyllidis[at]gmail.com

List of Publications

Psyllidis, Achilleas. (2015). *Ontology-Based Data Integration from Heterogeneous Urban Systems: A Knowledge Representation Framework for Smart Cities*. In: Ferreira J Jr, Goodspeed R (eds) Planning Support Systems and Smart Cities: Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015). MIT, Cambridgre, MA, USA, pp. 240-1 — 240-21.

Psyllidis, Achilleas, Bozzon, Alessandro, Bocconi, Stefano, & Bolivar, Christiaan-Titos. (2015). *Harnessing Heterogeneous Social Data to Explore, Monitor, and Visualize Urban Dynamics*. In: Ferreira J Jr, Goodspeed R (eds) Planning Support Systems and Smart Cities: Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015). MIT, Cambridgre, MA, USA, pp. 239-1 — 239-22.

Psyllidis, Achilleas. (2015). *OSMoSys: A Web Interface for Graph-Based RDF Data Visualization and Ontology Browsing*. In: Cimiano P, Frasincar F, Houben G-J, Schwabe D (eds) Engineering in the Web in the Big Data Era: 15th International Conference on Web Engineering (ICWE 2015). LNCS 9114, Springer International Publishing, Switzerland, pp. 679—682. doi: 10.1007/978-3-319-19890-3_56

Psyllidis, Achilleas, Bozzon, Alessandro, Bocconi, Stefano, & Bolivar, Christiaan-Titos. (2015). *A Platform for Urban Analytics and Semantic Integration in City Planning*. In: Celani G, Moreno Sperling D, Franco JMS (eds) Computer-Aided Architectural Design Futures – New Technologies and the Future of the Built Environment: 16th International Conference (CAAD Futures 2015) – Selected Papers. LNCS, CCIS 527, Springer, Berlin Heidelberg, pp. 21—36. doi: 10.1007/978-3-662-47386-3_2

Bocconi, Stefano, Bozzon, Alessandro, **Psyllidis, Achilleas**, & Bolivar, Christiaan-Titos. (2015). *SocialGlass: A Platform for Urban Analytics and Decision-making Through Heterogeneous Social Data*. In: Gangemi A, Leonardi S, Panconesi A (eds) 24th International Word Wide Web Conference (WWW 2015). ACM, New York, NY, pp. 175—178. doi: 10.1145/2740908.2742826

Psyllidis, Achilleas, & Biloria, Nimish. (2014). *OntoPolis: A semantic participatory platform for performance assessment and augmentation of urban environments*. In: 10th IEEE International Conference on Intelligent Environments 2014 (IE '14). IOS Press, Washington, DC, pp. 140—147. doi 10.1109/IE.2014.28

Psyllidis, Achilleas, & Biloria, Nimish. (2014). *From Big Data to Linked Data: Extracting and Interlinking Knowledge from the City through Semantic Web Technologies*. In:

Workshop on Smart Cities and Big Data International Research Workshop, Aarhus, Denmark.

Psyllidis, Achilleas. (2013). *SmartScapes: Big Data and Urban Informatics for Performative Cities*. ATLANTIS, 24(2):20—25.

Psyllidis, Achilleas, & Biloria, Nimish. (2013). *Urban Media Geographies: Interfacing Ubiquitous Computing with the Physicality of Urban Space*. In: Geiger, J., Khan, O., Shepard, M. (eds.) Media City 4: MEDIA CITIES, International Conference. The University at Buffalo, State University of New York, New York NY USA, pp. 302—309.

Psyllidis, Achilleas, & Biloria, Nimish. (2013). *The Adaptive City: A Socio-technical Interaction-driven Approach Towards Urban Systems*. In: Charitos, D., Theona, I., Dragona, D., Rizopoulos, H., Meimaris, M. (eds.) 2nd International Hybrid City Conference. University Research Insititute of Applied Communication (URIAC) & Klidarithmos Publications, Athens, pp. 371—378.