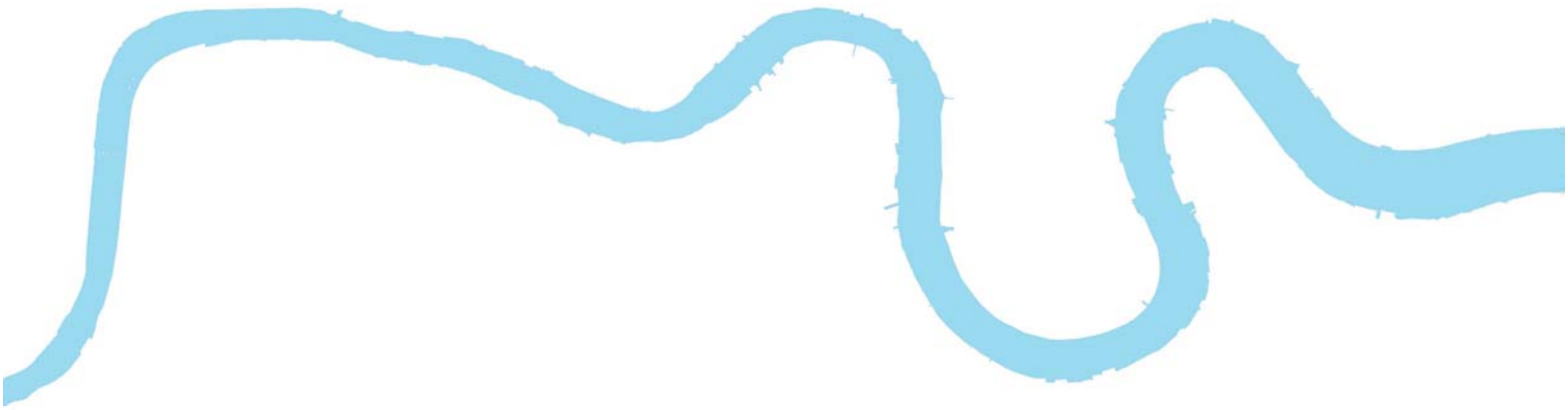
PhD Thesis

Azam Bahrehdar

Extraction of Place Descriptions from User-Generated Content



Extraction of Place Descriptions from User-Generated Content

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat)

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Azam Bahrehdar

aus

dem Iran

Promotionskommission

Prof. Dr. Ross Stuart Purves (Vorsitz)

Prof. Dr. Robert Weibel

Dr. William Mackaness

Zürich, 2020

Faculty of Science
University of Zurich

Extraction of Place Descriptions from User-Generated Content

Azam Bahrehdar
Geocomputation Unit
Department of Geography
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

"The stranger who finds himself in 'The Dials' for the first time, and stands Belzoni-like, at the entrance of seven obscure passages, uncertain which to take, will see enough around him to keep his curiosity and attention awake for no inconsiderable time. From the irregular square into which he has plunged, the streets and courts dart in all directions, until they are lost in the unwholesome vapour which hangs over the house-tops, and renders the dirty perspective uncertain and confined; and lounging at every corner, as if they came there to take a few gasps of such fresh air as has found its way so far, but is too much exhausted already, to be enabled to force itself into the narrow alleys around, are groups of people, whose appearance and dwellings would fill any mind but a regular Londoner's with astonishment.

On one side, a little crowd has collected round a couple of ladies, who having imbibed the contents of various "three-outs" of gin and bitters in the course of the morning, have at length differed on some point of domestic arrangement, and are on the eve of settling the quarrel satisfactorily, by an appeal to blows, greatly to the interest of other ladies who live in the same house, and tenements adjoining, and who are all partisans on one side or other."

- Charles Dickens, Sketches by Boz

ACKNOWLEDGMENTS

Carrying out this PhD at the University of Zurich's Department of Geography in Switzerland has been a truly life-changing experience for me. It would not have been possible without the support of many people whom I owe a great deal of gratitude for their help and support throughout my research. First and foremost, I would like to express my sincere appreciation to my supervisor, Prof. Dr. Ross S. Purves, for his consistent kindness, encouragement and scientific guidance along with a tremendous dose of patience, especially throughout the scientific writing process. Your support and belief in me encouraged me to carry on through the last years and enabled me to succeed and achieve my goal. I also thankful to my PhD committee members Prof. Dr. Robert Weibel and Dr. William Mackaness for their advice, constructive feedback and scientific experience.

I am very grateful to my friends and colleagues in the Geocomputation group for their continuous support and feedback. I am especially grateful to Olga for all the brainstorming sessions and discussions during my research and for her valuable feedback on my thesis. I am also thankful to Isabela, Katja and Flurina who helped me during the first phase of my PhD-life in Switzerland.

I would like to thank all past/present members of the GIScience and GIVA groups, with whom I have shared moments of deep anxiety but also of great excitement. I would like to thank Peter and Alex, whose support coated with their unique sense of humour took me through tough times during the last two years of my PhD. I would also like to thank Julia and Kiran, who always had a word of encouragement when dealing with all sorts of challenges. Also, a warm word for my great friend Michelle, who always managed to make me feel special and with whom I had the best sports breaks in my life!

I am grateful to my lovely family (my great grandparents, mamani, baba, Firouzeh, Parvaneh, Roya, Nahid, Soudabeh, Jalal, Javad, Siamak, Parham, Nazanin and Arousha) and friends (Peymaneh, Ali, Shima and Homi) for their great love, support, encouragement and patience. They never showed their surprise when I claimed my thesis would be finished 'in the next two months' for nearly a year, especially Annica, Saloumeh and Sepide.

Finally, I have to thank my husband and love of my life, Yashar, for keeping things going and encouraging me to do my best.

Zurich, December 2019

SUMMARY

The emergence of digital social data produced by an increasingly large number of people, in parallel with technological advances in the field of information retrieval, has led to a discovery that such fine-grained data on urban spaces (such as user-generated content) can provide significant insights about semantics ascribed to a particular place. An important challenge for any work attempting to operationalise place is the lack of definition of place. Researchers also need to face challenges regarding limitations and biases in user-generated content and challenges in aggregating data contributed by multiple people.

The overall objective of this research was to explore meaningful ways of capturing places through descriptive information extracted from user-generated content based on concepts of place. To achieve this objective, we set out three smaller goals: first, to explore studied dimensions of place to reflect possible implications with respect to the limitation of UGC as well as challenges of operationalising place (Publication I); second, to generate a continuous model that infers characteristics of places from georeferenced textual information (Publication II); and third, to characterise a city through user-generated content and based on a conceptual model to reflect perceived semantics of a city (Publication III).

In terms of methods, Publications I-III aimed to improve overall understanding of the potential and limitations of user-generated content (in the form of metadata attached to Flickr images) to characterise places. We investigated the implications of explored dimensions of place in literature for location-based services (Publication I) based on a list of application categories. To model and characterise places we used two approaches. First, we applied a purely data-driven approach whereby we aggregated our data using a grid network; furthermore, we employed an unsupervised classification method to compare grids with respect to linked textual information (Publication II). Second, we reinforced a conceptual geographical model: a street network that allowed us to reflect an "image of the city" representing people's perceptions of their experienced surrounding. We did so by measuring similarities between streets based on three place dimensions: semantics, users' behaviour and time (Publication III). We also used viewsheds of locations to model places with respect to a point of interest and produced their descriptions using three categories of place descriptions: elements, qualities and activities (unpublished work). Moreover, we applied a sensitivity analysis to assess to what extent places and their semantics are perceived as place descriptions by humans. We did so by studying the influence of inputs on our place model (Publication II).

In our results in Publications I-III, we showed that all different dimensions of place at three levels ¬– "specific of", "generic of" and "about" – can be extracted from metadata attached to Flickr images. In Publication II, we generated a continuous model in which all places can be described through topics. We were able to label the topics based on consisted tags and assign them to one of the four categories of place properties (location, activity, locale, and people). Through our sensitivity analysis, we showed that labelled topics have higher coherence values which can be considered as a predictor of the likelihood of humans being able to interpret topics. In Publication III, different dimensions of place (such as semantics, user behaviour and temporal aspects) were explored through similarity patterns between streets. Through four examples, we showed that streets are indeed natural units for capturing perception of cities. We modelled the city through paths and also could emerge other elements of the city such as districts, landmarks and edges. Our place model based on the visibility concept enabled us to find places that are related based on viewshed analysis.

There is great potential for our results in different scientific domains including GIScience, location-based services, urban planning, and map production. The descriptive information in user-generated content provides GIScience a collection of data contributed by multiple users which covers large scales and contains people's experiences (related to, for example, their activities in and their attachments to a place). Applied methods in this work can be used to model places such as "the image of a city" and to extract semantics related to locations, which can help us to operationalise place and offer the possibility of reasoning with place. Therefore, we believe that focussing on integrating different sources to capture notions of place can provide a better understanding of how people experience and perceive their surroundings.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

SYNOPSIS

1

INTRODUCTION

1.1 MOTIVATION

Think about what it would like to be lost in a city as a newcomer. You would start asking questions like: "Which way I should go? Where am I at all? Should I be worried that I am wandering around?" Having a positioning device (e.g. a mobile phone or a watch equipped with GPS) or a map in which more geographical context (e.g. place names) is given, you would be capable of comparing your current location and distance to potential target places at the level of geometry.

Knowing potential targets and possible paths towards each destination can help us with the first two questions (e.g. localising ourselves or identifying paths leading to a target location), but the last question remains unanswered. An usual map representing place names on a base map can provide us not only general information about the environment (e.g. railways and parks) but also what we should expect. For example, a path across Hyde Park conveys different feeling than a path along a railway.

Place names do not necessarily reflect the nature of the place itself (e.g. Whitehall street). In such cases, providing semantic information about places to users, which is not a simple task, becomes more important, since each location can have different, and potentially contrasting meanings for different people. For example, in spite of tourists' interests in visiting historic places, they prefer access to paths that include places attractive for specific user groups. (For example, families with children are interested in places that afford various kinds of activities like restaurants, zoos and beaches.)

A possible solution these days is to have maps or services to collect information from social media (e.g. in the form of images, descriptions, or ranked points of interest) about lived and experienced locations, and to provide specific or general information about different dimensions of a place, for example, visual appearance (e.g. how scenic a path is), social or cultural aspects of the populace, or information related to the crowdedness (e.g. popularity of the landmarks). It is a complex and time-consuming task to aggregate disparate documents describing a place. For instance, a person from the countryside's descriptions of a metropolitan city are different from a person from the city who is used to the noise and pace. Therefore, if a single group of people (e.g. country people or

townspeople) are the dominant users, it is more likely that the content is biased by its contributors and the platform reflects their perception of the city.

People's shared experiences of a place might be related to multiple and possibly different aspects, yielding a variety of meanings for a given place. For example, georeferenced photos taken in the middle of a forest can represent surrounding trees (what happens around the location itself) or portray mountains in the distance (reflecting the location's relations with other locations). Sometimes, the shared experiences are linked wrongly to the same place, since they are assigned to a place name which is ambiguous; for example, in two text entries each describing a place called "London", one is related to the city in United Kingdom and another one the city in Canada. Another reason for such mismatching is because of vagueness of location of a place; for example, the extent of "downtown" varies among different users. Therefore, linking descriptions to existing places provides an opportunity to investigate how places are semantically distinctive and varied among multiple user groups and communities. The overall objective of this thesis is to explore meaningful ways of capturing places through their descriptive information extracted from user-generated content, and by which different application domains can benefit from.

Phenomenological studies on place that focus on interpreting human experience under the level of conscious awareness can elevate geographic information systems (GIS). Therefore, instead of systems that model space with reference to coordinates, we can have systems that integrate human experience with spatial information. Such advanced systems represent the world as "meaningful locations" perceived by people [*Cresswell*, 2014]. Despite efforts to study concepts of place in geography in the last two decades, the term "place" has been used mostly in GIScience as a shorthand for locations or bounded regions. Place also has been described through limited properties, such as different place names of a region or place names associated with vague regions [*Montello et al.*, 2017].

Studies on place can be used extensively in different branches in GIScience [*Merschdorf and Blaschke*, 2018]: for studying the impact of spatial features (e.g. neighbourhoods) on people's behaviour (e.g. their preferences) in qualitative GIS; for identifying popular places through volunteered geographic information (VGI) to enable local participation GIS by using place as context for individuals' space-time analysis in location-based services (LBS); or for studying people's interactions with their spatial surrounding to inform spatial management and planning. In spite of a broad range of applications, progress is limited because of the difficulties in capturing and modelling notions of place in an unambiguous way that can be represented by computer systems (based on binary concepts, which require crisp definitions of represented features) [*Merschdorf and Blaschke*, 2018].

Another reason for the lack of progress in both to operationalise place and to integrate relevant information into systems such as GIS or LBS, could be the lack of an agreed definition of place. Places are connected and related to each other like a network [*Massey*, 1994], and not only their names but also their boundaries can be inherently vague. The meaning of place is also subject to change with respect to the time, scale, or actors who are experiencing the place. In fact, place is highly dynamic and contextual, and results of analysing place semantics are attributed to change based on the context [*Goodchild*, 2011].

Natural language is suggested as an artefact to study and understand human conceptualisations of the notions of place [*Bennett and Agarwal*, 2007]. Place and related concepts are commonly used in everyday communication about our geographical environment. For example, anytime we mention the location (i.e. positions or coordinates) of our daily practices, we either use their place names (e.g. London or city centre) or give a reference to a known location (e.g. a cinema near the train station). The term "place" can refer to locations varying in scale [*Cresswell*, 2014](e.g. a specific chair in a cafe, a building called home, or the whole planet), and a particular location can have multiple meanings for different groups of people (e.g. Starbucks as a workplace or as a place to enjoy friends' company) [*Davies et al.*, 2008]. Regions can either be vague (e.g. downtown) or well-defined like administrative areas (e.g. City of London), disregarding possible different perceptions of their boundaries [*Hollenstein*, 2008]. A GIS system that attempts to deliver geographically relevant information, especially where query or results are text based and use natural language, should be capable of dealing with such notions of place.

Natural language text (e.g. historical archives or news articles mentioning placenames or explicitly geotagged Wikipedia pages) is one way of identifying places and building place descriptions. In recent years, a broad range of methods (e.g. named entity recognition or sentiment analysis) and tools (e.g. MALLET Toolkit for topic modelling) have been developed and employed in the field of geographic information retrieval. For example, to study place names, we can explore possible links between place names and properties of types of places that the names label. We can also explore the geographic footprint of a place name to recognise its boundary [e.g., *Hollenstein and Purves*, 2010; *Kelm et al.*, 2013; *Vögele et al.*, 2003; *Jones et al.*, 2001; *Vasardani et al.*, 2013]. It is important to point out that automatic extraction of semantically and contextually relevant information from unstructured text is a long-standing task, especially in case of place-related information, since it requires facing the vagueness involved in both natural language and the concept of place being communicated by a language.

User-generated content (UGC) has increasingly drawn attention in GIScience due to its geographical element —ranging from volunteered geodata on OpenStreetMap.org to georeferenced unstructured text in the form of travel blogs or Wikipedia pages; georeferenced and tagged images on Flickr.com or Instagram;

location check-ins or reviews on social media sites such as Foursquare or Yelp; or microblogs in the form of Twitter. Hence, UGC as an optional and complementary data source has been used to explore conceptualisations of place as a lived and experienced location [*Arampatzis et al.*, 2006; *Lansley and Longley*, 2016]. A key potential of UGC is that a large number of contributors as individual citizens produce content reflecting contextual aspects of multiple perspectives on a place [*Goodchild*, 2011]. However, emerging descriptive information involves challenging tasks such as (1) capturing all different aspects of a place, if it is possible at all, (2) aggregating and summarising various, potentially contrary, opinions into coherent themes, and finally, (3) removing biases generated in the process of data production [*Haklay*, 2016].

Extensive investigations on place have been done from both conceptual or operational perspectives, however most of these studies have not been placed into a framework that can be used by others. In the course of this dissertation, I explain how multiple aspects of place-like locations have been explored through UGC and how representative descriptions have been extracted by addressing limitations of UGC data that minimises the impact of biases on our results. Moreover, I discuss each approach and places dimensions in the context of applications.

1.2 STRUCTURE OF THE THESIS

This dissertation consists of two complementary parts. Part I (Synopsis) provides a detailed overview of the research carried out in the scope of this thesis, additional to unpublished results. Following the introduction, Chapter 2 (Background) provides a summary of the information necessary to understand the current state of research, and the research gaps that led to the conducted work. Chapter 3 (Methodology) provides an overview of datasets used in empirical analysis, a detailed summary of the main steps of data preparation, and finally, the methodological approaches towards modelling place. The main findings from Publications 1-3 and unpublished work are thematically presented in Chapter 4. Chapter 5 (Discussion) discusses the quality of UGC data and their limitation with respect to applications as well as the characteristics of extracted place-based information and foreseen challenges. A summary of contributions and insights gained in the thesis, and an outlook of future research are given in Chapter 6. Part II (Publications) consists of the three research papers written over the course of this dissertation:

Publication I: Approaching location-based services from a place-based perspective: from data to services?

Bahrehdar, A. R., Koblet, O., and Purves, R. S. (2019), Approaching location-based services from a place-based perspective: from data to services?. Journal of Location Based Services, 1-21.

PhD candidate's contributions: Developed research ideas and annotating papers in collaboration with co-authors. Authored the main categorisations section and incorporated several rounds of feedback from the co-authors. Wrote the draft manuscript and co-authors' feedback.

Publication II: Description and characterisation of place properties using topic modelling on georeferenced tags.

Bahrehdar, A. R. and Purves, R. S. (2018). Describing and characterising place using topic modelling on georeferenced tags. Journal of Geo-spatial Information Science: Special Issue on Crowdsourcing for Urban Geoinformatics, 21(3):173-184.

PhD candidate's contributions: Developed research ideas in collaboration with co-authors. Conducted data processing and analysis in Java and Python. Wrote the draft manuscript and incorporated co-authors' feedback.

Publication III: Streets of London: Using Flickr and OpenStreetMap to build an interactive image of the city.

Bahrehdar, A. R., Adams, B., and Purves, R. S. (2019). Streets of London: Using Flickr and OpenStreetMap to build an interactive image of the city. Computers, Environment and Urban Systems (submitted).

PhD candidate's contributions: Developed research ideas in collaboration with co-authors. Gathered and processed data and analysed parts in Java. wrote the draft manuscript.

2

BACKGROUND

In this chapter, we present the background on relevant concepts and developed methodological approaches in two parts:

Notions of place: We introduce different ways of conceptualisations of place in a broader context, beginning with differentiating the term "place" from its competing term "space". We furthermore, continue with explaining general conceptual approaches in geography focussing on studying and discussing different ways of characterising a place with respect to different perspectives.

Place and GIScience: Secondly, we will focus on place-related studies in the context of GIS. We will introduce the theoretical and computational models of place introduced in GIScience. Furthermore, we will discuss approaches, in terms of data and applied methods that have been used UGC to formalise developed models in information systems. We then, will highlight both methodological challenges in formalising notions of place and challenges with respect to existing biases in data influencing on representativeness of extracted information. Finally, we introduce research gaps and research questions.

2.1 NOTIONS OF PLACE

2.1.1 *Place vs. space*

Place was originally discussed in classical Greek philosophy as "the starting point for all other forms of existence", but the concept of place as "a meaningful segment of geographical space" was only formed in the 1970s [*Cresswell*, 2014]. In an abstract and purely geographical view, place appears as a location with certain properties that distinguish it from space: places are contained within space. In essence, when moving to a new and unfamiliar “part of the world”, one is moving to a “specific part of space”, and only in the process of living, interaction with social system and environment, place is created [*Schneider*, 1987].

According to *Longman Dictionary of Contemporary English*, space is defined as a continuous area that can be free to use or occupied, and place is defined as "a particular point on a surface". Notions of place have been discussed in relation to space, since both place and space provide information about "where" things happen at a particular time [*Agnew*, 2011]. According to Agnew, what differentiates these two fundamental and contested concepts is "their relative invocation that has usually signalled different understandings of what ‘where’

means" [*Agnew*, 2011, p. 1]. Hence, it is best to study space and place together. Tuan, alternatively, conceived place in contrast with space [*Tuan*, 1977]. In his view, space is an unrestricted open environment, which allows to move between pauses. Places, conversely, are the parts loaded with human meaning.

2.1.2 *Approaches towards handling place in geography*

Place can be represented as an object that can be observed and studied, or as a "way of looking", which influences the ways we do research about other things. Thus, there are several approaches in geography to handle place [*Cresswell*, 2014]:

REGIONAL CONCEPTION OF PLACE: Abstract spatial analyses are used to recognise distinctive properties of regions by exploring particular combinations of natural environment (e.g. climate or soil type) and cultural forms (e.g. food and clothing). Here, places are not objects to be found; they are regions that are formed through chorological observations shared between different locations [*Harvey and Wardenga*, 2006]. This approach deals with drawing boundaries for regions based on similar natural properties [*Herbertson*, 1905] or human characteristics [*Fleure*, 1919], which reflects the importance of meaning in a given location.

PHENOMENOLOGICAL WORK ("A WAY OF UNDERSTANDING"): Human geographers like Tuan (1974) and Relph (1976) draw attention to the relationships and interactions between geography (especially "place") and, people. The research on complexity and depth of place was done through understanding different ways, in which people experience places (e.g. [e.g., *Relph*, 1976; *Tuan*, 1974]. Place as a way of understanding provides a context for all activities that humans do, and it is connected to individuals' memories and feelings [*Cresswell*, 2014; *Tuan*, 1977].

Relph (1976) highlighted three fundamental elements of a place: (1) its physical setting, including location and physical appearance; (2) the influence of physical materiality on people's actions and activities; and (3) the meanings attached to a place. From a psychological point of view, place can be recognised and characterised by a physical configuration that has an objective with respect to individual or social and cultural aspects of the place, affording various functionality (different activities people do) in different scales (the granularity of a place; room, building or a city) [*Canter and Groat*, 1977]. A fairly different psychological model argues three parts for identifying a place: the self (e.g. personal meanings and self-identification), others (related to social relations and the norms), and the environment (the physical characteristics and natural conditions) [*Gustafson*, 2001].

SOCIAL CONSTRUCTIONIST PERSPECTIVE ("A GLOBAL SENSE OF PLACE"):

Unlike the former approach, from a constructionist point of view, places are not conceived as permanent fixtures of space or remote containers where social interactions occur, nor are they only linked to the local [*Harrison and Tatar*, 2008]; places are networked space through the mobility of people or goods.

Massey (1994) presented a "progressive sense of place" and understood places as social constructs emerging from communities as ways of looking at, talking about and understanding the world. In this perspective, places are formed in the course of the movements of people and commodities. Places are not bounded, but are rather connected to the rest of the world and inherently relational and connected with people from other places at a global scale [*Cresswell*, 2014; *Massey*, 2012], which creates their heterogeneous identities.

Reviewing three approaches allows us to understand the complexity of concept of place and to identify shared characteristics of notions of place that are emphasised in each approach: (1) places are social products, in which a sense of place forms through the course of interaction with environment and people who live there; (2) places are subject to change because of dynamic characteristics of place, since they are connected to the other places; and (3) properties or qualities of places are diverse, even within their vague boundaries.

To go beyond conceptual work in human geography and study place from a practical point of view that allows us to identify and formalise dimensions of place and associated properties based on conceptual frameworks, we need more details about place. Based on an experiment asking residents to sketch their perceptions of a city, Lynch (1960) identified five elements used to represent their environment: **(1) paths** ("channels along which the observer customarily, occasionally or potentially moves"); **(2) nodes** ("strategic spots in a city into which an observer can enter ... and from which he is travelling,"); **(3) districts** ("medium-to-large sections of the city ...which the observer can mentally go inside of, and which have some common character."); **(4) edges** ("linear elements not considered as paths by the observer ... which are not only visually prominent, but also continuous in form and impenetrable to cross movement"); and **(5) landmarks** ("considered to be external to the observer ... the key physical characteristic of this class is singularity, some aspect that is unique or memorable in the context."). Each element can be defined as an aggregated location in the city [*Winter and Freksa*, 2012], which can be localised in space. Winter and Freska (2012) discussed how these elements combined with spatial prepositions often appear in place descriptions (e.g. on Kilburn High Road).

Agnew (2011) suggested a conception of place in geographical context by localising place in space, and categorised the nature of place into three elements *location*, *locale* and *sense of place*. Based on his proposed model, a place is a spe-

cific location with a name that has a physical setting (rooms in a building or streets and parks in a city, etc.) that facilitates everyday activities and social interaction for both individuals and groups. Within this setting, people develop emotional attachments to the place and its elements. Harrison and Tatar (2008) point out the importance of the contribution of two elements, people and events, in notions of place. They represent place as a particular context of people (actors who were involved in the course of place-making process), events (all the activities ranging from ringing a phone to having dinner), and loci (elements facilitating "place-meaning-making") as a semantic tangle.

Place dimensions of Agnew's model parallels with dimensions of *where*, a facet of the Pansofsky-Shatford matrix introduced in information science [*Shatford*, 1986]: *specific of* (related to named places or instances of places), *generic of* (properties or features of places), and the *about* (associated emotions and feelings). The matrix was originally developed to classify contents of art collection (containing images), and was furthermore used in broader context, for example in annotation tasks. The similarity between a place model and a model in information science, which has been widely used, suggests an opportunity for fusing the two as place models in information science.

2.2 PLACE IN GISCIENCE

This section presents a broad spectrum of place-related research in GIScience: from purely theoretical discussions and conceptual models on one end to computational models of place and operationalisations of single characteristics of a place on the other end. In the following, we first summarise efforts in providing conceptual frameworks for practical studies on place. We then, give a detailed overview of work focussing on operationalising place properties, particularly using UGC.

2.2.1 *Conceptualisation and formalisation of place*

Work from conceptual perspective that concentrates on deriving general models of place typically starts from the literature in human geography and some other fields (see section 2.1) that focus on describing a conceptual data model suitable for dealing with place in information systems. One common theory that has been used for conceptualisations of place is related to affordance theory. Affordance deals with how people perceive their environment based on existing "objects or things" or "activities" that the environment afford [*Jordan et al.*, 1998]. Affordance of a place can be realised only by looking at the things [*Gibson*, 1977] or through the course of cognition [*Norman*, 1988], in which individual experience matters. An early work based on the concept of affordance in the sense of Gibson (1977) is a methodology that models places in different scales —from an office to a city, within which individuals' experiences take

place [*Jordan et al.*, 1998]. This conceptualisation focusses on different subjective affordances of a place from an individual view (the agent's capability reflecting knowledge about users), the environment (e.g. services offered by a restaurant), and the task requirements (e.g. "stability" for a table-like object for eating lunch).

According to Gibson (1977), one of the properties of a place is its accessibility, which represents place as both a "container" in which objects and events are located, and as a "surface" on which the movement of elements occurs. Several models have been developed based on this accessibility property. While Kuhn (2001) used an experiential view to model a system of entities and affordable actions assuming activities as key to the context, Jorgensen et al. (2001) went one step further and focussed on a summary of experiences as "sense of place" to propose a multidimensional measure to compare places based on three subcategories: identity, attachments and dependence.

Scheider and Janowicz (2010) developed a model based on Jordan et al. (1998) and perceived places as a sub-category of affordance in the sense of Gibson, "perceivable action potentials in the meaningful environment of an observer". Regarding the place as a medium supporting its element's movements within a certain spatial relation to an identifiable piece of surface, they argue that their approach provides an insight into the ways in which places can be categorised and identified, and therefore, offers a robust basis for geo-ontologies.

Place is also conceptualised within natural language: for example, Tversky and Hemenway (1983) classified terms used to describe scenes in natural language. They used a subordinate categorisation of outdoor scenes (i.e. basic levels) such as "mountain", "beach", "park" and "city" [*Rosch and Lloyd*, 1978] and classified the associated words into three shared themes: parts, activities, and attributes. Furthermore, Winter and Freksa (2012) used cognitive concepts and language to capture notions of place through contrast. They argued that the five perceivable elements of a city, referring to Lynch's concept (1960), are typically used in human communication as reference points that characterise a place together with spatial prepositions. Their methods have been considered as the best approach [e.g., *Merschdorf and Blaschke*, 2018] to localise place names without considering their precise locations and geographical boundaries.

2.2.2 *Operationalisations of place properties and UGC*

The advent of social media has provided an opportunity to explore broad ranges and volumes of fine-grained data. Such data is the inspiration behind studies motivated by notions of place, which are often limited to operationalising a few attributes of a place [e.g., *Derungs and Purves*, 2016]. In the last decades, work conducting extraction of place-related information from UGC, or more specifically Volunteered Geographic Information (VGI), has increased. This work,

summarised briefly in the categories below, mostly aims at reflecting notions of place as lived and experienced parts of space [e.g., *Jenkins et al.*, 2016; *Lansley and Longley*, 2016; *Capineri*, 2016; *Hauthal and Burghardt*, 2016; *Shelton et al.*, 2015].

STUDIES RELATED TO THE NAMED PLACES AND INSTANCES OF PLACES: These studies apply contrasting approaches to explore different aspects of specific places by (1) depicting regions associated with names representing places, e.g. `hydepark`, `regentspark` [*Hollenstein and Purves*, 2010]; (2) deriving cognitive regions, e.g. historic centre of Vienna, or "NorCal" referring to North California [*Hobel et al.*, 2016; *Gao et al.*, 2017]; (3) investigating different ways that people use *vernacular place names* [*Hollenstein and Purves*, 2010], since place names that people use to communicate about their surrounding are not necessarily identical to the ones presented in administrative gazetteers (e.g. `downtown`); and (4) providing knowledge about boundaries of *vernacular regions*, which are more likely to be associated with place names at various scales.

STUDIES CONCENTRATED ON PROPERTIES OR FEATURES OF PLACES: These studies characterise cognitive regions by, for example, generating thematic characteristics extracted from text on social media or by identifying lists of words (known as topics). Drawing on an example from California, words are associated with areas perceived [*Gao et al.*, 2017]. Figure 2.1 demonstrates topics and their words characterising California which are dominated by outdoor physical features (e.g. desert, park and beach). The font size in word clouds represent the probability that words belong to a topic, and therefore, indicate how well a word describes the assigned area. Topics are a list of words ranked based on their probability of belonging to the topic. These ranked lists of words have been applied to characterise locations with a specific theme.

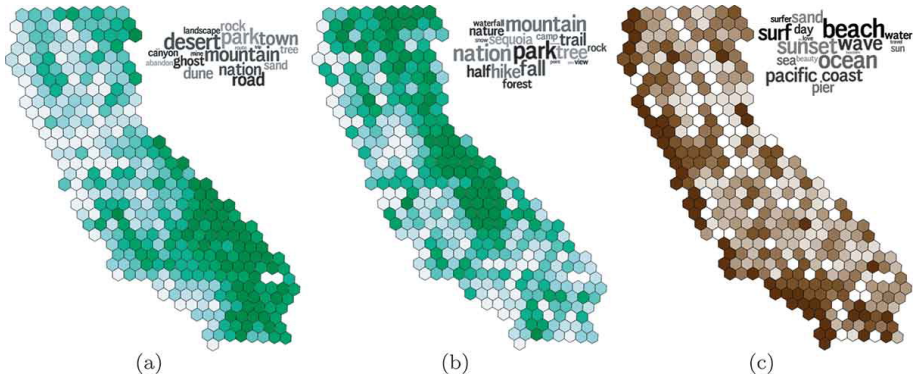


Figure 2.1: Three topics mapped to California along with their related word clouds. The darker the chromatic hue, the more prominent are the topics of terms in the postings from a particular cell [*Gao et al.*, 2017]

For example, Figure 2.2 illustrates a word cloud that has "war" as the top ranked word in the topic as well as regions associated with the topic. Using contours, they could calculate the degree to which a topic is related to identified regions.

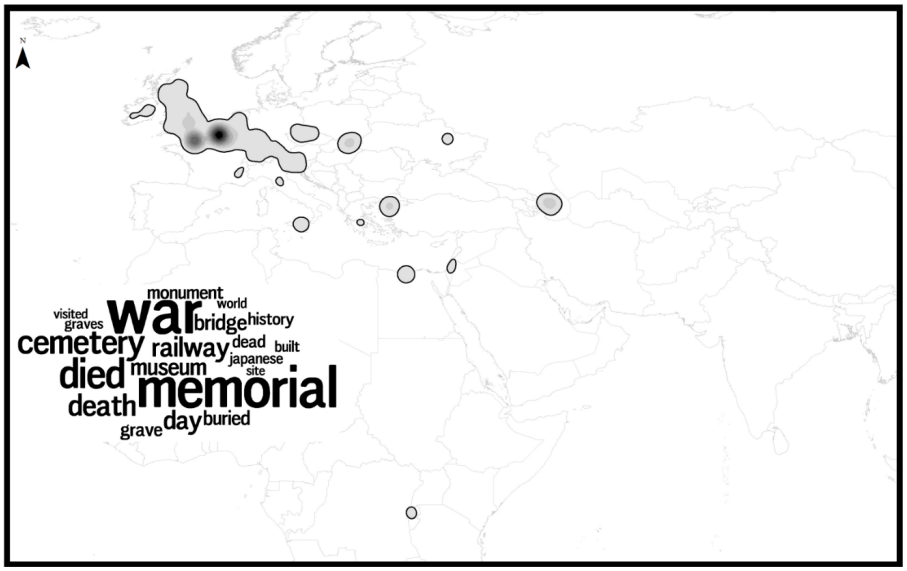


Figure 2.2: Regions associated with a topic that has the word "war" as the top ranked word [*Adams and McKenzie*, 2013]

Capineri (2016) also investigated different aspects of a place related to objects and activities (e.g. urban functions or services) using the theoretical framework of Agnew (2011). Dunkel (2015) also represented particular urban places with respect to scenery and infrequent or cyclic events by calculating frequency of terms, while Derungs and Purves (2016) focussed on natural landscape (e.g. forest, mountain, and ridge). Giving a vector that reflects prominent natural features shows how locations, in terms of grid cells, can be compared.

WORK RELATED TO FEELINGS AND EMOTIONS ASSOCIATED WITH PLACES: Generating maps that represent people's preferences for particular places is a common approach in such kinds of studies. Going one step further, different techniques or methods have been used to interpret associated semantics. For example, using a relatively simple visualisation techniques like word clouds [e.g., *Adams and McKenzie*, 2013; *Dunkel*, 2015] to demonstrate related words and perhaps their frequency of use. Another common approach is using a simple count of the number of users or photos, for example in non-urban areas, that can represent cultural values (e.g. enjoyment or social values) and capture an abstract notion of cultural ecosystem services [*Gliozzo et al.*, 2016]. Chesnokova, Nowak, and Purves (2017) investigated an abstract concept of place related to aesthetes and modelled landscape preferences through rates given to images.

Exploring such abstract notions of place can be done through linking emotions and feelings to locations in urban areas with respect to time (e.g. season or day of week) [*Hauthal and Burghardt*, 2016; *Resch et al.*, 2016] or physical setting of a place. Sentiment analysis, an advanced form of natural language processing, has been used to retrieve semantic information. For example, Lim et al. (2018) measured the degree to which a Tweet is negative or positive, and took one step further and explore the correlation between the nature of emotions (e.g. fear or joy) and different urban settings (e.g. parks and road junctions) based on a psychological theory. They demonstrated that typically fewer negative emotions are associated with green spaces compared with large transport infrastructures; however, these emotions are subject to change over time.

Shelton, Poorthuis, and Zook (2015) sought to understand places with respect to social aspects through people's mobility. However, interpretation appears as an inevitable step in these works. Sometimes, an abstract understanding of place could be perceived only through a high level of interpretation and contextual knowledge contributed by the analysts. Capineri (2016) also approved the complexity and challenge of analysing properties like feelings and emotions, which are not typically and clearly "expressed in single words like happy, unhappy, love or hate" [*Capineri*, 2016, p. 137], and are often a combination of several statements which expose the ultimate meaning.

The state of the art in GIScience demonstrates that different aspects of a place can be captured using UGC data – ranging from delineated regions associated with places and their membership values in terms of their associations with place names to thematic regions representing places that facilitate comparing place similarities, to linking simple counts to more abstract concepts like aesthetics. Despite existing data sources and methods developed for exploring dimensions of place, extracting semantics associated to abstract notions of place (i.e. sense of place [*Agnew*, 2011]) is highly subjective and requires elaborate interpretations.

2.2.3 *User-generated content*

According to the state of the art, UGC is a recent approach to capture the diversity of ways of experiencing and understanding places. One reason for increasingly focussed attention to UGC is the significant number of users who contribute to the content. UGC potentially offers a great variety of ways to describe a given place-like location. Such data are provided by four broad categories of social services:

PHOTO SHARING/HOSTING SOCIAL NETWORK SERVICES: Images and their attached metadata, have been discussed to open up the possibility of an

immediate and direct link to place [*Fisher and Unwin*, 2005]. Flickr and Instagram are two common sources of such data that have been explored with respect to place-relevant information in GIScience [e.g., *Hausmann et al.*, 2016; *Gliozzo et al.*, 2016; *Boy and Uitermark*, 2017]. Among the photo-sharing communities, Flickr gained much attention because of, arguably, straightforward access to public images and their associated metadata by implementing queries to an Application Programming Interface (API) and fetching both spatial (e.g. using a bounding box) and textual (e.g. using a term like "downtown") data. Flickr has been used to capture various conceptualisations of place with respect to types of visitors (e.g. tourists vs. locals) [*Straumann et al.*, 2014]; however, Instagram has a broader community and potentially a wider range of place descriptions [*Di Minin et al.*, 2015; *Gao et al.*, 2017]. In 2018, Instagram shut down its API to download public data [1].

Tagging systems that focuss on linking pieces of information (in the form of words) to the contents of images facilitate the process of indexing content. Tagging system make images more searchable and, therefore, visible [*Mountain and MacFarlane*, 2007]. Moreover, they make it possible to access a wide range of geographical information, both in the form of geometry and semantics (often dominated by place names) [*Rattenbury and Naaman*, 2009]. Exploring geographical footprints of images and their associated semantics demonstrate that geotagged images on Flickr (or other image-sharing platforms) are not randomly sampled; they represent popular places [*Crandall et al.*, 2009], portray events [*Davies*, 2007], reflect the basic level (e.g. building, city, and dog) [*Rorissa*, 2008], or indicate aesthetic aspects [*van Zanten et al.*, 2016]. Hence, tags provide sufficient information to generate meaningful descriptions for capturing different aspects of a location [*Dunkel*, 2015]. The lack of syntax in the list of tags (as a free list) [*Wartmann et al.*, 2018] relatively simplifies text analysis process but makes the disambiguation process more challenging. For example, the word *bank* in a list of tags (e.g. "london", "bank", "thames", "shopping") might refer either to "Thames River bank" or to a bank branch that affords monetary withdrawals.

Although a large number of Flickr images are geotagged [*Antoniou et al.*, 2010], there are some uncertainties about both accuracy and precision of coordinates and image locations, whether referring to the location of the photographer or the subject of the photograph [*Zielstra and Hochmair*, 2013].

MICROBLOGGING AND SOCIAL NETWORKING SERVICES: Twitter is a very well-known and popular microblog in research because of easy accessibility through an API. Unlike on Flickr, historical data are not easily available,

1 https://www.instagram.com/developer/

which limits the proportion of original Twitter datasets available to most researchers. Tweets are short (on average about 33 characters in English Tweets [*Rosen*, 2017]) and have a relatively simple language structure [*Dittrich et al.*, 2015] that covers a broad range of topics from different domains [*Go et al.*, 2009; *Kwak et al.*, 2010]. These topics have proven to be a suitable source for broad scale patterns to emerge, for example, social-spatial segregation in cities through language and users' mobility analysis [*Shelton et al.*, 2015].

Twitter recently disabled precise its geo-tagging option (in the form of coordinates). A large proportion of Tweets in the past did not have explicit locations, leading to shortcomings in fine-grained analysis. Attempts to address this limitation through georeferencing the Tweets typically fail at fine resolutions except in cases where sets of points of interest were selected [*Zheng et al.*, 2018]. Several other additional challenges need to be faced while using Tweets to extract semantics. Tweets may consist of bots that typically produce geocoded Tweets not related to human activity [*Chu et al.*, 2010; *Compton et al.*, 2014]; the uncertainty of the Tweets location, since users' location can be different from the location of content that interests them [*Hahmann et al.*, 2014]; tweets contain a high frequency of slang and typos [*Go et al.*, 2009]; and Twitter messages are also often composed by language mixing [*Hong et al.*, 2011].

RATING AND RECOMMENDING SERVICES: Today's mainstream of online sharing systems offers useful information about users' ratings and therefore, their preferences of shared items (e.g. in the form of reviews of books or locations) [*Kim and Yoon*, 2016]. Foursquare, Yelp and the now-defunct Whrrl [*Ye et al.*, 2011; *McKenzie and Adams*, 2017] are exemplary sources that have been investigated with respect not only to place properties, but also to place geometry, as points of interest are shared by/through the services. Since users share their location via "check-ins" that refer to an instance of a place type (e.g. a hotel, airport or restaurant), natural places are underrepresented. For example, McKenzie and Adams (2017) demonstrated that instances of beaches in Foursquare are typically officially designated public beaches. Comparing these sources with ones where content is spatially geotagged (through coordinates rather than locations of points of interest), the spatial footprints and related content might help to capture where beaches are and what people think about them.

ONLINE THEMATIC BLOGS: The last category of source of data are thematic blogs like travel blogs, TripAdvisor entries, Wikipedia pages and the Text+Berg corpus [e.g., *Adams and McKenzie*, 2013; *Hobel et al.*, 2016; *Gao et al.*, 2017; *Derungs and Purves*, 2014]. These provide unstructured texts, which require more complex methods to localise and link the content to specific places. However, many of these examples are already linked to places, for example, TripAdvisor entries and Wikipedia pages where con-

tent associated with specific locations is presented. They also present relevant information about the place. At issue here is the availability of such texts and their terms of use. Content from Wikipedia pages is freely available under an open licence. In contrast TripAdvisor content is copyrighted and only available under specific terms.

Studies have often taken a pragmatic approach to source data, selecting data sources that are both free to access and relatively easy to collect. These data sources have been used in different studies with various purposes, but they share similar characteristics: (1) heterogeneity of the nature of data, which provide the opportunity to explore various opinions through content with a relatively explicit link to places (e.g. geotagged images and reviews about a POI) or a more implicit link (e.g. Tweets or unstructured text in Text+Berg corpus, in which places names or some locative expressions are referring to specific locations); (2) inconsistency in granularities captured in such data, and therefore, (3) inconsistency in the scales of the places described, particularly in unstructured text and microblogs.

2.2.4 *Biases in user-generated content*

User-generated content has been used as a source to discover the knowledge of the crowds and answer questions related to people's experiences or opinions of/about the world with respect to, for example, locations or events. To be able to answer such questions, we need to be aware of the impact of the quality of datasets on our results. Olteanu et al. (2019) recently suggested a framework to identify different sources of biases in social media: population biases (related to user demographics, which might influence the representation of a specific population), behavioural biases (related to users' behaviour across the platform), content biases (any type of distortion in content due to user behaviour), redundancy (caused by any duplicates in datasets), linking biases (a kind of behavioural bias caused by different attributes of users' networks, for example, a user connection network that affects their behaviours), and temporal variations (caused by changes in user behaviours or the population over time). Olteanu et al. (2019) discuss how each bias and probable resulting distortions should be investigated with respect to research questions and research goals.

Earlier, Nielson (2006, p. 1) introduced the "90-9-1 rule" related to "participation inequality" in social media. This parallels the population bias introduced by Olteanu et al. (2019), where a large volume of content is produced by a small proportion of contributors. The impacts of biases related to participation inequality have been discussed in geographic information science. Several approaches have been applied to reduce/remove them [see *Van Mierlo*, 2014; *Haklay*, 2016; *Purves et al.*, 2011]. Authors often express their desire to collect information (or people's opinions) about locations that are shared among most of the users rather a small group of people, not information generated by prolific users or users who ran-

domly contributed to the content out of curiosity. Hollenstein and Purves (2010) identified two groups of people who produce large distortion in Flickr metadata, and filtered them out from their data-set: `testers`, who are assumed to be experimenting with the system and had only one single image in their profile, and `prolific users` whose contributions are a noticeable proportion of the data. Gliozzo, Pettorelli, and Haklay (2016) used the number of users instead of number of contributions to avoid the impact of prolific users with specific interest sharing many photos. These users were recognised as "outliers" by Olteanu et al. (2019).

Redundancy can occur in the process of bulk uploading, in which data points (e.g. images) with identical metadata (e.g. coordinates or tags) have been uploaded. Regarding Flickr photos, it has been argued that bulk uploads are typically tagged with a generic list of tags covering the content of all the images [*Senaratne et al.*, 2013a]. These tags hardly provide any specific information about each image.

There are some behavioural patterns in using tags in different platforms concerning all different communities involved, the themes and their services. For example, place names (e.g. London or San Francisco) are the most common tag in Flicker images [*Kennedy et al.*, 2007] and can be investigated through a relatively simple count of the tags. Individuals can influence the content by overusing a tag, and therefore reducing the functionality of a simple global tag frequency as a metric of tag representatives.

One way of reducing such biases is to visualise how a particular tag was used among all users and compare different patterns. Hollenstein and Purves (2010) generated tag profiles to study and then measure the popularity and representatives of unique tags within each dataset. Figure 2.3 shows tag profiles generated for "london", a very popular tag, which is commonly used by both prolific and non-prolific users in their dataset, and the tag "innercity", which is used by few users.

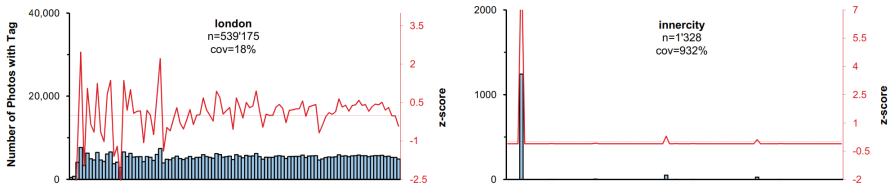


Figure 2.3: Tag profiles for “london” and “innercity” showing absolute tag counts and associated z-scores. The z-scores are indicated by lines; the histogram shows the absolute number of images with this tag ranked by contributor [*Hollenstein and Purves*, 2010]

The amount of produced content varies in space and follows geographic inequalities. For example, Graham, Hale, and Stephens (2012) point out that most

content is produced by, and is about, the Global North; the Global South is underrepresented. However, the availability of demographics of users is often very limited in social media. Therefore, recognising which groups or communities are represented by a particular dataset is very challenging. Research is mostly limited to differentiating between residents and tourists based on activity span (sometimes combining with number of contribution) in a particular location [*Girardin et al.*, 2008].

2.2.5 *Different ways of exploring place properties through UGC*

One key issue of characterising places through UGC data concerns the ways in which the data themselves are localised in space and are linked to places. These data are either (1) explicitly geotagged with coordinates, whether through a link to a specific place with a name (e.g. reviews of hotels), or by an indirect link (e.g. geotagged images) to places with different granularities; or (2) they have an implicit location due to their content. The latter is the subject for geographic information retrieval (GIR), in which different processes and methods are applied to first identify place names from text using natural language processing (geo-parsing) [*Jones et al.*, 2008; *Leveling and Hartrumpf*, 2008] to thereafter assign them to locations or places (geo-coding) [*Larson*, 1996].

Explicit georeferenced content have been often studied with respect to their geographical footprint and delineating the associated regions. Hollenstein and Purves (2010) explore used geotagged Flickr photos which were tagged by the word "downtown" to delineate spatial footprints of vague regions using Kernel Density Estimation (KDE) or to represent region associated with a given name like "hydepark". Gao et al. (2017) used a hexagon-cell-based representation and calculated membership values to identify regions associated with the words "NorCal" and "SoCal" (referring to North and South California) by using georeferenced text from five different sources. Rattenbery et al. (2009) used a k-mean clustering method to group spatial footprints of geotagged photos. By combining the TagMap method with the TF-IDF technique, they could recognise regions and their representative tags based on data. Similarly, Hu et al. (2015) used a clustering method (DBSCAN) to draw regions associated with areas of interest in urban areas based around representative tags from Flickr photos.

Work dealing with data with implicit locations in the form of place names, in which different sets of methods and techniques are required to first recognise placenames, disambiguate the data and link them to places. For example, Adams and Mckenzie (2013) used topic modelling for place name recognition and then linked coordinates to relatively coarse grid cells. Similarly, Kessler et al. (2009) developed a clustering method based on Delaunay Triangulation to construct spatial footprints. One common approach for disambiguating place names and geo-referencing them, once they are recognised, is to compare them

to entries in digital gazetteers [*Jones et al.*, 2008; *Purves et al.*, 2018], in which a dictionary of placenames containing of the name, coordinates, type of place and country is provided [*Hill*, 2009].

Once the data is localised, some spatial analysis can be done to capture multiple aspects of a location or place by simple counts of, for example, users (e.g. the number of people who visited a village) or users' contributions (e.g. the number of Flicker photos assigned to a place) to measure the popularity of a place with respect to different aspects, such as the aesthetics aspect of a landscape or the popularity of a landmark [*Gliozzo et al.*, 2016]. This approach, combined with temporal information (e.g. the proportion of users/contributions at different time stamp), facilitates some spatio-temporal analysis with respect to behavioural patterns.

One way of understanding what people typically "think" about those locations and build place descriptions is through metadata in the form of text like tags attached to Flickr photos [*McKenzie and Adams*, 2017]. Textual content has been extensively explored regarding place semantics, since it provides insight into ways people interpret and conceptualise places. Content in the form of natural language text is typically unstructured (i.e. text without any predefined schema), and therefore, users freely produce the text (e.g. travelblog entries or descriptions attached to Flickr photos). Given language flexibility and text ambiguity, handling and processing such data is more challenging than structured text [*Hu*, 2018]. Various text mining methods have been developed in GIR to extract place-related information. Sentiment and emotion analysis is a very common family of methods used to examine UGC with respect to place properties [*Hauthal and Burghardt*, 2016].

Topic modelling is a commonly used method concerning a large amount of geotagged natural language data describing the same location [*Blei and Lafferty*, 2006]. This method clusters a large number of documents, each consisting of words, in a corpus based on common co-occurrences. A mixture of topics, each representing a multi-nomial distribution of words, is assigned to each document. The most common approach to topic modelling is Latent Dirichlet Allocation (LDA) [*Blei et al.*, 2003] that facilitates a simple way of summarising a group of documents. The number of topics and labels attached to topics, however, are chosen and interpreted by people. It has been used to characterise and compare places through identifying coherent themes; for example, Adams and McKenzie (2013) used LDA to identify places and compare them based on topics capturing terms related to activities, features, or localities (what local areas are called).

Content describing an image arguably gives insight into the characteristics of associated location. Edwardes and Purves (2007) used a framework for categorisations terms describing images, as proposed by Sara Shatford (1986) in information science. Table 2.1 shows a summary of the framework consisting of four

Table 2.1: The Panofsky-Shatford facet matrix (*Shatford* [1986], p. 49).

Facets	**Specific Of**	**Generic Of**	**About**
Who?	Individually named persons, animals, things	Kinds of persons, animals, things	Mythical beings, abstraction manifested or symbolised by objects or beings
What?	Individually named events	Actions, conditions	Emotions, Abstractions manifested by actions
Where?	Individually named geographic locations	Kinds of places e.g. geographic or architectural	Places symbolised, abstractions manifest by locale
When?	Linear time; dates or periods	Cyclical time; seasons, time of day	Emotions or abstraction symbolised by or manifest by

facets used to infer multiple subjects of an image (what, who, where, and when), each classified to different level of information abstraction. Different aspects of the "where" facet arguably correspond to Agnew's model [*Purves et al.*, 2019]: *specific of* (related to named places or instances of places), *generic of* (properties or features of places), and the *about* (associated emotions and feelings).

Table 2.2: Exemplar lists of *elements*, *qualities* and *activities* identified from Geograph and Flickr that are provided in the taxonomy of place description [*Purves et al.*, 2011]

Elements		**Activities**		**Qualities**	
Flickr	*Geograph*	*Flickr*	*Geograph*	*Flickr*	*Geograph*
church	road	party	walk	architecture	old
city	farm	music	grazing	night	new
sky	lane	gig	running	city	built
water	church	wedding	golf	art	centre
river	bridge	birthday	work	blue	square
building	hill	travel	cycle	light	small
park	river	christmas	fishing	red	water
street	house	concert	construction	sunset	wood
people	park	holiday	run	urban	high
garden	street	festival	walking	winter	main

Edwardes and Purves (2007) used Shatford's theory to underpin their experiment and identify a list of characteristics (*related to generic of*) associated with basic level scenes by exploring co-occurrence patterns of terms happening together. In this way, they could classify them to three categories: elements (i.e. parts), activities or qualities. Later, they could provide a taxonomy of place descriptions [*Purves et al.*, 2011] associated with these categories, annotating tasks

with nouns from Flickr and Geograph used to describe large spatial extents. Table 2.2 presents a summary of most frequent terms associated with each category from different sources.

2.3 RESEARCH GAPS

In spite of long-standing discussions about the importance of the notions of place in different research domains and recent calls demanding a place-based GIS capable of reasoning place [*Goodchild*, 2011; *Elwood et al.*, 2013], operationalising place in information systems remains to be accomplished. According to our knowledge, one example is the work by Gao et al. (2013) that suggests replacing typical distance/directional operations in classic GIS with "platial" operations (e.g. join and buffer) based on relations between places and semantic descriptions in Linked Data. They discussed how the platial join compared with spatial join might be more effective for merging the attributes of objects to target places near boundaries. [*Fjørtoft*, 2001] also used affordance theory in the sense of Gibson (1977) and explored the affordances of landscape for children's play and the influence of landscape on children's behaviour. These research works are facing a shortage of links between the broad conceptual models and the operationalisations. Studying this gap and the potential link allow us to identify which, and how, dimensions of place are currently capable of being described through data-driven approaches, and which dimensions of place remain neglected.

2.4 RESEARCH QUESTIONS AND SCIENTIFIC APPROACHES

The overall objective of this thesis is to explore meaningful ways of capturing places through their descriptive information extracted from user-generated content, and by which different application domains can benefit from. Considering biases in social media and potential limitations with respect to different implications, we aimed to measure to the validity of our results. The three main research questions, which are addressed in this thesis in the context of modelling and reasoning with places, are as follows:

Research Question 1: Which dimensions of place can be explored through user-generated content, and what are the challenges and limitations towards operationalisations of place?

Since the term "place", has been often used as a shorthand for location, and since different properties of place have been studied while neglecting the concept of place, we therefore used a combination of purposive and snowball sampling to collect a body of place-related literature.

Research Question 2: How can we model places based on bottom-up descriptions that emerge from data?

Places are either known locations with a given name that have heterogeneous characteristics, or they are regions that emerge from data with shared characteristics. A method to model specific places requires reasoning about which data are spatially and semantically relevant to a given place. Therefore, a clustering method can help to identify sub-regions with homogeneous properties with respect to a given named place. Emerging places with similar properties can be identified through exploring thematic inferences from shared experiences. Assigning topics to locations is a common approach for describing space and inferring characteristics of places from georeferenced textual information.

Research Question 3: What are the ways of characterising and comparing places based on a geographical conceptual model?

Despite the importance of Lynch's model (1960) of a city in urban studies, it has been ignored in the work of characterising cities through UGC. To bridge the gap and capture Lynch's idea that paths are the most important elements to our understanding of a city, linking UGC data to segments of streets is a meaningful approach to aggregate individual data points. Browsing between multiple maps is crucial to simultaneously explore patterns of similarity between segments across different dimensions.

3

METHODOLOGY

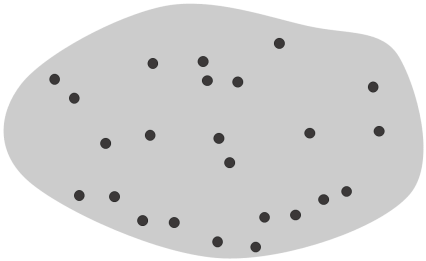
This research project mainly focussed on modelling and characterising places in urban areas by capturing shared meanings ascribed to Flickr images. To investigate spatial footprints of tags associated with images and aggregate and link them to places, we adopted three different approaches that are summarised in Figure 3.1: (1) linking a set of possibly separate regions to a given object from which the regions can be seen to therefore assign tags of images to the object (Figure 3.1I); (2) using a grid network where we treat each cell as a document containing all tags of all images located in the cell (Figure 3.1II), and (3) using a street network to link images within a specified distance from streets to the streets themselves (Figure 3.1III).

Furthermore, we identified places in three ways: firstly, by using two conceptualisations of place focussing on properties associated with the nature of place itself and the role of actors in a given place [*Agnew*, 2011; *Harrison and Tatar*, 2008]. This allowed us to explore different aspects of a place: *location*, *locale*, *sense of place*, and *people* (see section 2.1.2). Secondly, we applied a taxonomy developed by Purves et al. (2011) (summarised in Table 2.2). Lastly, we adopted the Shatford-Panofsky facet matrix [*Shatford*, 1986] (see section 2.2.5) to categorise extracted properties and show the ways in which we operationalised both geographical aspects (the *where* facet in terms of Shatford's model) and the aspects of place related to the context of the use of a place (such as *who* and *when*) (Table 2.1).

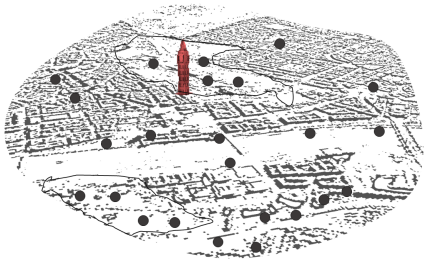
We reinforced a geographical model to capture places in visibility and street-based approaches. In the grid-based approach, we used predefined cells together with an unsupervised machine learning method to derive places. The identification of places from a grid-based approach requires a further step to assess to what extent places and their semantics are perceived as place descriptions by humans.

3.1 DATA

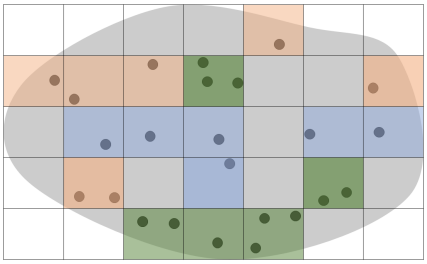
Our study focusses on, first, identifying place-like locations (such as visible parts of a scene, clusters of grid cells sharing the same meanings, or similar parts of streets with the same behavioural patterns) (Figure 3.1). Second, we characterised place-like locations in space and time (for example, we characterised geographical location and geographical boundary of a place or various



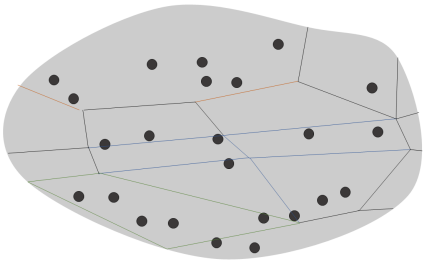
(Dataset) Assess the initial dataset: geospatial footprints of images and associated tags



(I) Link images to an object (e.g. a landmark) by linking them to areas that are seen from the object. A digital surface model was used to distinguish visible areas



(II) Consider each grid cell as a document containing tags associated with images within the cell



(III) Link tags associated with images (those within a buffer of the streets) to streets

Figure 3.1: Three approaches to aggregate individual tags shared through images and link them to places.

potential meanings of a place for different user groups in different times). We implemented our data-driven approaches for two overlapping study areas in London, United Kingdom, presented in Figure 3.2. The first study area is shown by a bounding box (as shown in Figure 3.2 using a red patch) located in central London around the river Thames includes very commonly photographed places in London such as Tower Bridge and Big Ben [*Crandall et al.*, 2009]). The second study area includes 33 boroughs of Greater London, which allow us to explore the ways, in which London was experienced through its structural elements like streets.

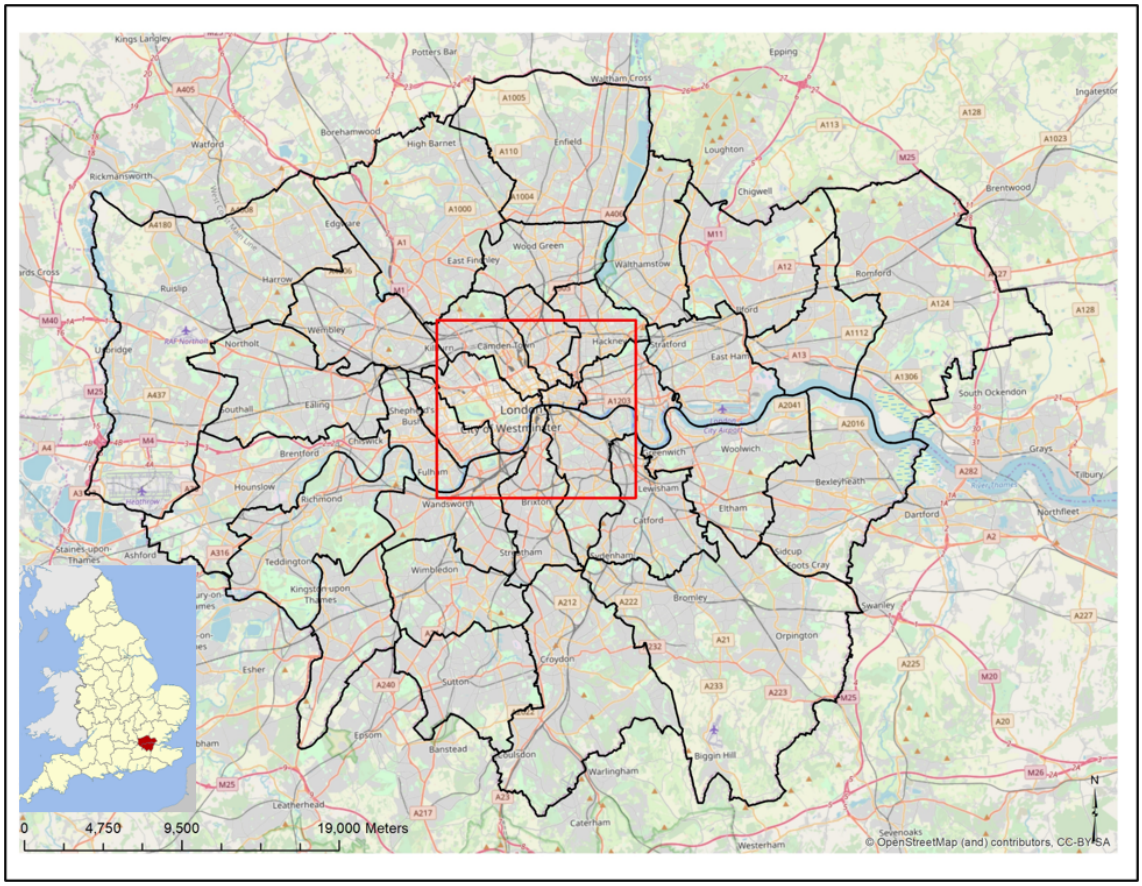


Figure 3.2: Study areas. The red patch located in central London around the river Thames includes most photographed places in London; the black lines show 33 boroughs of the Greater City of London.

To conduct our studies, we mainly used two data sources: (1) *OpenStreetMap* as a reliable and complete source with respect to both geometric and semantic data [*Haklay*, 2010], and (2) *Flickr*, which is used as a source for photos taken in urban areas [*Crandall et al.*, 2009] to explore properties of a city [*Straumann et al.*, 2014; *Girardin et al.*, 2008]. The OpenStreetMap dataset for Greater London region was downloaded from Geofabrik [1]. The package consisted of 12 layers (e.g. places, roads, or naturals), each providing information about different feature classes. Table 3.1 shows a sample of attributes provided in the "place" layer that capture geometries and attributes like name and population.

To collect Flickr data, we used queries to an available API that specified the geographical extent of each study area. Subsequently, we only gathered geotagged images and retained attached metadata (such as precision (called “accuracy” in Flickr), temporal stamps (that record the time an image was taken or uploaded), textual descriptions (tags, titles and descriptions) and a unique user identifier (Figure 3.3).

1 http://download.geofabrik.de/europe/great-britain/england/greater-london.html

Table 3.1: A summary of attributes provided for place layers in the Geofabrik package

FID	osm_id	code	fclass	population	name
0	107775	1,005	national_capital	8,416,535.00	London
177	31036374	1,002	town	66,292.00	Hounslow
208	1.24E+08	1,010	suburb	56,668.00	Beckenham
218	2.07E+08	1,010	suburb	71,552.00	Peckham
269	4.24E+08	1,002	town	58,449.00	Wimbledon

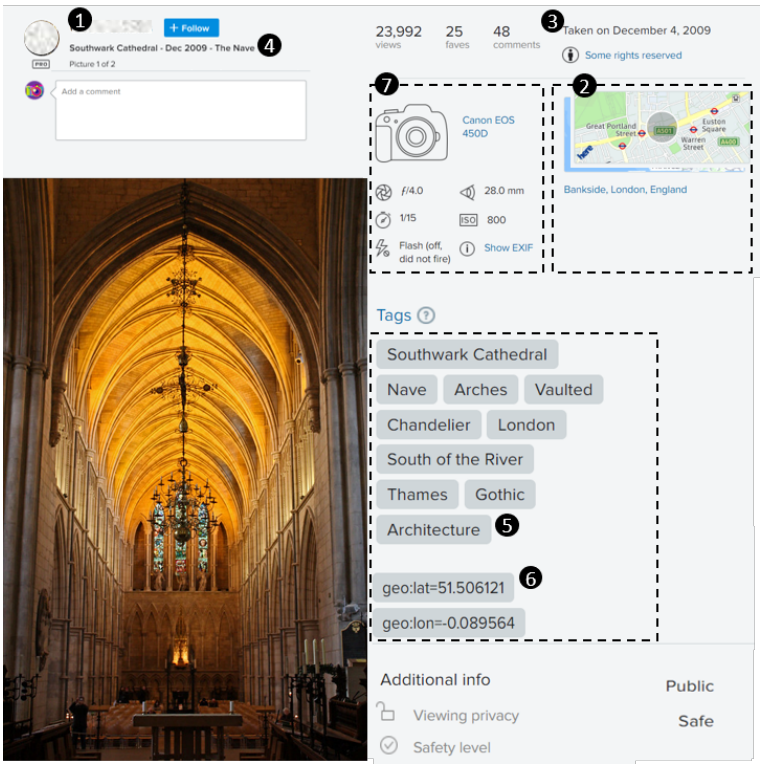


Figure 3.3: An example of metadata attached to an image on Flickr: (1) a unique user identifier, (2) image coordinates, (3) time stamp, (4) title, (5) tags given by user, (6) geo-tags, (7) tags generated by Flickr, (8) camera information *(source: https://www.flickr.com/photos/gareth1953/4163897488/. Accessed 05 December, 2019)*

Before carrying out our analysis of tags, we first carried out a range of filtering steps based on section 2.2.4. Note that we only analysed English natural language text. As a prerequisite step for text analysis, and concerning the goal of characterising places through shared tags by the majority of users, we removed (1) outliers (images from both very prolific and inactive users) [*Hollenstein and Purves*, 2010], (2) noises in content (tags generated by either Flickr (e.g.

Table 3.2: A summary of biases considered in this study and methods applied to reduce their influence

bias	source of bias	filtering method	goal	applied approach
Temporal	profile removal	check users' current status	remove users who no longer have an account	I, II
Redundancy	bulk-uploads	identify images of a user with identical tags and coordinates	remove duplicates in content	I, II, III
Content production	tagging behaviour	generate tag profiles; calculate coefficient of variation	measure "popularity of tags"	I, II, III
		fuzzy matching on a set of GeoNames	remove placenames	III
	geographic context	generate topics associated with locations using topic model (LDA); calculate the probability of tags per topics	measure "importance of a tag" in various location	III

geo:lat=51.6555) or camera (e.g. IMG001)), and (3) duplicate tags (in the tag list of a single image).

Table 3.2 provides a summary of biases to minimise their influence on our results with respect to each approach. We reduced biases in this research by using the following approaches.

TEMPORAL: We collected the first dataset before July 2014 but explored it only after almost two years. We therefore decided to eliminate users who had deleted their profile from Flickr by the time of our analysis.

BULK UPLOADS: To reduce the influence of the bias caused in the process of bulk uploading, we chose to remove bulk-upload images that had both identical coordinates and tags from both datasets.

TAGGING BEHAVIOUR: This first filtering step to select popular tags among all users was applied for datasets used in all three of our modelling approaches. According to Hollenstein and Purves (2010), we used four steps to generate a tag profile: (1) all tagged images in our dataset were listed based on number of images per user; (2) images and their subsequent tags were binned according to prolificness of users (each bin corresponds to one-hundredth of the total number of images in the dataset); (3) for each particular tag, the absolute number of images containing the given tag was counted; and finally, (4) z-values were computed to normalise the counts

of tags in each bin and to compare different patterns of contribution. z-values for each tag were calculated as follows:

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

,

where x is the count of a given tag in a given bin, μ is the mean of the tag per bin, and σ is the standard deviation of the entire population with respect to the tag. The popularity of each tag is expressed by the ratio of the standard deviation σ to the mean μ of the population (the coefficient of variation (COV) per tag), which allows us to measure whether a tag was equally used by both prolific and non-prolific posters. Finally, tags with a high COV (> 200) were removed.

The second filtering step was only applied for the "street-based approach" for eliminating the influence of place names on our results. We did this step by searching a GeoNames dataset for 33 boroughs of London in the list of tags. We identified a tag matched to the entries:

- if the tag was identical to a name in GeoNames.
- if we found a match in tags if all the spaces in the name in GeoNames were removed, for example, ealingbroadway matches English Broadway [e.g., *Alazzawi et al.*, 2012].
- if tags matched the name with only one-character transpositions (except the first one) using Damerau–Levenshtein distance [e.g., *Samal et al.*, 2004].
- if we could find the tag using the following regular expressions: $[a - zA - Z]\{2,\}$,`array`, where the array contained any of the following words: "street", "station", "road", "museum", "avenue", "square", "cathedral", "bridge", "centre", "underground station".

In the following section, we explain different methodologies used to model places through user-generated content.

3.2 MODELLING LOCATION OF PLACE

After applying filtering, we conducted three different approaches to extract descriptions of place properties, each employing different range of methods:

1. An *object-based* approach assumes that visible locations from a landmark are semantically relevant to the landmark (Figure 3.1I). Therefore, by assigning nominal regions to a landmark, we can distinguish locations that share some properties. By using an existing list of vocabularies describing multiple dimensions of a place such as elements, qualities and activities

[*Purves et al.*, 2011], properties of regions assigned to a given landmark were extracted. Finally, we performed a density-based clustering method to identify place-like regions or sub-regions sharing the same property.

Approach	**Main goal**	**Methods**
Object-based	Exploring descriptive information for regions visible from specific places	viewshed analysis, taxonomy of elements, qualities and activities, and dbscan clustering
Grid-based	Extracting similar places based on semantic topics and exploring different dimensions of places	topic modelling (LDA), measuring coherence values and corpus distance, and annotating topics based on dimensions of place
Street-based	Calculating similarities between streets segments based on three contrasting dimensions: semantics, time, and user behaviour	weighted TF-IDF cosine similarity measuring, binary cosine similarity measuring, and Euclidean distance

Table 3.3: Summary of methods applied to each approach to explore descriptions of place properties

2. In *grid-based* approach, we intended to go beyond descriptions of specific locations and present a continuous spatial model in which all locations could be characterised. To do so, as represented in Figure 3.1II, we used a grid network to model geographical aspects of a place. Then, by aggregating content of images within each cell, we were able to compare them according to collections of assigned co-occur tags. Having a list of most probable vocabularies assigned to each cell in the form of a thematic topic, we categorised words based on four elements of place [*Agnew*, 2011; *Harrison and Tatar*, 2008]: *location*, *locale*, *sense of place* and *people*. And, finally, through the labelling together with a quantitative measure, we inferred that descriptions resulting from aggregated tags within a grid cell generate meaningful descriptions of places.

3. The *Street-based* approach was inspired by the book *The image of the City* [*Lynch*, 1960] that introduced the most salient elements representing people's perception of a city. Here, we focus on paths from which the environment can be experienced. In our study, we assumed that streets are places which might share similar semantics. Similar to the first approach, we started from specific locations (named popular locations vs. named streets) to explore similarities between major streets. To identify places and characterise them, we then analysed information related to three contrasting dimensions (such as user behaviour, semantics and time). To do so, we

explored computationally the ways in which a street or place has been described or been used with respect to users who were visiting the place and also time of the visit (Figure 3.1III).

Table 3.3 provides a summary of main goals and methods for achieving the goals for each approach. Having a variety of methods, all approaching the same goal —exploring descriptions of place properties —we chose to compare our approaches to the Panofsky–Shatford facet matrix (as shown in Table 2.1) and demonstrated different ways in which we addressed multiple aspects of places based on the *where* facet representing geographical aspect of a place, the *when* related with the time in which a place was visited, and the *who* that relates to the users who visited a place. We also explored descriptions based on different dimensions of *where* facets.

3.2.1 *Object-based approach*

Landmarks are salient objects in their environment and accordingly are often used in everyday communication about navigation and way finding. They are linked to locations [*Purves et al.*, 2019], and presumably in our research, landmarks are linked to regions from where they are visible. In our first attempt to model a place based on textual descriptions, we focussed on geographical objects, either landmarks or popular locations. We chose ten places in London: four places among the top seven photographed landmarks in the world (such as `Trafalgar Square`, `Tate Modern`, `Big Ben`, `London Eye`); three places among the top seven landmarks in London (`Piccadilly Circus`, `Buckingham Palace`, and `Tower Bridge`) [*Crandall et al.*, 2009], and finally, three prominent touristic attractions (`St. Paul's Cathedral`, `the Globe Theatre` and `Hyde Park`). By using geographical footprints of given places from OpenStreetMap as shown in Figure 4.1, we started from place names (*where/ specific of* in the sense of Shatford (1986)) to explore properties and features of the places after filtering out biases (Table 3.2) discussed in subsection (2.2.4).

Similar to the approaches focussing on textual descriptions, we need to find text relevant to the geographic coordinates at some semantic and spatial granularity [*Derungs and Purves*, 2014]. Images which are tagged by a place name are more likely to be semantically relevant to the place. Therefore, we searched for place names in the textual metadata of each image (such as tags, title and descriptions). We used Damerau–Levenshtein distance [*Brill and Moore*, 2000] concerning spelling variations or typos in place names (e.g. Bukingham Palace instead of Buckingham Palace) [*Levenshtein*, 1966]. We then performed viewshed analysis using a freely available 1m resolution digital surface model [2] and OSM geometry (either point or polygon) of each location. Having identified nominal regions related to a given location [*Senaratne et al.*, 2013b; *Fisher*, 1996], we could

2 https://data.gov.uk/dataset/6a117171-5c59-4c7d-8e8b-8e7aefe8ee2e/lidar-composite-dtm-1m

model the geographical aspect of a place or the *where* facet of Shatford's model by keeping all images and their tags located within the viewsheds linked to our ten geometries.

Having a list of tags linked to each location and filtering out biases shown in Figure 3.4, we used an existing taxonomy [*Purves et al.*, 2011] to identify terms describing elements (e.g. river, road, hill), activities (e.g. music, festival, birthday), and qualities (e.g. summer, urban, sunset) reflected in images. These provide descriptive place information at the level of generic of (*sensu* Shatford (1986)) related to the *where* facet.

Since places are effectively experienced as regions [*Montello et al.*, 2003], we identified sub-regions sharing a particular tag within linked areas to a named location. To do so, we used a collection of existing *elements*, *qualities* and *activities* for each location and then retrieved the coordinates of associated images. Furthermore, we could perform a dbscan clustering [*Elsner and Kara*, 1999] per tag to distinguish regions captured by a single tag.

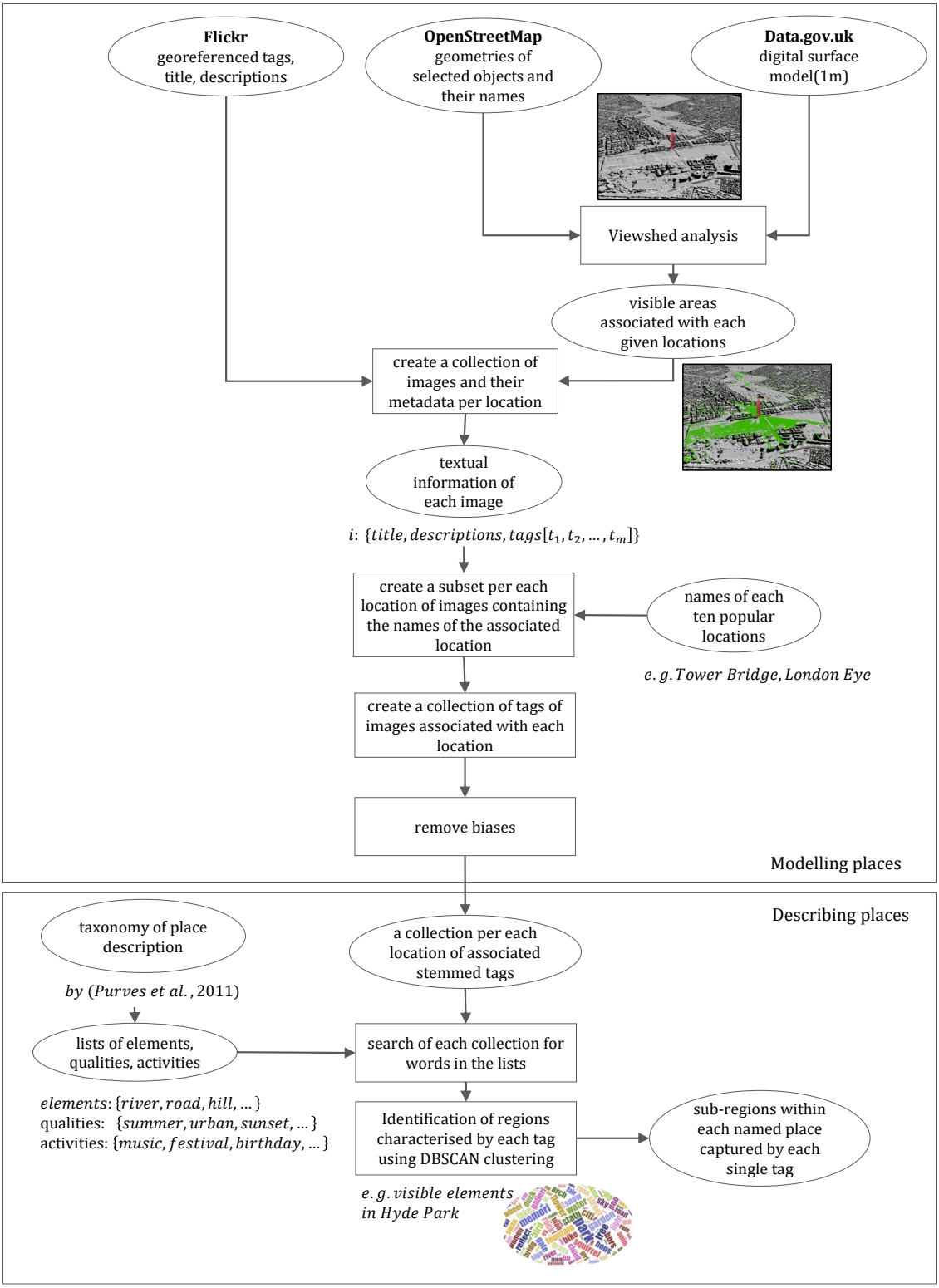


Figure 3.4: Flowchart describing the overall process in the object-based approach

3.2.2 *Grid-based approach*

A grid-based approach consists of two main sets of methods (as shown in Figure 3.5): (1) *computational* methods for modelling places, in which we test our first hypothesis that using tags of Flickr images enables us to extract semantics of unnoticed locations in UGC [*Goodchild*, 2007] by exploring similar places using a topic model; and (2) *interpretation* methods for characterising emerged places that allow us to explore our second hypothesis about the quality of topics that can be expressed in terms of their interpretability by humans [*Newman et al.*, 2010; *Mei et al.*, 2007] as well as coherence value of a topic.

As a prerequisite step, biases with respect to users participation and contributions (see Table 3.2) are removed.

3.2.2.1 *Computational methods: Topic modelling*

To model geographical aspects of a place or *where/generic of* aspects (*sensu* Shatford (1986)), we used a grid network overlaying our study area and then linked the content to grid cells. The grid resolution defines the spatial granularity of our text analysis. We will discuss the process of selecting grid resolution in subsection 3.2.2.2. We chose to run LDA [*Andrienko et al.*, 2013; *Blei et al.*, 2003] to explore both generic and specific properties of locations. We did so through the following steps:

1. We generated a document per grid cell, each associated with a vector of all occurrences of each of the identified unique tags after filtering:
2. We used the documents as inputs to the MALLET LDA toolkit [*McCallum*, 2002] and calculated the optimised hyper parameters for a given number of topics n [*Cao et al.*, 2009].

LDA produces the following vectors:

- for each of n topics, a vector of all tags and their probabilities of belonging to that topic;
- for each grid cell (document), a list of all topics and the probability of the grid cell belonging to each topic; and
- for each topic, a set of measures describing topic quality.

We the assigned the most probable topic to each grid cell and grouped the cells associated with the same topics into regions. Tags in a topic were ranked based in their usefulness in characterising an individual topic [*Aletras et al.*, 2017]. Thus, by exploring cumulative probability tag curves of each topic and applying them as a mask, tags with less effectiveness were removed.

3.2.2.2 *Computational methods: sensitivity analysis*

Despite the spatially continuous model which allows us to characterise every location within each cell in terms of tags, the influence of the resolution of the grid over which data are aggregated presents a limitation in this approach [*Openshaw*, 1983]. We addressed this problem by analysing the sensitivity of our results from four different resolutions: 50, 250, 500, and 1000m. Another challenging decision is selecting an appropriate number of topics as inputs for the topic model. To investigate both the sensitivity of the model to resolution and the appropriate number of topics, we used a range of measures produced by MALLET. These metrics help us to express the quality of generated topics. Moreover, we use them to explore the semantic quality of the topics, which we explain in subsection 3.2.2.3. Therefore, three different measures were chosen for investigation: (1) corpus distance which measures the similarity of a topic to the corpus as a whole, (2) number of tokens as an indicator for the number of tokens or tags associated with each topic, and (3) coherence values to characterise how semantically coherent a topic is.

We utilised two measures, corpus distance and number of tokens, to optimise grid resolutions and number of topics. Documents with higher corpus distance are distinctive from the corpus. In other words, places have different characteristics and are distinguishable from the whole area [*AlSumait et al.*, 2009]. It is more likely that the number of tokens assigned to topics decreases as the number of topics (or the grid resolution) increases, because with smaller area and less data, the need for generalisation over cells and topics is less. Hence, the number of tokens should be large enough to characterise individual topics but also small enough to be distinctive from the corpus (c.f. corpus distance) [*Mimno et al.*, 2011]. The coherence value is calculated based on the probability of tags in a topic co-occurring in cells belonging to the topic:

$$\text{coherence} = \sum_{i} \sum_{j<i} \log \frac{D(w_j, w_i) + \beta}{D(w_i)} \tag{3.2}$$

, where β is a parameter to prevent log zero errors, $D(w_j, w_i)$ is the number of co-occurrences of two terms in a document, and $D(w_i)$ is the number of occurrences of the more probable terms.

Large negative values indicate that the tags in a topic rarely co-occur in grid cells, whilst zero values indicate that topics and associated tags are semantically coherent [*Stevens et al.*, 2012].

3.2.2.3 *Interpretation methods: Topic labelling and annotation*

To test the second hypothesis, firstly, we identified coherent topics: in other words, tags describing a topic (e.g. Kings Cross, railway and Paddington) express a theme (e.g. railway-related cluster) that is interpretable in terms of Lon-

don's geography. We did so by assigning a label to each topic manually, a typical step in any unsupervised classification method. It is important to point out that local knowledge is very crucial at this stage. Concerning the difficulties in labelling topics, only some of them that we were confident about our interpretation got labelled. Secondly, we compared coherence values of labelled and unlabelled topics.

Finally, we adopted two notions of place concerning both the nature of place itself [*Agnew*, 2011] and the important role of people in a given place [*Harrison and Tatar*, 2008]. We then annotated the labels with four elements of a place and the combination thereof: *location* (labels related to specific locations and place names), *locale* (labels reflecting generic information about a place, like explicit activities or the features or objects describing a place), *sense of place* (labels characterising the "about" aspect of place associated with emotions and feelings), and finally, *people* (labels related to another *who* facet characterising individuals or groups associated with a place).

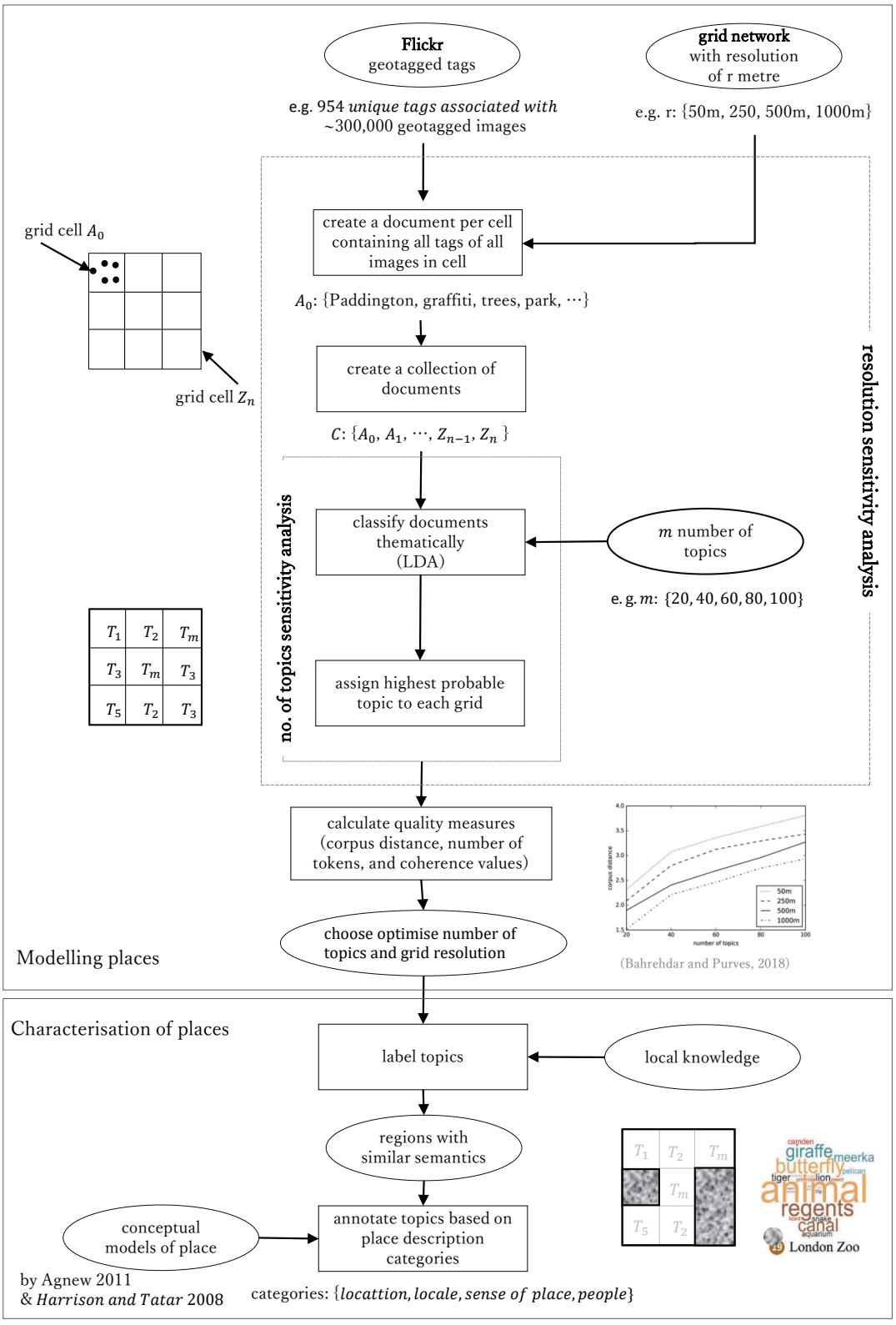


Figure 3.5: Flowchart describing overall process in a grid-based approach

3.2.3 *Street-based approach*

To characterise London based on Lynch's idea (1960), we used the metadata of Flickr images as the source of georeferenced textual information and a subset of OpenStreetMap road layer to model place at the level of streets or paths in which people experience a city (Figure 4.13). We do so in three main steps: first, by modelling paths as fundamental units (i.e. segments of a street) for our analysis; second, by identifying relevant attributes to each segment; and third, by measuring similarity between segments based on multiple dimensions of place such as semantics, user behaviour and time relating to *where*, *who*, and *when* facets sensu Shatford (1986).

3.2.3.1 *Modelling paths and assigning attributes*

Concerning gaps in data with respect to user activity in terms of generating content and with the aim of characterising London through paths [*Lynch*, 1960], we chose to use only major roads (primary, trunk and secondary) according to their types and references to UK national road classes. As a consequence, some pseudo-nodes appeared where there was no longer a junction. We removed these nodes from individual segments with the same name, type and class. We furthermore filtered out segments with lengths of less than 200m. Doing so, we were left with 3,406 segments with a median length of 519m. Finally, paths were street segments longer than 200m with a name but no topological relationships.

As discussed in subsection 2.1, there are two types of bias: (1) bias related to users' presence at a location as well as their loads of contributions and (2) bias related to user behaviour in terms of content which can amplify individual voices and over-represent their opinions. To avoid influence of individuals on street semantics, but still consider their presence and activity in analysis with respect to time and user behaviour, we decided to use two datasets as shown in Figure 3.6: one dataset for both user behaviour and temporal similarities, in which we filter biases based on participation behaviours, and the second dataset for calculating semantic similarity that is strongly filtered based on tagging behaviours as well as participation bias. Henceforth, we call these datasets the user and content datasets.

The initial dataset consists of metadata of Flickr images (such as unique user ids, tags, image coordinates and the timestamp at which a photo was taken) for 33 boroughs of Greater London. To create user datasets linked to paths, first of all, we removed outliers and bulk uploads. We then used a 100m buffer around the segments to identify geographically relevant images. The remaining dataset consisted of 1,250,205 images. 61,184 were used for calculating similarity according to users and time. To retain semantically relevant tags, we also removed both ASCII characters (e.g. 伦敦) in tags and tags transferred through Instagram links (e.g. Valencia, which is a photo filter) because their subjects

rarely related to locations. After selecting popular tags using tag profiles, we again used a 100m buffer around street segments to find images related to the street segments.

We then performed an LDA to measure the importance of a single tag for different locations (or segments). LDA enables us to identify topics, and consequently, vocabularies that are specific to a set of paths. The reasons to perform this method are, first, to select important tags with respect to the geography of London and based on calculated probabilities of words belonging to a topic; and second to choose the most important popular tags to reduce the dimension of our measure for better performance.

We did so by performing LDA based on the steps we explained in section 3.2.2.1 with an exception: that each cell includes segments and tags of all images within the buffer of street segments. According to our sensitivity analysis explained in subsection 3.2.2.2 and resultant evidence that we will present later in section 4.2.1, we chose a 500m grid resolution to generate documents and 40 topics as input for LDA.

At the output, each topic was represented by a list of all tags and their probabilities of belonging to that topic. In addition, each grid cell was represented by a vector of 40 topics and the probability of the grid cell belonging to each topic. We then assigned the most probable topic to each grid cell. Tags associated with high probabilities represent the more useful and important tags in characterising an individual topic [*Aletras et al.*, 2017], and consequently, the corresponding grid cell. By selecting tags predicting 80 percent of the cumulative probability per topic, we could remove tags which are neither very influential nor representative. Finally, remained tags associated with each grid cell transferred to all segments that intersect with the grid.

Since using place names is common in tagging behaviours [*Sigurbjörnsson and Van Zwol*, 2008], we decided to remove place names from the lists of tags associated with segments. Therefore, we applied a fuzzy matching algorithm (explained in detail in subsection 2.2.4) of extracted place names from a collection of GeoNames for Great Britain. Finally, we were left with 1,605 unique tags and 4,268,980 tags describing 671,207 images by 36,486 users. The list of 1,605 unique tags associated with grid cells were passed to the segments and then were used to calculate semantic similarities.

3.2.3.2 *Measuring similarities*

Semantic similarity: To calculate the semantic similarity between each two street segments, we compared a collection of lists of tags describing each segment using the following steps:

1. Each segment (S) is presented as a vector $V_S = [t_1^s, t_2^s, ..., t_n^s]$, where each member of the vector t_1^s correspond to the usage of a tag in a segment.

Therefore, n is equal to the number of unique tags identified from the previous stage. We used the term frequency-inverse document frequency (TF-IDF) measure to compute the value of each t_1^s. Using TF-IDF, we measured how important a tag is for a segment. The size of the content for each segment is different due to differences in number of images and tags. We then normalised the values with respect to the size of content for each segment. The values increase proportionally with the number of occurrences of a tag in a segment. Lastly, we considered how frequent the tag is in all dataset.

2. The similarity between each two segments was calculated using a cosine similarity measure between the TF-IDF weighted term vectors. Since the size of all the vectors are the same (equal to the number of all unique tags), the similarity can be computed as a dot product of two vectors as follows:

$$\text{Sim}(v_{s_1}, v_{s_2}) = \cos(\theta) = \frac{v_{s_1} \cdot v_{s_2}}{\|v_{s_1}\| \, \|v_{s_2}\|} \tag{3.3}$$

Similarity measures range from 0 to 1, with 1 indicating that the two semantics are identical and 0 representing complete dissimilarity. Computing the dissimilarity between each of the two segments produces a ranked set of streets, we are able to identify the most similar segments to each one.

User behaviour similarity: To compare similarity between two segments in term of the number of shared users (whose photo is linked to the segments), we used a cosine similarity measure. It is assumed that each segment was represented as a binary vector with the length of all number of users in the user dataset. An element in the vector was assigned a value of 1 if the corresponding user was present in the segment; otherwise the value was 0. The method looks at each of the two vectors and finds the incident where both values are equal to 1. The resulting value reflects how many 1:1 match occurs in comparison to the total number of users. Furthermore, we differentiated between tourists, whose total images were taken within only two weeks, and global users who were not tourists. Thus, we can explore patterns of both groups of people.

Temporal similarity: We calculated temporal similarity according to the day of the week a photo was taken. To do so, a histogram per segment was created, each bar demonstrating the distribution of images over days of week, where the number of images taken on each day was counted and normalised to the scale of 1. Finally, we calculated temporal similarities between both segments by measuring the Euclidean distance between a seven-dimensional vector, where we treated the proportion of images taken on each day of the week as an independent dimension.

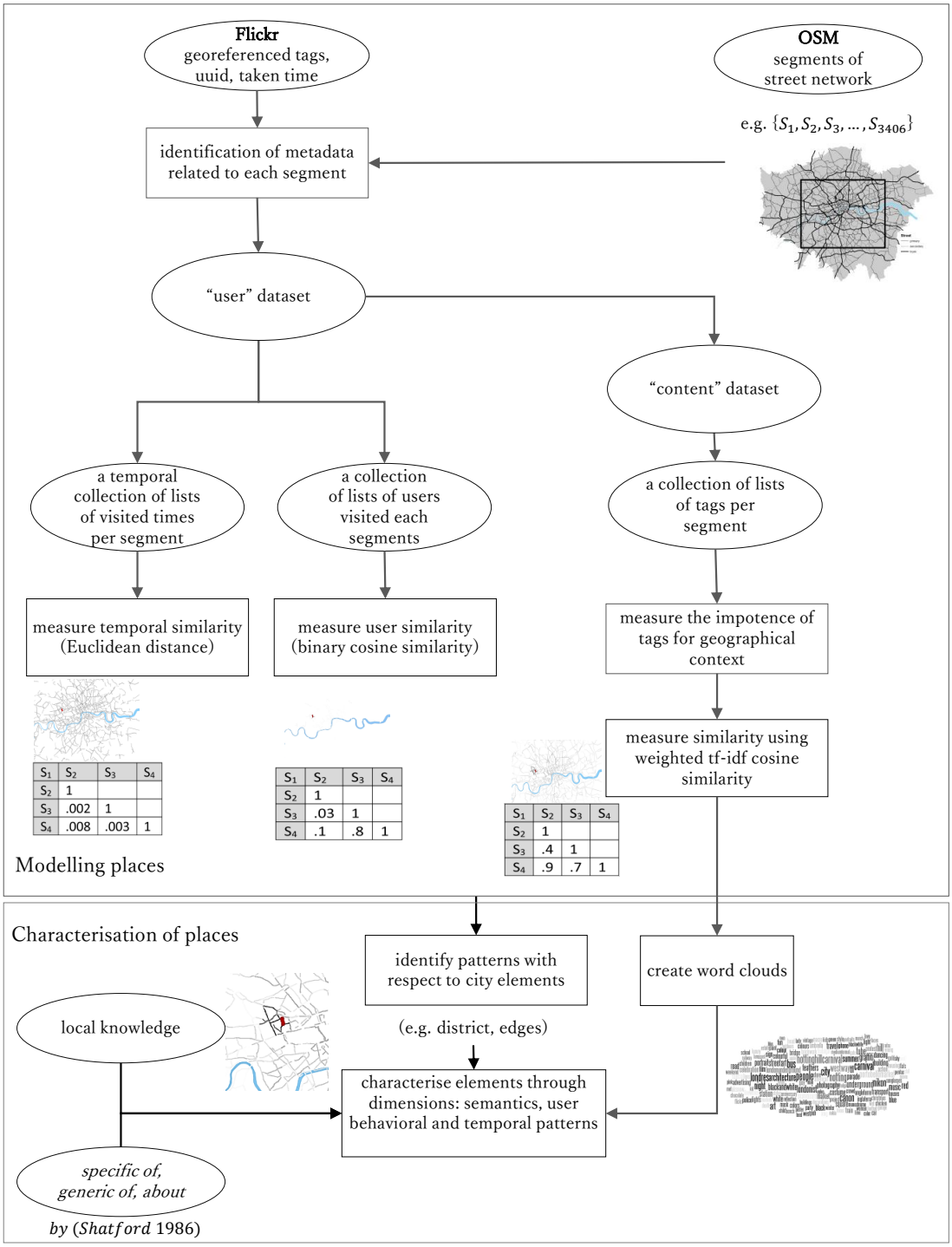


Figure 3.6: Flowchart describing overall process in street-based approach.

4

RESULTS AND INTERPRETATION

4.1 LANDMARKS AND PLACES IN LONDON

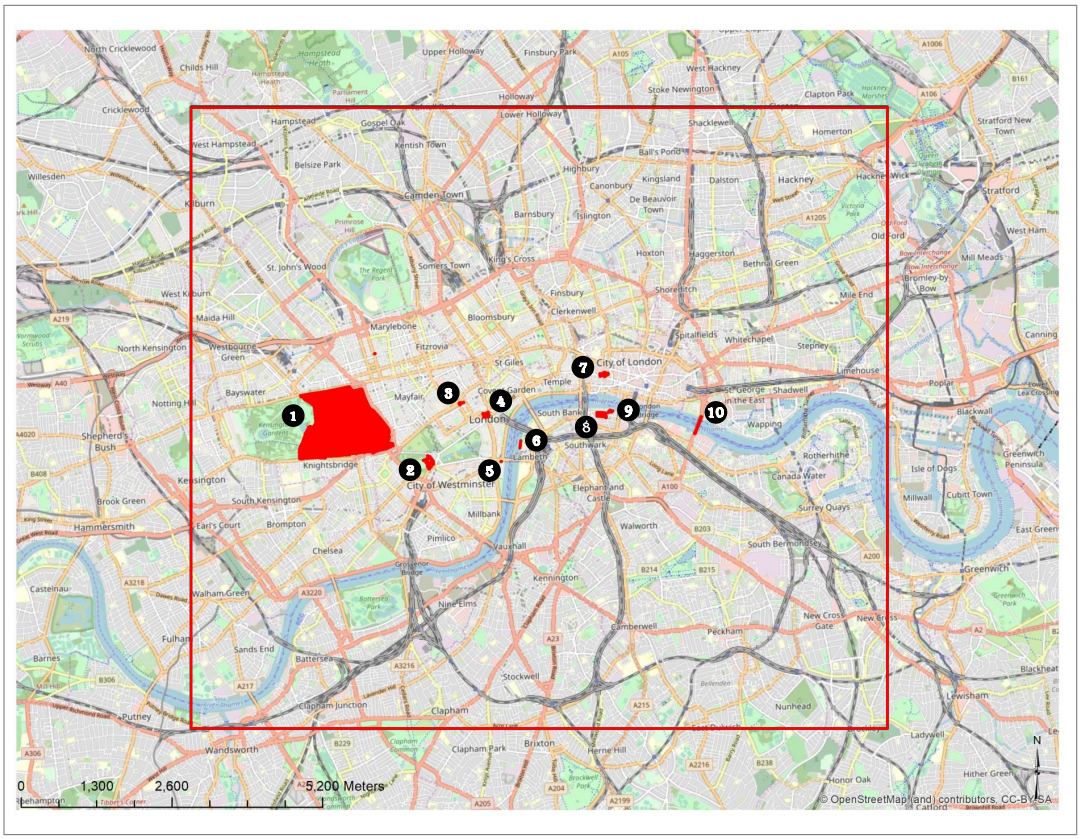


Figure 4.1: Our ten selected locations in central: (1) Hyde Park, (2) Buckingham Palace, (3) Piccadilly Circus, (4) Trafalgar Square, (5) Big Ben, (6) London Eye, (7) Saint Paul's Cathedral, (8) Tate Modern, (9) Globe Theatre, (10) Tower Bridge.

Figure 4.1 shows the locations of the ten selected places in London retrieved from OpenStreetMap. The initial dataset of georeferenced images within the given bounding box consisted of 3,105,544 images shared by 49,130 users, all taken before July 2013. The number of images reduced by two-third after removing images where the georeferencing precision was less than street level. Finally, we were left with approximately 10% of original number of images (371,752) that could be potentially used for characterising areas associated with objects.

Table 4.1 presents the effect of each filtering steps on the number of images and users.

function	#images	#users
original dataset	3105544	49130
accuracy filtering	1047003	31092
bulk-upload filtering	839822	31080
camera-generated content	571241	30377
inactive users	503536	8143
prolific users	404329	8060
null tags	371752	7753

Table 4.1: Remaining number of images within the given bounding box remaining at each stage of filtering.

We used the cleaned dataset to explore three different aspects of each place according to *elements*, *qualities* and *activities*. Figure 4.2 illustrates the influence of tagging behaviour on tags referring to, for example, *elements* perceived around `Tower Bridge`. Stemmed tags like `villag`, `railroad`, and `lake` with high coefficients of variation are frequent among lists of tags associated with the location, but not popular among all users contributing to the subset. Therefore, we only retained tags with low coefficients of variation (< 200): for example, `bank`, `station`, and `train` as descriptive information for Tower Bridge (Figure 4.2). The statistics of remaining tags capturing different aspects of the locations and illustrating the richness of such data to describe the locations are presented in Table 4.2. `Trafalgar Square`, `Big Ben`, `London Eye`, and `Tower Bridge` were visited by more than 50% of the remaining 7 753 users, all among the top ten most photographed places in London. Except for the Globe Theatre, all the locations were described through more than 1000 images. Except for `Piccadilly Circus` and the Globe Theatre, all locations are characterised by more than 1000 users.

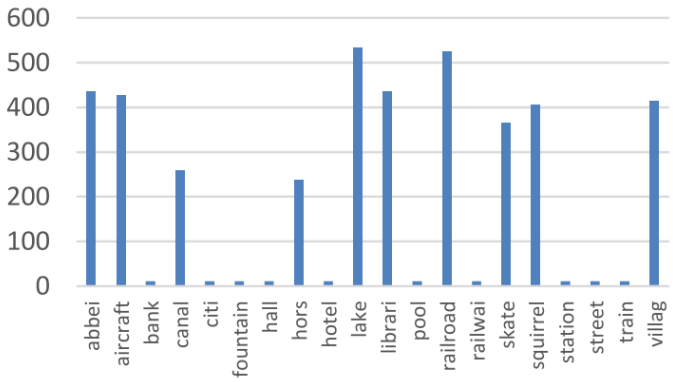


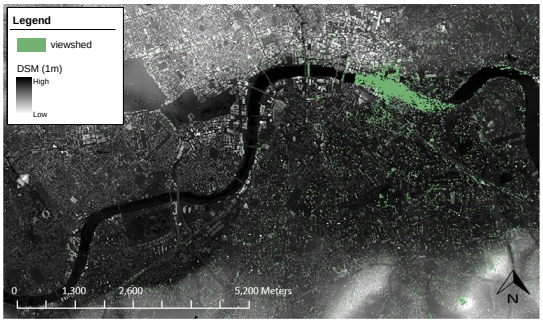
Figure 4.2: Coefficient of variation of 20 selected elements around Tower Bridge.

location	#images	#users	#elements, #qualities, #activities
Trafalgar Square	12801	3586	232, 142, 85
Tate Modern	7249	2490	227, 135, 74
Big Ben	8257	4335	229, 134, 72
London Eye	12645	4817	233, 132, 80
Piccadilly Circus	1102	691	124, 88, 49
Buckingham Palace	5733	2209	204, 131, 59
Tower Bridge	7738	3621	227, 136, 70
St. Paul's Cathedral	5430	2179	221, 133, 42
Globe Theatre	592	415	141, 80, 42
Hyde Park	6901	1761	229, 133, 76

Table 4.2: Number of images, users, and categorised tags for the ten locations after filtering biases found in the visible areas [*Bahrehdar and Purves*, 2016]

Word clouds in Figure 4.3 and Figure 4.4 are two typical examples of our results. Each word cloud demonstrates the 100 most popular tags, with respect to the identified aspects (such as *elements*, *qualities* and *activities*) of regions associated with each given location (green areas in Figure 4.3a and Figure 4.4a). The font size of words reflects the popularity of each tag among all users contributed in associated regions. Figure 4.3a shows that areas relevant to `Tower Bridge` are alongside the river Thames due to its visual and physical salience. The most representative *elements* associated with `Tower Bridge` (presented in Figure 4.3b) are related to the visual appearance of the place (e.g. `brick`, `trees`), the geographical features (e.g. `river`, `bank`) or to the *elements* related to either the landscape (e.g. `clouds`, `sky`) or the affordances in the location (e.g. `boat`, `market`). Figure 4.3d) presents the perceived attributes of things capturing *qualities* like colours (e.g. `"blue"`) or the word `"beauty"`.

By contrast, viewsheds of `Hyde Park` are almost limited to its boundaries in OSM (Figure 4.4); however, there are images tagged as "Hyde park" located in Kensington Gardens, where we assume that people wrongly perceived two different locations, `Kensington Gardens` and `Hyde Park`, as one place. The most representative tag describing physical aspects of `Hyde Park` is, unsurprisingly, `"park"`, which is followed by objects or things inside like `memorial`, `flower`, or animals like `horses` or `squirrel`. Based on an *activities* word cloud, `Hyde Park` is a place for events like `festivals`, `carnivals`, race and concerts, which are associated with being `fun` and/or related to `art`.



(a) Computed viewshed for Tower Bridge

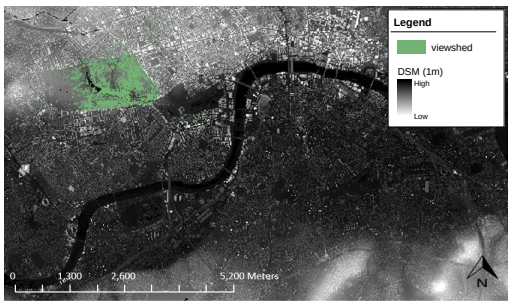
(b) Words describing *elements* associated with Tower Bridge

(c) Words describing *activities* associated with Tower Bridge

(d) Words describing *qualities* associated with Tower Bridge

Figure 4.3: Various aspects of regions associated with Tower Bridge.

Clusters of images associated with `Tower Bridge` and tagged by the word "bank" illustrate that the southern bank of the `Thames` was preferred as a view point (Figure 4.5).



(a) Computed viewshed for Hyde Park

(b) Words describing *elements* associated with Hyde Park

(c) Words describing *activities* associated with Hyde Park

(d) Words describing *qualities* associated with Hyde Park

Figure 4.4: Various aspects of regions associated with Hyde Park.

4.2 TOPICS DESCRIBING PLACES

We used the same dataset as the previous experiment to test our second approach, but, the sequence of filtering steps differs from a visibility approach, since we used different methods to aggregate the metadata and to model place (see section 3.2.2). After all filtering steps for removing contribution biases (Table 3.2) and removing tags with high coefficient of variation (> 200) using tag profiles, we were left with 956 unique tags to generate documents for topic modelling.

Figure 4.6 demonstrates the density of contributors of our dataset. From Central London to the West End is more crowded than other areas. One reason could be the tourists and leisure attractions (e.g. `Tower Bridge`, `Big Ben`, and `Tate Modern`). According to the results of a linear regression ($r_2 = 0.95$), the number of users is highly correlated with their contributions. Performing a spatial auto-regressive regression (SAR) model yielding to a correlation value of 0.96 suggests a limited influence of spatial auto-correlation in our model, which means that contributions in a grid cell are a function of the number of con-

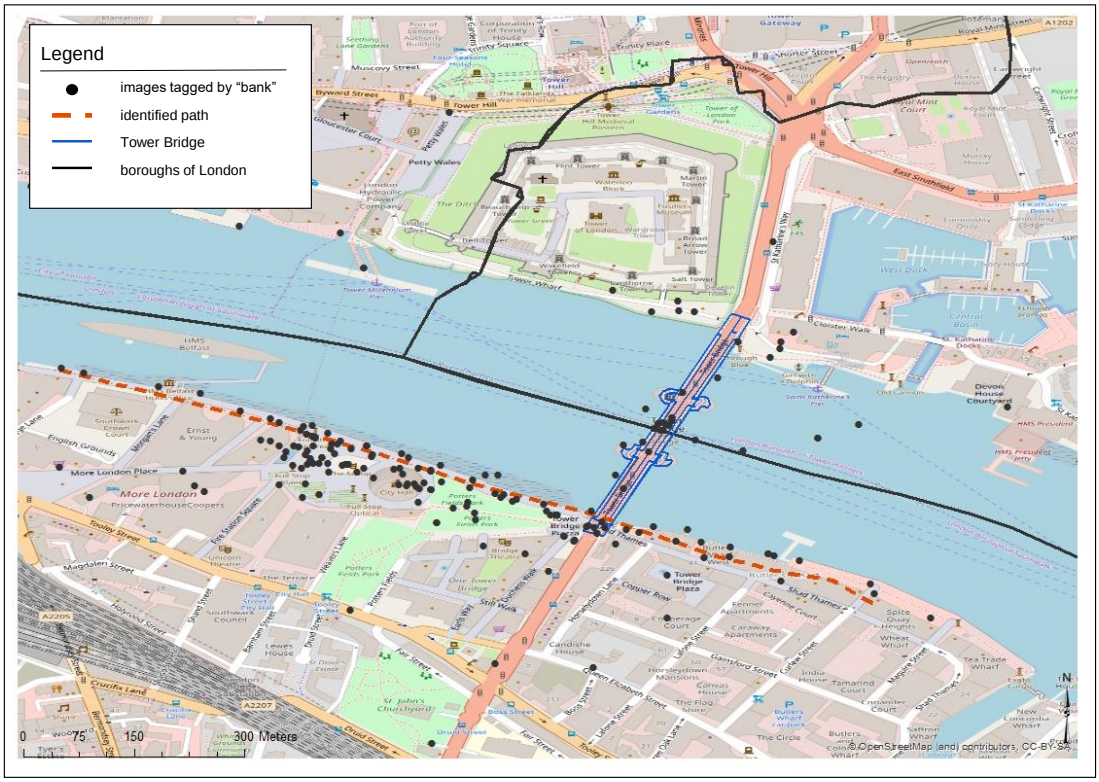


Figure 4.5: Cluster of images associated with the tag "bank" and its potential use in exaggerating the representation of the river bank south of the Thames.

tributors. Simply put, individual users do not affect the spatial distribution of images. Therefore, their influence on the produced content, either geographically or semantically, is minimised and shows the effectiveness of our filtering steps of removing tags with low coefficients of variation on minimising biases in content related to tagging behaviour.

4.2.1 *Sensitivity test*

Here we explain the influences of the size of grid cell and the number of topics on our final results. Table 4.3 summarises the calculation of two measures, median number of cells and median corpus distance, for five different grid resolutions and the number of topics. The results of the median number of tokens have no correlation with grid resolutions, therefore, we only present median corpus distance.

The strong correlation between mean median corpus and grid resolution (Pearson correlation: $r_2 = 0.95$) suggests that high resolutions yield the most distinctive topics. When the resolutions become finer, the number of cells associated with a topic decreases, because the number of cells without tags increases. We

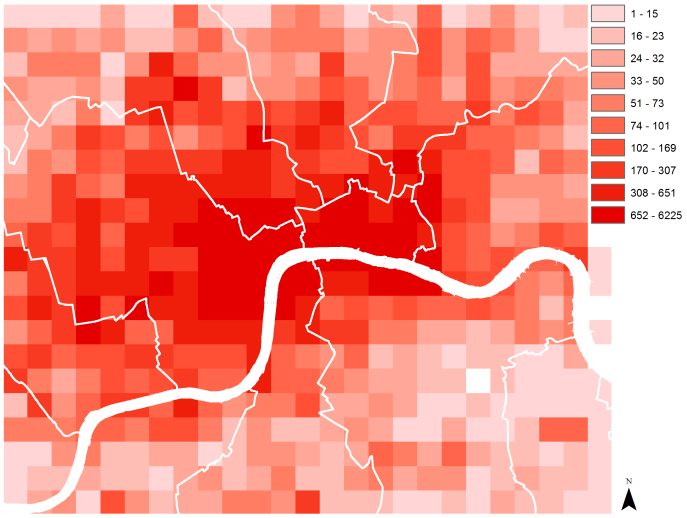
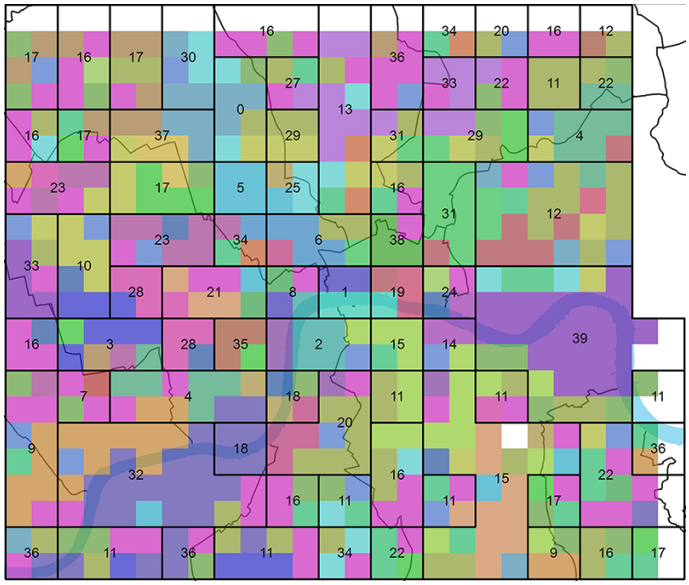


Figure 4.6: The number of contributors taking photos in each grid cell with a resolution of 500 metres. Figure from Bahrehdar and Purves (2018) – Publication II.

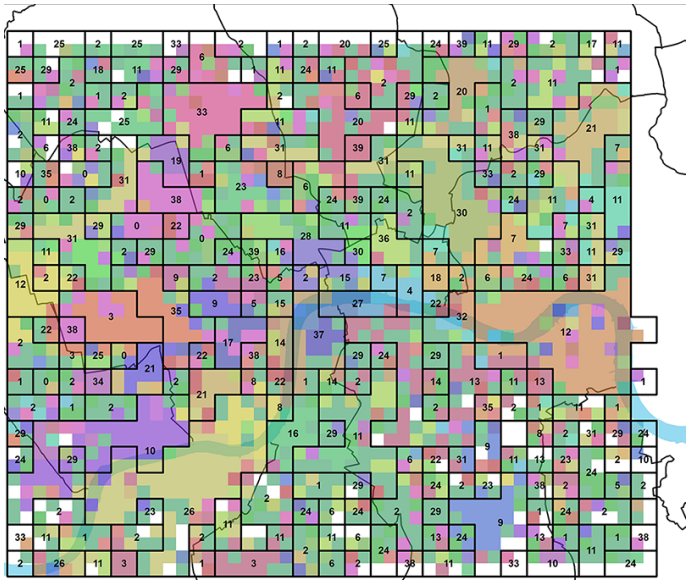
can see the same effect in Figure 4.7 whereby colours show clusters of higher resolutions compared to clusters of 40 topics in lower resolutions. According to the graph presented in Figure 4.8 when the resolution increases, the number of cells associated with topics drops. For example, white grids could not be allocated to a topic because of insufficient number of tags in the grid. We chose a 500m grid resolution, to balance between very coarse resolutions (where meaningful places are not delineated) and fine resolutions (where we have insufficient data to describe places for many cells) as represented in Figure 4.7a and Figure 4.7c.

After establishing an optimum grid resolution, we analysed the influence of the number of topics on our topic model results. Measures of corpus distance in Figure 4.8 show that the biggest change happens at 40 topics, irrespective of resolution; therefore, the biggest change in distinctiveness of our topics is likely to occur if we increase the number of topics from 20 to 40. As with resolution, simply increasing the number of topics results in higher corpus distances and thus more distinct topics. Moreover, we investigated the sensitivity of our results with respect to the number of topics. Therefore, we explored both the number of cells associated with each topic and the relationship between number of topics and the number of cells.

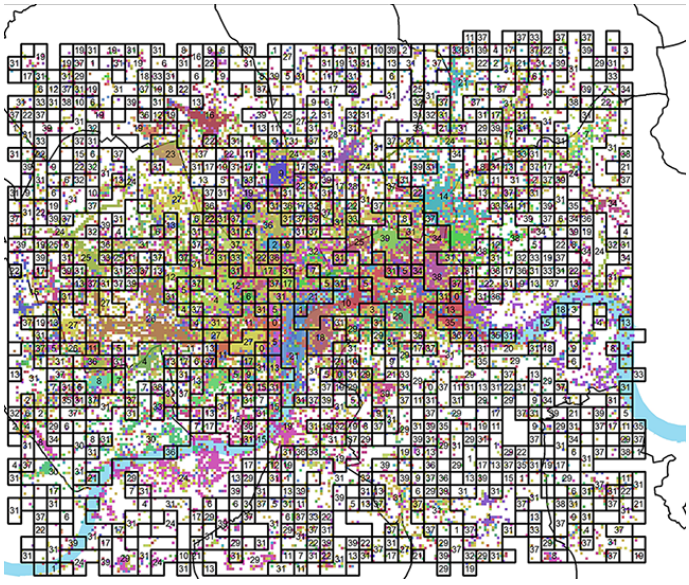
To study the number of cells allocated for each topic, we plotted corpus distance in Figure 4.9 to explore its variation. We observed that corpus distance varies as a function of both the number of topics and the number of cells, or area, associated with a topic. We did desire that our results, the distinctiveness of our topics, would not strongly vary as a function of area, which could mean that topics associated with single cells are not much more distinctive than those associated with large areas or vice versa. We realised that 40 topics seems to have



(a) 1km vs. 500m



(b) 500m vs. 250m



(c) 250m vs. 50m

Figure 4.7: Comparison of clusters of 40 topics with respect to the grid resolution. Figure from Bahrehdar and Purves (2018) – Publication II.

Resolution(m)	No. of topics	Median no. of cells	Median corpus distance
50	20	678.0	2.32
	40	349.0	3.07
	60	228.0	3.36
	80	166.0	3.59
	100	131.0	3.81
250	20	64.0	2.90
	40	32.5	2.80
	60	18.5	3.13
	80	14.0	3.30
	100	11.0	3.44
500	20	16.0	1.90
	40	10.0	2.41
	60	5.0	2.69
	80	4.0	2.96
	100	3.0	3.27
1000	20	5.5	1.52
	40	2.0	2.21
	60	2.0	2.46
	80	1.0	2.75
	100	1.0	2.93

Table 4.3: Median corpus distance and number of cells per topic as a function of the number of topics for different grid resolutions. Figure from Bahrehdar and Purves (2018) – Publication II.

the most stable corpus distance as a function of the number of cells associated with a topic.

We plotted the relationship between the number of cells assigned to each topic and the number of topics in Figure 4.10. We again, observed that the most stable behaviour appears to occur at 40 topics – in other words, we have a roughly equal distribution of topics with the areas in ranges of 0.25–1km2,1–2km2, and 2–3km2.

According to presented results of our detailed sensitivity analysis, a 500 m resolution was selected as best suited to model places in our study area and also maximised corpus distance for topics. In addition, we selected 40 topics as input for LDA, since it allowed us to generate topics with a roughly constant corpus distance as a function of area. Thus, we could demonstrate that the emerged

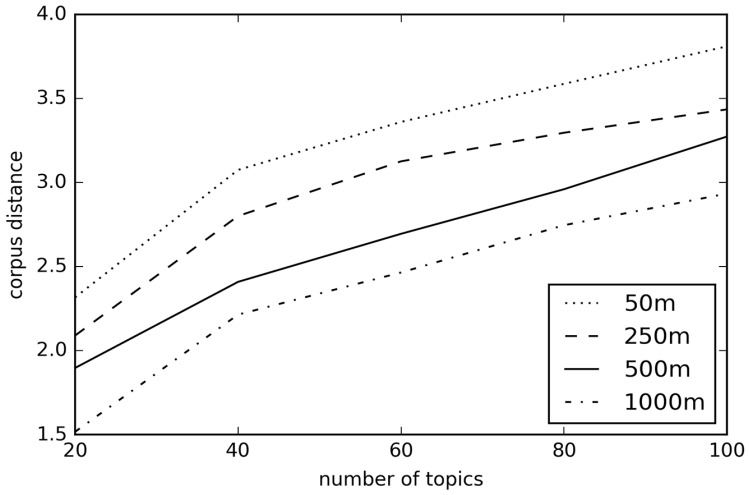


Figure 4.8: Change in median corpus distance for different numbers of topics with respect to grid resolution. Figure from Bahrehdar and Purves (2018) – Publication II.

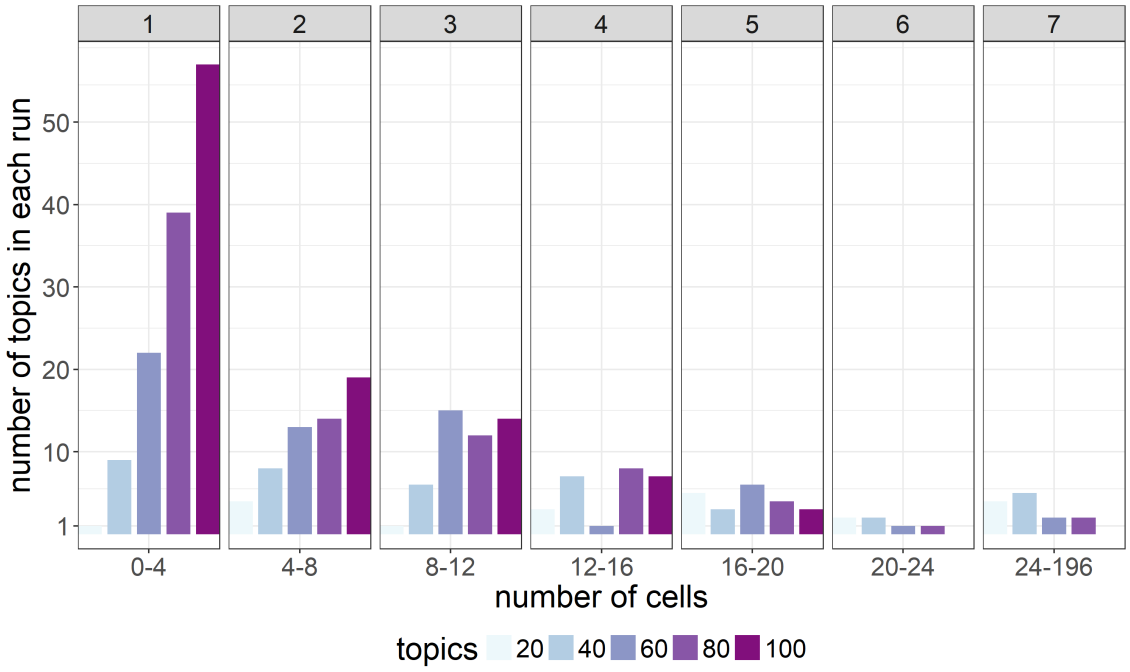


Figure 4.9: Corpus distance for topics associated with different numbers of cells at a resolution of 500m. Figure from Bahrehdar and Purves (2018) – Publication II.

places and their characteristics are not biased to either topics covering only very large or small areas.

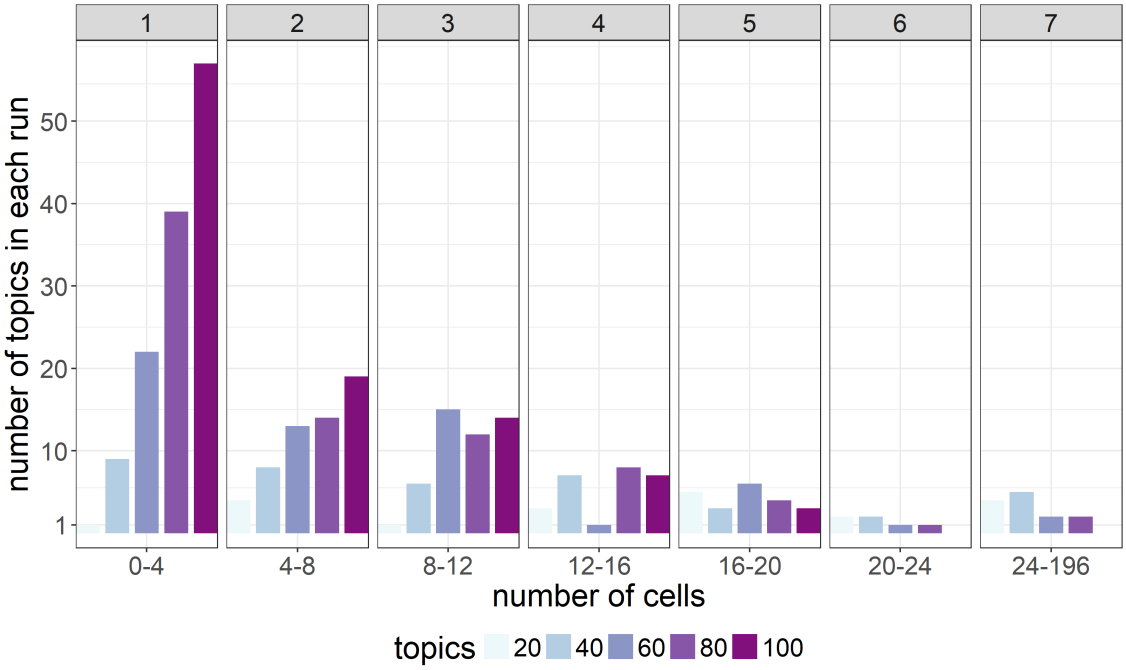
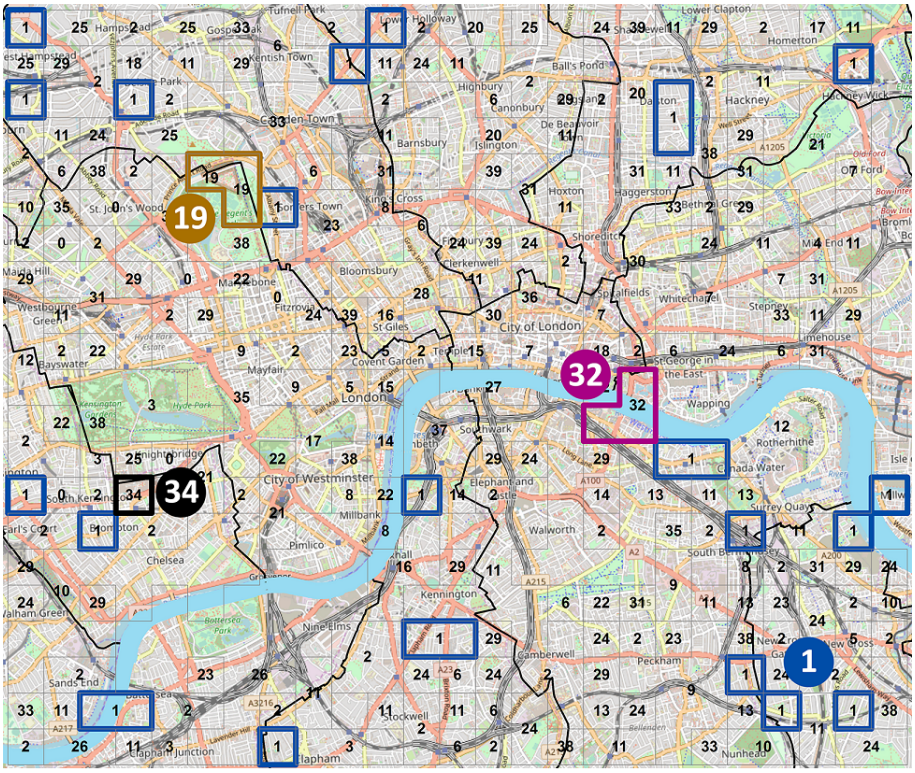


Figure 4.10: Number of topics associated with the 500m cells for each implementation of the model. Figure from Bahrehdar and Purves (2018) – Publication II.

4.2.2 *Labelling and exploring topics*

After selecting an optimum grid resolution and number of topics, we then analysed the meaning of the topics generated. Exploring cumulative probabilities associated with tags in each topic allowed us to select lists of representative tags. Furthermore, we tried to label each topic based on a selected tags list, typically containing 15-25 tags, and using our local knowledge of London. We could label only 30 topics out of 40. Based on a previous study [*Chang et al.*, 2009], we hypothesised that it is more likely to be able to label the topics with low coherence values. Exploring coherence values of the 30 labelled topics and 10 unlabelled ones confirmed that the coherence value is a good indicator of the likelihood of topics being interpretable by humans.

In Figure 4.11, we observed different ways in which properties of place are captured by our grid-based approach. We presented four examples of delineated places as labelled as `view`, `London Zoo`, `along the Thames`, and `South Kensington and museums`. The first example topic was distributed over several locations in London (Figure 4.11b); mostly represent semantics related to general features of scenes like `clouds`, `sunset`, and `skyline` (Figure 4.11a) and indicate generic views of London. Locations affected by this topic, are spread throughout our study area and highlight places from which London is seen and



(a) Labelled topics (numbered cells)

(b) Example topics, labels, and tags

Figure 4.11: Labelled topics (numbered cells) and example topics, labels, and tags (with size as a function of probability).
Figure from Bahrehdar and Purves (2018) – Publication II.

photographed. Therefore, these places can be characterised by what is found and what can be seen from these places. As Opposed to the first example that presents generic semantics of similar locations, in the other three examples, we could characterise specific locations in the form of either a cluster of adjoining cells like topic 19 (London Zoo) and topic 32 (along the Thames) or a single location like topic 32 (South Kensington and museums) (Figure 4.11b). We removed the two most probable tags from topic 19 (zoo) and topic 34 (natural) to increase clarity. Exploring tag clouds, we observe a mixture of mostly proper nouns in the form of place names and building names (e.g. `southbank`, `londra`, `gherkin`), nouns (e.g. `butterfly`, `cloud`, `skyline`, `family`), and more abstract terms (e.g. `assembly`, `authority`) (Figure 4.11a).

We furthermore, associated our labels with a simple taxonomy based on previous studies [*Agnew*, 2011; *Harrison and Tatar*, 2008], which allows us to understand the nature of terms with respect to aspects of a place. We chose to represent the labels using five dimensions of places: location, locale, activity, sense of place, and people. Figure 4.12 illustrates the contrast between, for instance, topics based around locations (e.g. Barbican, Piccadilly), locales (e.g. canals, trains, and stations), and combinations of locations and locales (e.g. Hyde Park, which contains both location and locale information). We could not find a topic that could be interpreted as "emotions and feeling" to further be classified as related to sense of place. However, we could propose a mixture of different dimensions.

As has been shown in previous research [*Sigurbjörnsson and Van Zwol*, 2008], toponyms provide an important way of describing images, and thus can be effectively used as labels for topics. However, the map also allows us to see that such topics can extend beyond the actual location associated with a toponym (e.g., as occurs for Piccadilly), thus suggesting that such topics describe both the place Piccadilly and other similar places. We suggest that some of the classes also seem likely to reflect different sorts of users. Views, canals, trains, and stations are distributed across London and seem likely to be indicative of locals interested in certain sorts of views and narratives about the city, rather than visitors characterising tourist attractions (e.g. London Zoo or the museums in South Kensington). However, this visualisation also illustrates some of the challenges of extracting semantics from tags, where we can only assign labels by interpreting and making assumptions about associations between tags. In general, we also note that most of the activities are leisure activities, suggesting that Flickr is typically used to document a mixture of tourist and leisure activities. This hints at what might be missing in such characterisations (e.g. more mundane activities and those with less positive associations).

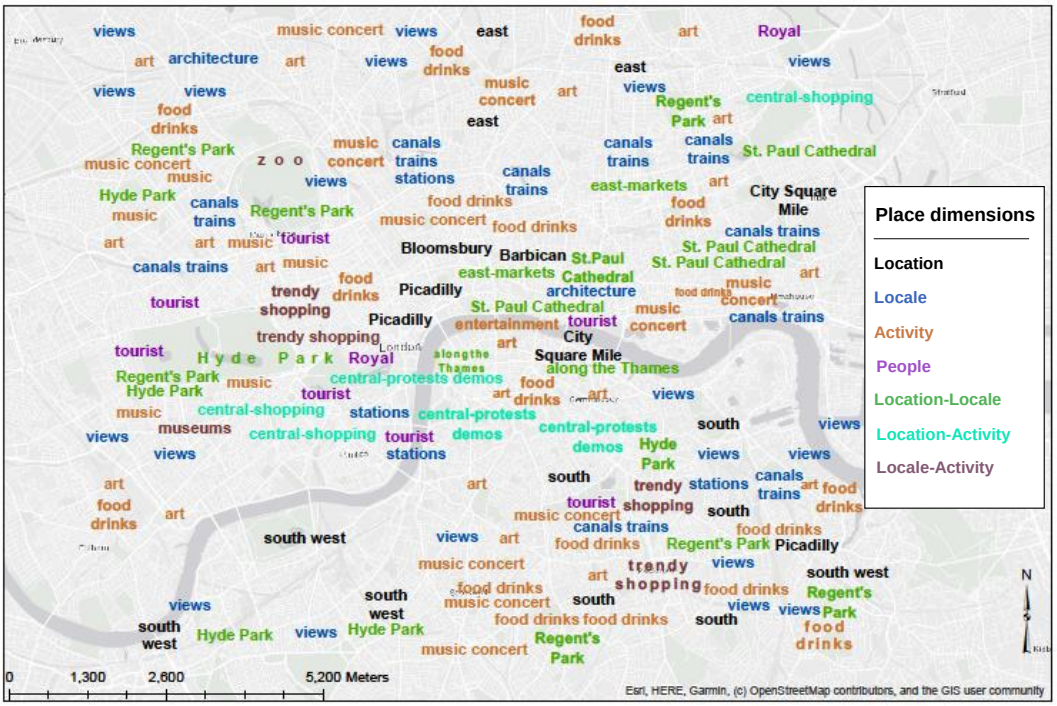
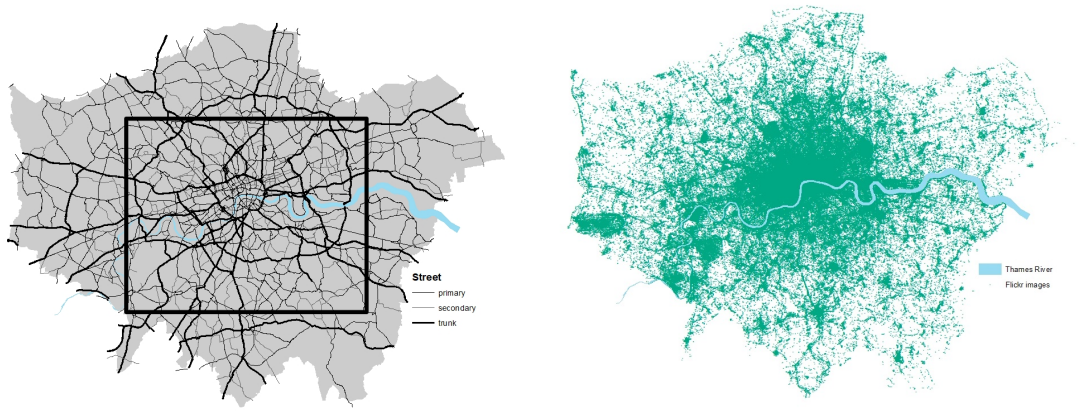


Figure 4.12: Map of London describing users' perception of the space as places. Figure from Bahrehdar and Purves (2018) – Publication II.

4.3 STREETS OF LONDON

The map in Figure 4.13a presents the study area and major roads network of London. The network consisted of 3406 street segments with a median length of 519m. A map of Flickr image footprints in Figure 4.13b) shows the structure of the street network that we used to analyse the location of associated images. The footprints of images illustrate that they are often associate with streets and are concentrated in open spaces (which our model could not capture and represent).

As we explained in section 3.2.3, we created two different datasets from the initial dataset —a collection of metadata of 5,119,629 geotagged Flickr images for 33 boroughs of London downloaded before September 2018 —to test our last approach. The first collection comprised a collection of 2,537,941 images shared by 72,407 users, filtered based on biases related to participation inequality (Table 4.4). 1,250,205 images were taken by 671,207 users within a 100metre buffer of segments. The second collection was strongly filtered with respect to shared semantics. Biases with respect to the shared tags and their associate images and users were deleted. Finally, we were left with a collection of 1,605 unique tags



(a) Street network of major roads longer than 200m

(b) UGC footprints, which follow street network and open spaces

Figure 4.13: Our study area within 33 boroughs of Greater London.
Figure adapted from Bahrehdar et al. (submitted) – Publication III.

from 671,207 images photographed by 36486 users (Table 4.5) to perform our semantic similarity measure.

function	#images	#users
original dataset	5119629	105021
accuracy filtering	4825534	97547
inactive users	4800395	72408
prolific users	4617460	72407
bulk-uploads	2537941	72407

Table 4.4: Remaining numbers of Flickr images and users after removing outliers, noises, and influential biases.

To explore patterns of similarity between four dimensions, we first created a map of correlations between segments with respect to each dimension (semantic, all users, tourists and temporal) and a tag cloud capturing the segments shared by at least 12 of the 30 segments most similar to the given location. We also created a histogram illustrating the ten most similar segments in terms of proportions of images taken on different days of the week. To explain our results, four examples were selected; each example allows us to effectively describe different aspects of our approach according to their particular property.

The first example of our result is `Tower Bridge`, a very popular and well-known location in London (Figure 4.14). The most semantically similar segments to Tower Bridge are along the banks of the Thames, which are linked by bridges that form a sinuous path which can be understood as a path through the city *sensu* Lynch (1960). We also observe, through the tag cloud, that the Thames

function	#images	#users	tags	unique tags
null tags	2062441	58950	20127673	—
duplicated tags	2062441	58950	20127130	—
machine generated tags	1906229	54080	16936308	483960
tag popularity	1726670	51282	8967337	4744
within 100m buffer	853171	41836	8257677	4738
tag importance	671207	36486	4268980	1605

Table 4.5: Remaining numbers of Flickr images, users and representative tags after applying each filtering step to create a "content" dataset

Path and Thames River are indeed tags shared by many of the most similar segments. Some other tags, such as `bridges`, `boats`, `tides` and the `river`, reflect different aspects of this location, for example physical aspects. We could also find more specific tags referring to named locations along the Thames like `victoriaembankment`, `theshard`, and `bankside`, which we desired to remove in our place name filtering step, as described in section 3.1. Tags like `reflection` and `fog` captured different properties of images that are likely to be related to water; and tags like `canon`, `blackandwhite`, and `nightshot` related to photography itself, which could arguably be filtered out. Based on our semantic similarity measure, we could capture a district which can be interpreted meaningfully as a path.

The other two maps in Figure 4.14 illustrate similarities based on users and tourists show weaker correlations consistent among all four examples. Users in general clustered around Tower Bridge and to the west and north of the river; however, some users crossed the Thames to the south. But tourists, as we expected, moved in a smaller area, mostly in central London and nearby places north of the river. These maps reflect the Thames as a barrier to people, with users much less likely to visit seemingly similar regions.

Figure 4.18 shows that correlations for many segments are high, but we see few spatial patterns. The associated histogram demonstrates that, on average, more pictures are taken on Saturdays and Sundays than other days of the week at Tower Bridge and similar segments. This behaviour indicates the typical usage of images in our study area as a whole, therefore shows limited spatial pattern in terms of temporal dimension.

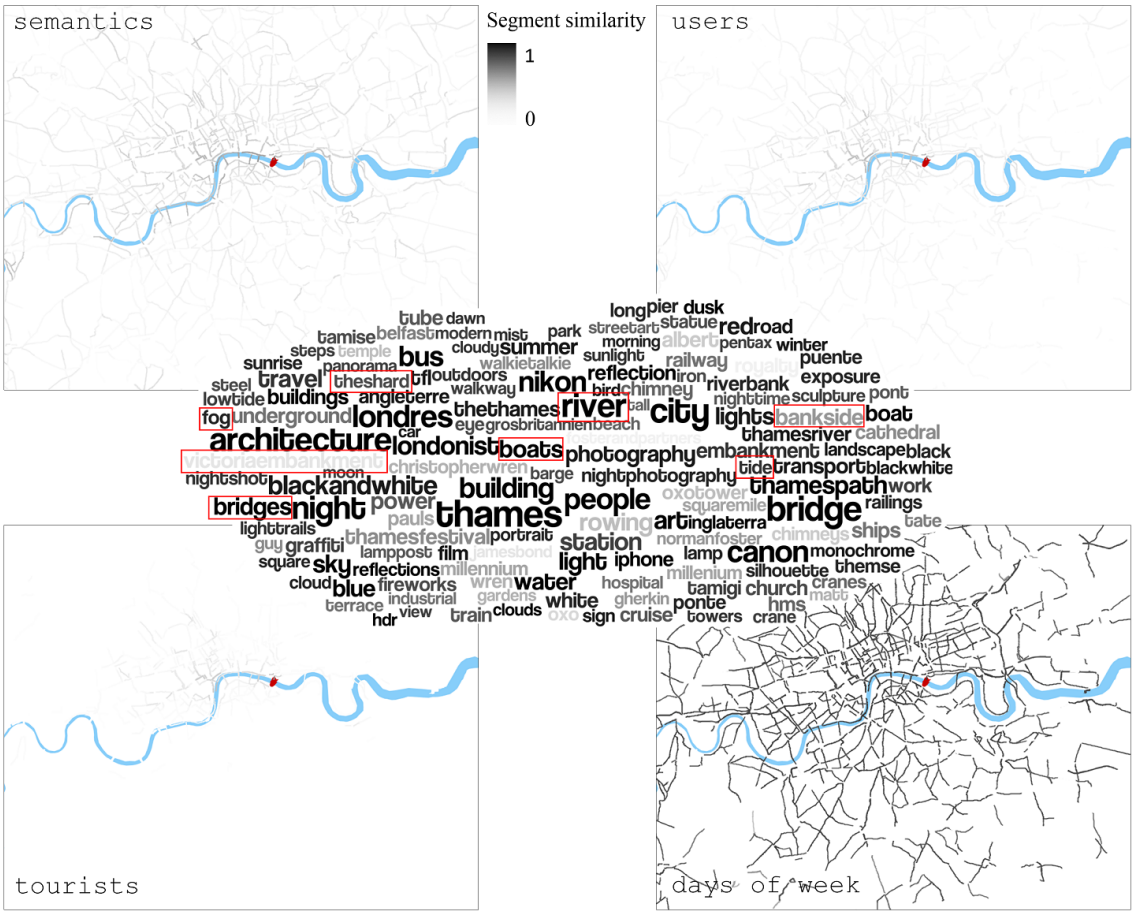


Figure 4.14: Four maps representing the similarity between the queried street in red and all other streets in London. Darker segments are more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Darker tags are shared by more of the top 30 segments and tags highlighted in red are discussed in the text. Figure adapted from Bahrehdar et al. (submitted) – Publication III.

The second example, Chepstow Road, exposes a very small region around Notting Hill through segments with a very strong spatial correlation (Figure 4.15). Exploring the tag cloud associated with this emerged region, we realised that it is the location of the annual Notting Hill Carnival (tagged as `nottinghillcarnival`), and included shared tags like `carnival`, `dancing`, `parade`, and `party`. Therefore, we could capture a district (*sensu* Lynch (1960)) through an event (In the former example, we captured a district through an affordance, e.g. the banks of the Thames and the Thames Path.) The user correlation map shows that people visiting Chepstow Road roam further than touristic places such as Tower Bridge. On the other hand, there is not much similarity between segments based on shared tourists. The histogram in Figure 4.18 presents low temporal correlations between segments and captures the dates of carnival (Sunday and Monday).

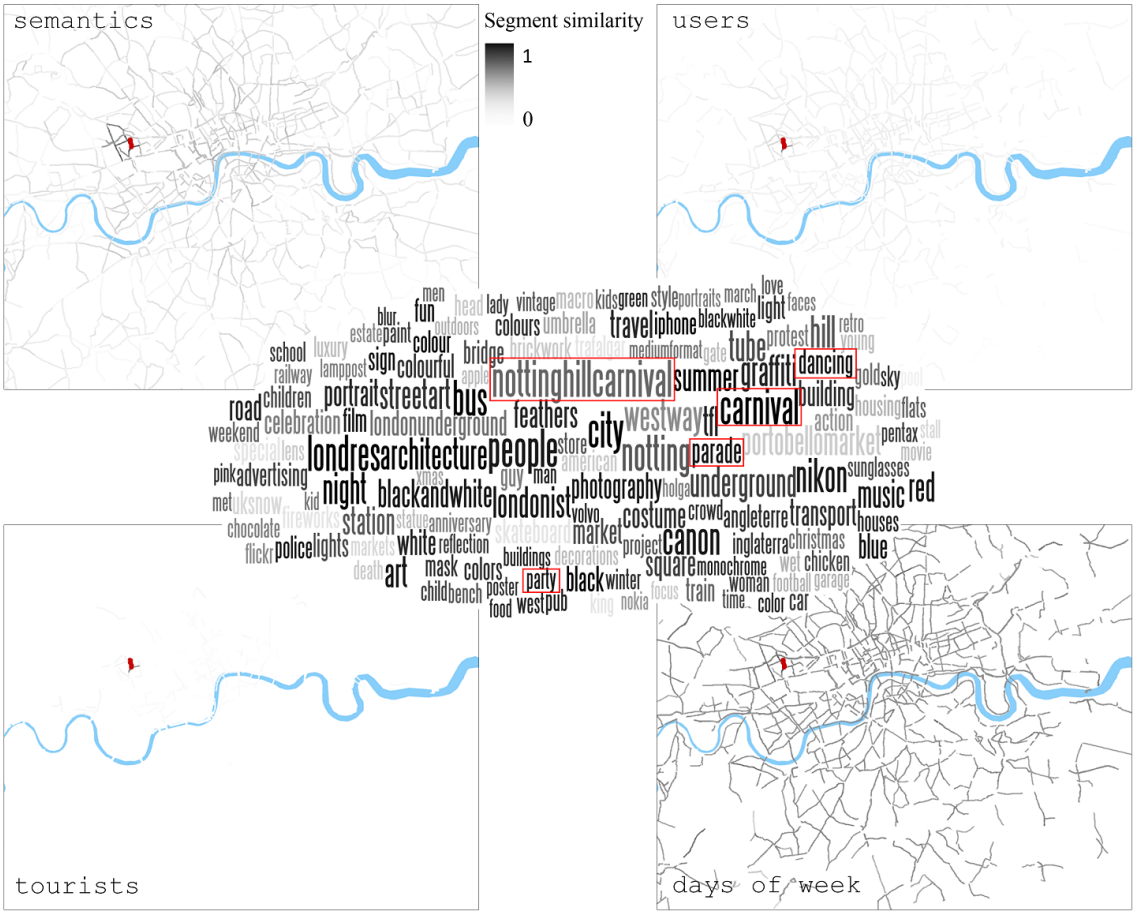


Figure 4.15: Four maps representing the similarity between the queried street in red and all other streets in London. Darker segments are more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Darker tags are shared by more of the top 30 segments and tags highlighted in red are discussed in the text. Figure adapted from Bahrehdar et al. (submitted) – Publication III.

Our third example, Whitehall, located in the heart of London, is associated with both political and ceremonial events (Figure 4.16). Semantically, we can distinguish a region around central London that spans both sides of the Thames. Tags describing Whitehall, similar to tags related to Tower Bridge, are related to both photography and named locations in area region (e.g., `oxfordst`, `stjamespark`, `hydeparkcorner`). Many tags capture different ways, in which this region is used in terms of recurring and rare events (e.g., `celebration`, `royalwedding`, `parade`, `protest`) and their participants (e.g., `soldier`, `queen`, `guards`). Similar to Chepstow Road, users visiting this segment capture larger areas than those photographing Tower Bridge. Tourists focus on the northern side of the river: again, it seems that the Thames creates a division for people's movement. Aside from Chepstow Road, the temporal correlation for Whitehall is similar to the general pattern of all locations, where most pictures are taken on Saturday and Sunday.

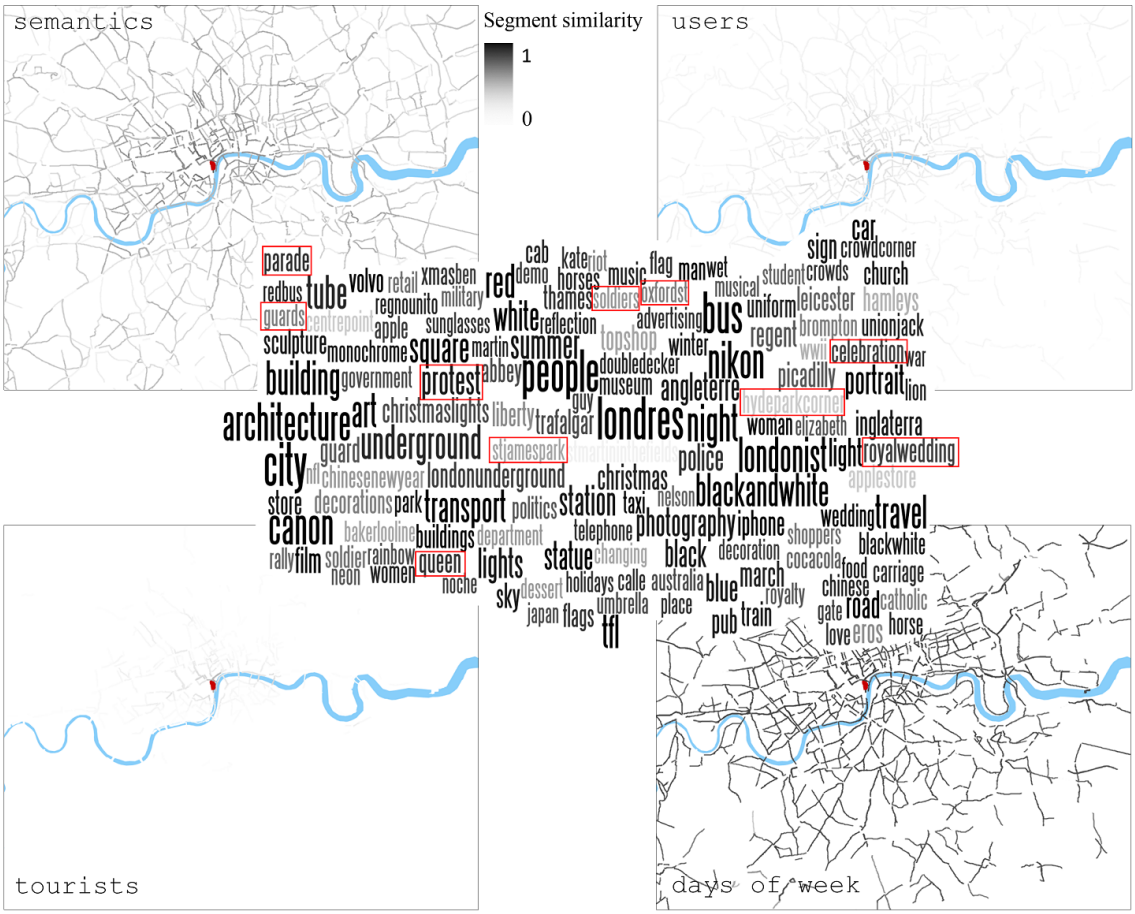


Figure 4.16: Four maps representing the similarity between the queried street in red and all other streets in London. Darker segments are more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Darker tags are shared by more of the top 30 segments and tags highlighted in red are discussed in the text.
Figure adapted from Bahrehdar et al. (submitted) – Publication III.

Our final example (Figure 4.17), Crystal Palace Parade, reveals a very different pattern to the previous three examples that allowed us to identify coherent regions associated with semantically similar segments. In this case, the Crystal Palace Parade segments are distributed, seemingly randomly, across all of London. However, the tag cloud associated with the most similar segments reveals the reason for this pattern. Other than common tags related to photography, we find here many tags related to transport, including buses, types of buses (e.g., `scania`, `plaxton`, `routemaster`, `mercedes`, `volvo`) and providers of public transport (e.g., `arriva`, `londontransport`, `stagecoach`, `abellio`). Semantic similarity in this location is thus defined by photographs of a particular type, taken by a specialist group interested in public transport. Users present at this location travelled not only in south London, but in north London as well, demonstrating an asymmetry in the barrier effect of the Thames: it apparently limits movement from north to south more than south to north. Note that since semantic similarity and user similarity have very different patterns, that our method implicitly shows that different photographers are interested in the same subject matter. Tourists taking pictures at this location appear to be rare, and thus have a very limited local spatial spread. Temporally, we note similar patterns to Tower Bridge and Whitehall, though with a noticeable secondary peak midweek.

Different dimensions of our datasets provide us an opportunity to identify districts (coherent regions of semantic similarity or dispersed locations, in the form of street segments that share the same identity captured through semantics). In cases like Whitehall, these districts are similar across different dimensions. Contrarily, they might have significant differences like Tower Bridge or Crystal Palace. Tag clouds capturing semantic similarity, including place names, reflect both landmarks and concepts related to less salient locations due to their identity, e.g. the buses of Crystal Palace. We demonstrate that the Thames River emerges both as a path (in the case of Tower Bridge) and an edge (again for Tower Bridge, but also Whitehall) dividing the city in two.

Figure 4.17: Four maps representing the similarity between the queried street in red and all other streets in London. Darker segments are more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Darker tags are shared by more of the top 30 segments and tags highlighted in red are discussed in the text. Figure adapted from Bahrehdar et al. (submitted) – Publication III.

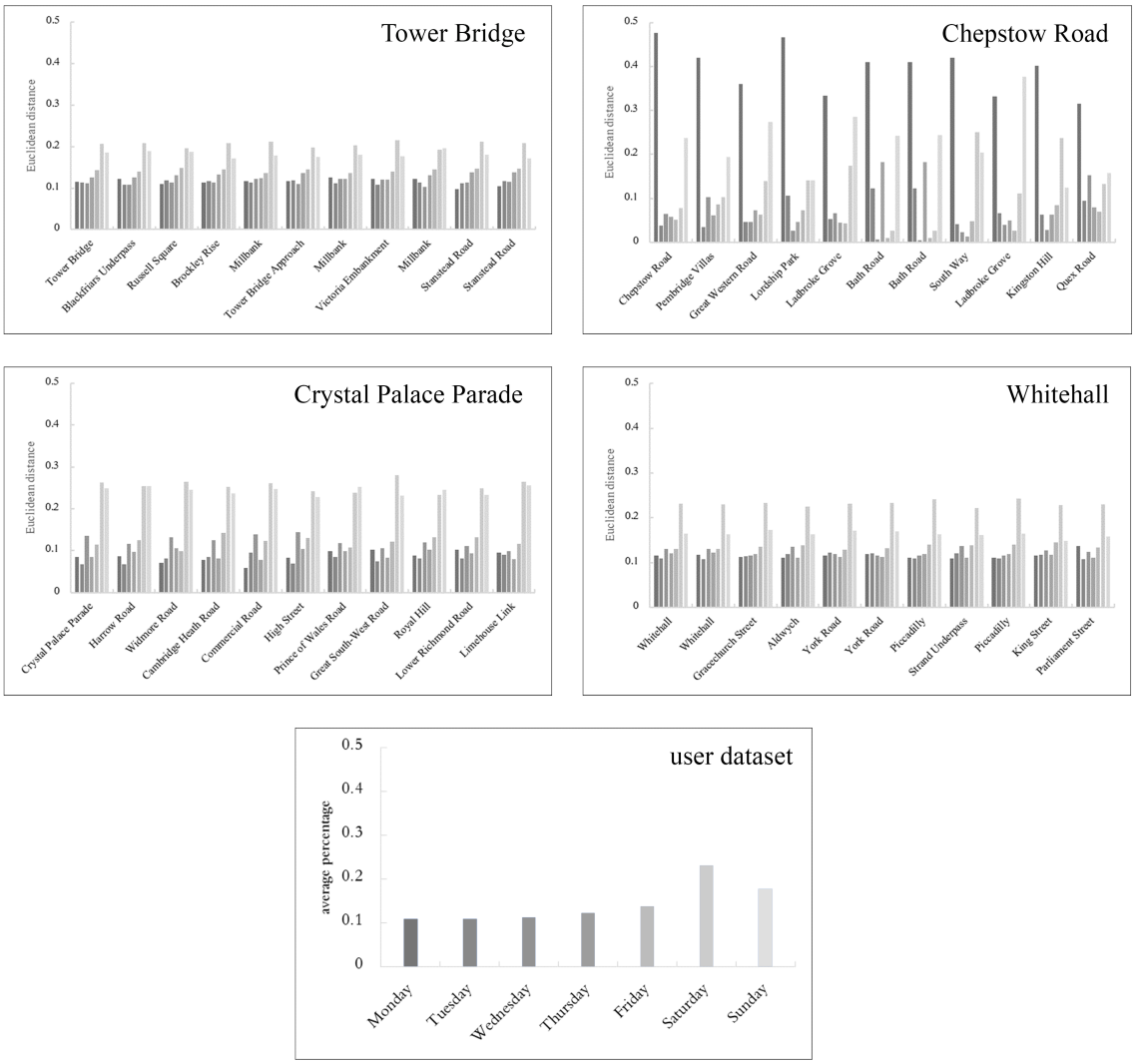


Figure 4.18: Histogram of daily images taken for the ten most similar segments to query segments as captured by temporal similarity for each of the four examples, and the overall temporal distribution of images in the collection. Figure adapted from Bahrehdar et al. (submitted) – Publication III.

5

DISCUSSION

Our aim was to demonstrate that user-generated content in the form of metadata, associated with geotagged Flickr images, contains sufficient information for exploring various aspects of a place in urban area such as spatial, temporal and user behaviours in urban areas. Moreover, we aimed to explore different ways of extracting such information based on both geographical models or a pure data driven approach. We could then furthermore present how place-relevant information can elevate the performance of geographic information systems by integrating human experiences into the models based on coordinates and geometries.

To achieve these goals, we performed three place-based models (object-, grid-, and street-based) on our two datasets of Greater City of London, which were collected in different time spans: one contains all metadata before July 2014, and the other before September 2018. Our datasets covered two overlapping study areas: a defined bounding box overlaying the central part of London and all 33 boroughs of London. The first dataset was used for object and grid-based approaches and the second was used for a street-based approach. To capture descriptive information of places, we conducted analysis only on English language data.

In the following, we first discuss our choices concerning time spans and study areas associated with our datasets. Next, we discuss the richness of Flickr metadata for extracting information in the level of *specific of*, *generic of* and *about* [*Shatford*, 1986]. Then, we compare adopted approaches with respect to methods used for linking individual descriptions to places by aggregating information assigned to locations to extract shared descriptions, the granularity of extracted information. Finally, we explain the limitations of our study.

As appointed by Delafontaine et al. (2012), the temporal property of a place (e.g. opening hours, cyclical events) limits human behaviours and accordingly, influences the meaning of a place [*Lansley and Longley*, 2016; *Resch et al.*, 2016]. Our work is based on multi-year collections of images (explained in detail in section 3.1), which allowed us to capture short temporal dynamics in London that are reflected in human activities.

Analysing long sampling periods that include natural language text associated with locations can also be used to study changes in language use [*Nguyen et al.*, 2013] rather than simply changes in the ways people perceive places. Natural language in the form of tags is not as rich as narratives [*Wartmann et al.*, 2018],

but it can be very useful to show how particular vocabularies have been used to explain concepts or things related to places (e.g. city core concepts [*Hollenstein and Purves*, 2010]). One possible opportunity for future studies is to explore the variation of tags being used for specific regions or places over time in addition to the tagging behaviour in different places at the same time [*Marlow et al.*, 2006].

Similar to related work [*Capineri*, 2016; *Resch et al.*, 2016; *Huang*, 2016], we focussed on places within a relatively small geographical extent, such as a region around the river Thames in inner London and boroughs of London. We argue one of the reasons why modelling places at a small scale is necessary is due to the inequality of data production [*Graham et al.*, 2012], which can lead to issues related to data biases.

Collected individual data points were georeferenced explicitly through coordinates attached to images and through place descriptions, either administrative or cognitive [e.g., *Gao et al.*, 2017] and with different granularities (e.g. a path along the river Thames or a neighbourhood in London). We argue that, in cases where geographical location is presented as metadata, we should pay attention to the ways in which these points are then linked to places. In this study, we used different ways to localise our dataset with respect to the selected study area and selected approach.

5.1 COMPARING THE WAYS OF CHARACTERISING PLACES

We used two approaches to explore place-related concepts and their properties: (1) a purely data-driven approach [*Adams and McKenzie*, 2013], grid-based, for both linking data points to space and deriving places from data —both for specific places (e.g. instances of places like `Hyde Park`) and for generic classes of places characterised by an activity (e.g. where people go cycling) or an element (e.g. river); (2) an approach in which meaningful geographical models (such as visibility of a location or street network) were reinforced in the process of linking data and capturing places [*Shelton et al.*, 2015].Having data linked to named places (either landmarks, tourist attractions or named streets), combined with our data-driven approaches, we were able not only to explore properties associated with specific places (e.g. `Piccadilly Circus` or `Whitehall`)), but also to emerge similar places sharing the same properties (e.g. instances of places similar to Whitehall, where people protest) [*Adams and McKenzie*, 2013; *Dunkel*, 2015].

Our pure data-driven approach allowed us to explore parts of London with no need of having prior knowledge about London and physical settings: for example, we did not need to know locations of landmarks or street names ahead of time. To localise individual images and their metadata, we treated cells as documents, including all tags of images within the cells. Having a generated

corpus containing cells covering all study areas, we performed a topic model (e.g. LDA), which allowed us to take into account co-occurrences of tags within the study area and to generate examples of coherent topics inferring spatial (e.g. park, bar, zoo) and social (e.g. protest, concert) aspects of a place (presented in Figure 4.10) [*Adams and McKenzie*, 2013]. In our grid-based approach, we examined the potential of extracting thematic descriptive information (e.g. words describing zoo or park-like places); therefore, we ignored temporal dynamics of place semantics. This, however, can be done by using attached timestamps [*Lansley and Longley*, 2016; *Adams and McKenzie*, 2013].

In contrast, approaches to delineating places and characterising them using geographical urban models —such as digital surface models or a street network —require accessibility to sources of infrastructure data, both for localising data and linking the localisation to places. In our study, we used a VGI source like OpenStreetMap to collect geometries (such as landmarks, tourist attractions, and road layers). Note that the OSM dataset similar to most UGC sources suffers from uneven geographical coverage at the global scale [*Haklay*, 2010]. Such inequality in data production limits the use of OSM in research.

Using a geographic model based on a visibility concept enabled us to turn the problem of uncertainty of the location of an image —whether the attached geotag refers to the content of a photo or to the photographer [*Kisilevich et al.*, 2010] —into an advantage. This model reinforced the idea that locations are linked through their viewsheds. Such links demonstrate the impact of a place on another [*Hu*, 2018] and reflect characteristics of both locations. With the aim of describing specific landmarks, we used polygonal geometries (capturing areas from which a given location can be seen). Together with its place name as a keyword (e.g. Capineri (2016)), we linked images and their textual information (such as titles, descriptions and tags) to the regions, and finally, to a given place (like Big Ben). Similarly, starting from streets as places, we took into account the structure of London and linked metadata to segments of the streets. While grid cells and clustered words provided an insight into the ways in which a city was understood (for example, the cells representing a general view of London or the ones portraying the Thames River), a street network reinforced a spatial concept that places are experienced and perceived through paths [*Lynch*, 1960] that people walk along through a city.

From the granularity perspective, we explored place properties at various levels. Visible areas associated with entities (i.e. geometries assigned to named places) varied from a single region overlaying the geometry of a given location (e.g. viewsheds of `Hyde Park`), to a set of regions, possibly separated, covering a larger proportion of study area (e.g. areas visible from `Big Ben`) due to their elevation compared to their surroundings. We characterised each set of regions linked to a location whether it was only one small region or multiple regions distributed over the study area using lists of *elements*, *qualities* and *activities* [*Purves et al.*, 2011]. It is important to point out that these entities did not have contigu-

ous properties [c.f. *Gao et al.*, 2017]; for example, sub-regions associated with `Tower Bridge` contain both vantage points and walking paths. In the case of `Tower Bridge`,a location of high elevation, this issue becomes more important, since several regions were described through such lists (elements, qualities, and activities). Differentiating and assigning such properties to each region is a complex task. However, in the case of `Hyde Park`, it was pretty straightforward to capture a coherent and distinctive theme due to the relatively homogeneous geographical context.

Moreover, we explored properties associated with grid cells through four resolutions. Similar to a previous approach whereby we found that the extracted properties of big or disperse regions were less coherent or distinctive, the topics assigned to coarse granular cells were not also coherent enough to annotate as one single theme. However, this is a valid problem in the sense that the topics become too general. We would argue that the challenge here is more about choosing an optimum cell size that is big enough to contain data while being small enough to distinguish places with distinctive semantics from their surroundings.

Carrying out a sensitivity analysis concerning the effect of grid resolutions and the expected number of thematic subjects as topics, we studied the quality of generated topics and whether they are understandable as a coherent theme by humans. To do so, we used local knowledge and labelled each topic. Furthermore, we measured corpus distance per topic. Comparing corpus distances of both labelled and unlabelled topics showed that topics with higher corpus distances are labelled. This observation indicates that corpus distance is a suitable indicator for quality of topics. Despite semantic relationship between words in each topic, it is important to note that all the words in a topic do not necessarily occur in all associated cells (similar to the extracted properties associated with visible areas). Thus, having explored topics and assigned words in different resolutions, we could explore very detailed and specific properties about locations at a very fine granularity. We could furthermore zoom out to gain a very generalised overview of bigger areas.

Finally, we focussed on capturing three contrasting spatial, temporal, and user behavioural patterns in London with respect to segments of main roads (such as primary, secondary and truck) that were longer than 200m. Removing small segments yielded some pseudo-nodes that we deleted. Therefore, we lost some junctions or nodes which could help us to capture popular locations like `Trafalgar Square` [*Crandall et al.*, 2009]. We also missed open spaces like `Hyde Park`, since we aggregated tags associated with streets.

Extracting semantics related to places often requires interpretation by authors in useful and thought-provoking ways. In terms of interpreting the results of all three approaches, we first used visual summaries in the form of tag clouds [*Keim et al.*, 2008], such as the most probable tags associated with each topics

[*Rattenbury and Naaman*, 2009] or histograms of temporal behaviours [*Lansley and Longley*, 2016] to represent details associated with different locations. Secondly, we benefitted from general knowledge about London. For example, having prior knowledge about the relationship between `Whitehall` and ceremonial events helped us to interpret semantics and detect the effect of the Thames as a barrier to north-south movement. Finally, to meaningfully interpret extracted semantics as place properties, we supplemented our knowledge with research, for example by identifying the potential relationship between `Cheptstow Road` and the `Notting Hill Carnival`. In case of topics associated with locations, we used an annotation task [*Chang et al.*, 2009] to interpret our results and further explored the semantic coherency of words in each topic by measuring associate corpus distance. Our work that modelled places based on topics showed that labelled topics had high coherence values and provided a general and comprehensive understanding of associated locations, which is missing in the visibility approach. Performing similarities based on streets allowed us to interpret the results even when there was no data. For example, studying user behaviours shows where users often go and take photos. Explaining the reason for what is happening, however, requires deep knowledge of a city.

Concerning implementation, calculating a viewshed for a given location requires both the geometry of observers (i.e. objects like landmarks or tourist attractions) and digital surface models of study areas. Another concern regarding viewshed calculation is that the computation time increases as both the extent of the study areas and the complexity of geometry of observer locations grow. In the street-based approach, we modelled places based on road layer of OpenStreetMap, which was freely available.

To calculate similarities, a 3406×3406 matrix was generated, which was impossible to understand without visualisation, which is not a simple task. This complexity grows when three semantic, temporal and user behavioural patterns needed to be simultaneously visualised. To model places based on grids, we were independent of other sources than Flickr.

5.2 PLACE DIMENSIONS AND PROPERTIES EXTRACTED FROM UGC

Our study is consistent with the discussed state of the art (section 2.2.2) and shows the richness of UGC data in the form of metadata attached to images for describing different parts of London. Results from each modelling approach can be used to characterise a different aspect of a place (discussed in detail in section 2.1) related to the *where*, *when* and *who* facets at three level of information: *specific of*, *generic of*, and *about* [*Shatford*, 1986]. At the level of *specific of*, we were able to delineate regions associate with named places like `Hyde Park`, `London Zoo` or `Tate Modern` through either footprints of images that are tagged by place names or clusters of cells associated with most probable topic related to a place name. Furthermore, we could explore the relationship between places

from which people walk or have a viewshed, for example a pedestrian path related to Tower Bridge (see Figure 4.3).

At the level of *generic of*, we emerged regions which shared thematic characteristics [*Adams and McKenzie*, 2013]. These regions can be spatially dispersed like separated clusters where we can have a view of London (e.g. the topic `Views` in Figure 4.10) or they can be located in one single cluster (e.g. the topic `Zoo` in Figure 4.10) that represents a particular context. Tags describing elements located in such regions are often dominated by basic level features [*Roche*, 2016] like parks, bridges, rivers or canals. Thus, we applied various text analysis methods to identify representative tags associated with locations. Finally, by comparing such information, we could detect regions characterised by events [*Andrienko et al.*, 2010]: either periodic events like the `Notting Hill Carnival` or irregular events, like the protests in `Whitehall`, where a district emerges.

We also explored locations with respect to *about* information. We showed the popularity of locations using a simple count of the number of images or users (see Table 4.2), which potentially can indicate abstract notions like of cultural ecosystem services [*Gliozzo et al.*, 2016]. By classifying users in our dataset into two classes of locals and non-locals and studying their movements patterns around London, we were able to explore the semantics associated with preferences for segments of streets [*Adams and McKenzie*, 2013; *Dunkel*, 2015]. Exploring temporal dynamics at the level of weekdays, we demonstrated increased activity over the weekend related to leisure activities, but showing little variation in space.

5.3 IMPLICATIONS: OPPORTUNITIES FOR PLACE-BASED MODELLING

Our work demonstrated that-user generated content provide sufficient data to extract properties of experienced locations, with different granularities (e.g. visible regions and street segments), in different levels (such as specific of, generic of, and about). We studied spatial footprints that reflect the visual salience of places in relation to geometries representing places in OpenStreetMap using our object-based approach. We argue that such findings could influence performance of map generalisation operators to better represent place related information on maps by focussing on not only geometry, but also on semantics contributed by large numbers of users. For example, we demonstrated that `Kensington Gardens` is often perceived by visitors as a part of `Hyde Park`. Therefore, the aggregated geometry assigned to `Hyde Park` representing the perceived region containing `Kensington Garden` as well as both place names which are recognised by a group of users can reflect their perceptions in a place-based map for London, which can be used to enlarge or exaggerate an object in all/some directions (i.e. enlargement/exaggeration operators), if needed. We suggest that future work conduct research on the ways of integrating generalization operat-

ors in a holistic generalization process which is the key challenge for generating maps based on place semantics.

Our extracted place properties in the form of specific place names that are commonly used in an area can improve the performance of navigation in LBS to identify a user's destination. Navigation is a very common service [*Basiri et al.*, 2015] in LBS, whereby users actively seek information. Having highly precise and accurate information in relation to both real-time user locations and contexts, these services are able to provide an appropriate route. Assuming that users often travel from the current location, specifying the location of destination in the form of exact coordinates or an absolute address is challenging, since humans communicate about locations through natural language. For example, an alternative route for "let's go to downtown" could be a main street which possibly has low cognitive load, rather than a complex path navigating through streets to reach a particular address.

Extracting properties associated with grid cells allows us to build a hierarchical structure of semantic information and adjust the structure to user needs by zooming in and out. For example, by combining a grid approach and street approach, one service can first benefit from grid topics and identify the location of a place in question (e.g. `Notting Hill Carnival`) on a bigger scale. Furthermore, in case of offering routing systems, it can navigate a user to a more complex target based on semantics of street segments. Selecting a generalised geometry (e.g. a bounding rectangle or an alpha shape [*Twaroch et al.*, 2009; *Keßler et al.*, 2009]) that represents such an initial destination is arguably a valid approach for dealing with the vagueness inherent in vague regions. Potentially, future work could work on a topological street network, in which the connectivity of streets is considered for route finding based on user preference or in combination with temporal information to deliver more detailed place information with respect to temporal dynamics of a place.

Users' geographical footprints can be also analysed for generating required context in tracking services. For example, our findings presented in Figure 4.14 demonstrated movement patterns of both locals and non-locals, reflecting the barrier effect of the Thames on users' mobility. It shows that non-locals (i.e. tourists) are more active in the northern part of the river. Such information suggests a more meaningful way of aggregating users' information based on semantics than by purely using geometry. For example, we can represent places visited by different groups of users based on a bottom-up model [c.f. *Huang*, 2016].

Location-based services in the form of navigation and tracking demand having a suitable data structure to store place information that relatively matches the ways humans (rather than machines) reason qualitatively about places, such as the place graphs originally proposed by Vasardani, Winter, and Richter (2013) and used by Kim, Vasardani, and Winter (2017). Such models are suited to both capturing some notion of vagueness and hierarchy. The key challenge is there-

fore, lying in mapping such data back onto the more precise geometry and network used in typical routing systems. Information classified as *generic of* can be used as additional context in both navigation and tracking. Our work can be used to annotate user behaviours, for example, by classifying all users who visited similar places [c.f. *Adams and McKenzie*, 2013]. We measured similarities between places, not simply based on the type of points of interest, for example, based on semantics —a vector of tags describing locations [*Janowicz et al.*, 2011]. Our work showed the potential use of such information in routing; identified segments associated with a pleasant route possibly can be identified or identifying routes where tourists flows [*Prelipcean et al.*, 2015; *Alivand and Hochmair*, 2013].

Marketing is a major domain area that uses location-based services and products based on a combination of current, past and predicted location combining with context. For example, having both place information at the level of *generic of* related to locations (either in the form of regions, grids or segments of a street), and previous and current location of a user, we can assign activities to both location and users' location. Thus, we are able to generate movement profiles which suggest likely activities (and thus can trigger location-based advertising) [*Kõivumägi et al.*, 2015]. Our work demonstrated that starting from a place-based model has several potential advantages for location-based services: first, having a place model, linking place properties to points of interest (for example a named location like a named hotel) is not essential. We are able to link properties to, for example grid cell.

Using a continuous grid model or a street-based network, properties are therefore, aggregated footprints which allows us to protect individual privacy [*Nussbaum et al.*, 2017], since we use an aggregated version of individuals' contributions. The second, is our place models also provide a more meaningful way of geofencing approaches [*Rosenkrans and Myers*, 2018] —for example, locations, from which Tate Modern can be seen, cells associated with zoo or segments of street in which shopping malls are located —as trigger of advertisements, which are based on the ways places are experienced, rather than administrative boundaries.

Based on our approaches, we suggest different ways of generating context for recommendation systems [*Huang*, 2016; *Ye et al.*, 2011], since we could analyse both attributed places and identify them. We argue incorporating place models for context retrieval in any kind of LBS have fundamental advantages:

1. Our place-based models of locations enabled us to shift the focus away from precise geometric information in LBS such as navigation and tracking, and to generate query footprints. Our work of comparing places based on semantics, temporal and user behavioural patterns demonstrated that UGC captured information about different types of users [*Gao et al.*, 2017], and by filtering users with respect to their behavioural patterns [*Huang*,

2016], we are potentially capable of generating query footprints for user groups and therefore, by exploring their footprints, represent places like central London that are tourist attractions.

2. Both grid-based and street-based approaches allow us to build place-based hierarchies through aggregating similar adjoining cells or street segments. Such hierarchical representation of place, which requires both geometries and semantic of places [*Gao et al.*, 2017] reflect a notion that a place is contained or adjacent to another and thus, we can make proximity queries. Our results demonstrated that such approaches shift the focus away from geometric representation and increases the importance of semantics of places. For example, having semantics associated with grid cells or segments of streets, we could emerge perceptual districts in London by aggregating adjacent cells or segments sharing the same semantics. Having built an appropriate hierarchy of places can be queried for other contained or overlapping places.

3. Place-based models containing place properties related to the *generic of* and *about* (sensu Shatford (1986)), potentially allow us to index documents with respect to specific context in LBS like tourism. For example, identified places characterised as green space, a tourist LBS can return information about such places that evoke positive sentiments [*Lim et al.*, 2018] or place with beautiful views in a location-based context.

In a broader context of implementation of digital earth, such data-driven approaches can contribute to bottom-up citizens perception that flows into a representation [*Craglia et al.*, 2012]. Our results capturing multiple dimensions of a place and linking these dimensions to locations —ranging from dispersed regions linked to a vantage point, to paths through the city and emerged locations with shared semantics —enabled us to analyse both space and places and identify the relations between places [*Salvini and Fabrikant*, 2016]. With the significant increase in the number of users and the amount of user generated content [*Vickery and Wunsch-Vincent*, 2007], we could extract multiple perspectives towards locations (e.g. related to user behaviour). It also allowed us to explore temporal dynamics of places by analysing different time spans of data.

5.4 LIMITATIONS

Our study has several limitations related to both data and our methodological approaches. The most important limitations in our study are the inaccuracies, biases and gaps in data used [*Olteanu et al.*, 2019; *Graham et al.*, 2012; *McKenzie et al.*, 2015; *Zook et al.*, 2010], which might lead to deceiving results [*Boyd and Crawford*, 2012; *Kıcıman et al.*, 2014]. Our study, similar to other place-related research, is inherently vague and subjective. When combined with our data-driven approach, it is prone to such issues [c.f. *Shelton et al.*, 2015]. Despite

efforts to characterise the population based on user profiles (for example, in terms of gender) [*Resch et al.*, 2016], the lack of demographic information is an important limitation of Flickr or any kind of UGC data source [*Olteanu et al.*, 2019; *Kienast et al.*, 2012], and yield to a gap between the target groups (resident and tourists of London.) As we discussed in detail in section 2.2.4, our datasets are not representative of the population, and we do not have any particular information about users aside from a unique identifier.

Concerning the influence of such biases on our study, as we discussed in detail in section 2.2.4, we conducted a set of filtering steps in each approach for extracting desired properties (summarised in Table 3.2). In cases that we explored different dimensions like temporal and use behaviour rather semantic patterns. We chose to generate different datasets and apply different sets of filtering steps. For example, we chose to filter data based both on behaviour (e.g., taking account of participation inequality, bulk uploads and so on) and semantic biases (primarily seeking to retain only tags used by a broad group of users) for exploring various place dimensions (e.g. temporal aspects and user behaviours). Furthermore, in the case of exploring semantics, we filtered the content concerning distinctiveness of semantics related to geography (through topic modelling) or language. We think that the influence and implications of the filtering process have been neglected in analyses of UGC, and argue that the attention given to biases in artificial intelligence studies [*Zou and Schiebinger*, 2018] is equally important in place-related research in GIScience [*Shelton et al.*, 2015].

Non-urban areas are under-represented in Flickr, since they are more focussed on human environments and activities [*Gliozzo et al.*, 2016]. By contrast, according to the literature discussed in section 2.2.4 and based on our results of Greater London, urban areas are represented by more people, which allowed us to better capture the shared views. The bias is not limited to the geography and appear in the content as well. Flickr images —associated with either urban areas or non-urban areas—reflect more positive experiences compared to other sources [*Cox et al.*, 2008], and therefore, we cannot capture negative aspects of places [*Boyd and Crawford*, 2012]. However, it has been argued that exploring populated places that more likely are visited by tourists can lead to further negative feedback in terms of the representation of such places [c.f. *Graham et al.*, 2012]. To overcome the later issue in a very simple way, we can differentiate between residents and non-residents [*Huang*, 2016]. An alternative to approaching such problems is to fill the gap by identifying and integrating various data containing such notions. Some of the sources might be less related to the space being described [*Hahmann et al.*, 2014], or are populated by different user groups in different locations [*van Zanten et al.*, 2016]. For example, we found out that images shared in Flickr through an Instagram link were more related to social activities, while "pure" Flickr images were more likely to describe locations.

We argue this can result in a lack of work on modelling place in global scale [*Graham et al.*, 2012], since there is no a particular source that fits all aspects

of modelling of all dimensions of place. One arguably important challenge facing integration of such heterogeneous data sources is data availability, which became increasingly more critical in any UGC related study concerning user privacy.

A second group of limitations is related to our methodological approaches. Despite a broad range of studies on textual content associated with locations with respect to descriptive information, the analysed language(s) typically were not explicitly specified. English language was dominant, even in locations where it is not the everyday language and does not represent the population as a whole. Selecting English in our study from an English-speaking country, has, some influence: we mostly captured place in the context of English speaking culture, and this language choice might have led to relatively homogeneous place-related concepts due to more advance natural language processing methods in English. Therefore, we might have neglected the diversity present in reality. Despite advances in text analysis, particularly related to natural English language, tag disambiguation is more challenging [*Liang et al.*, 2009]. In our study, we assumed that images are not randomly tagged in different locations, and thus, we used LDA as a classification method for considering word co-occurrences in different locations [*Adams and McKenzie*, 2013]. We then generated semantic themes as context, which minimised the resulting ambiguity. Furthermore, we manually labelled the topics [*Chang et al.*, 2009] and calculated the coherency topics containing tags [*Mimno et al.*, 2011] to assess how representative they are.

A key limitation of any grid-based approach is the Modifiable Areal Unit Problem [*Rattenbury and Naaman*, 2009]. In our study, we conducted a detailed sensitivity analysis based on three different grid resolutions to test the influence of the scale on our results. Based on the results, we assumed that our results are relatively insensitive to the shape and origin of our grid. An alternative for addressing the Modifiable Areal Unit Problem is by using an adaptive grid [*Derungs and Purves*, 2014] and by testing the sensitivity of the results on the grid's origin.

Finally, it is important to consider the impact of not only publishing place-based information that emerges from the content shared by individual users [*Shelton et al.*, 2015], but also of the consequences of integrating such information and some algorithmic solutions in any place-based product. For example, generating pleasant or beautiful routes which avoid certain parts of a city [c.f. *Shelton et al.*, 2015] might reinforce prejudices.

6

CONCLUSION AND OUTLOOK

In this work we explored the potential of user-generated content (UGC) in the form of metadata attached to Flicker images to identify places and to extract their descriptive information. We hypothesised that metadata in the form of tags reflects ways people communicate about places and allows us to capture conceptualisations of a place as a lived and experience location. By drawing upon georeferenced collections of sets of tags, a user identifier, temporal information and local knowledge, we could meaningfully interpret characteristics assigned to modelled places with respect to various place dimensions, suggesting the potential of future implications of UGC for providing place-based products and reasoning with place.

In an initial study, motivated by enriching existing geometries by linking some semantics (e.g. linking descriptive information of Flickr images to OSM geometries), we used a simple geographical concept (viewshed visibility analysis) and explored the relation between known places and their surroundings as well as the richness of attached textual metadata to describe such places. Linking semantics to geometries of places provided us with a novel way of generalising data holistically that shifted away from top-down, administratively generated topographic data to more semantically rich place-related data for map representation of lived locations. Using tags for generating place properties is a valid approach in urban areas, since places are not randomly tagged and the number of images (and accordingly the number of tags) correlates with the popularity of places.

By performing a visibility study, we took into account the impact of a place on another (e.g. vantage point). Furthermore, we moved away from our exploratory study towards a more data-driven approach, analysing the similarity of places modelled as a function of language, using spatially distributed topic models to identify semantically similar regions. This study showed that places indeed emerged by taking into account their semantic similarities (e.g. park-like locations). An important contribution here was made by conducting a detailed sensitivity study on inputs of the model and assessing the utility of a range of functionality measures in describing the quality of place semantics. Using such measures, we studied thematic information concerning the likelihood of humans being able to interpret and label the themes with respect to geographical context.

In a multi-dimensional study capturing semantic similarity, user behaviour and temporal patterns based on a perceptual geographical model of a city (a street network), we explored users' perception of a city not only by detecting streets, but also districts, landmarks and even edges. Reinforcing a street network for organising our data, we shifted away from issues of aggregating data points based on administrative boundaries or imposed tessellations such as grids.

Despite all the limitations of UGC with respect to inequalities and biases, our data for London was sufficiently rich enough to detect various patterns (e.g. using words related to transportation or by using data from users who visited more places in the north of the river Thames) in relation to different place dimensions (e.g. information related to activities or events happening in a location). It is crucial to point out that interpretation of place-related information requires both investigating the data in great detail and using external, preferably local, knowledge.

We suggest that future work explores suitable ways of integrating heterogeneous data from different sources that have various communities and user groups; however, the implications of data biases and data gaps should not be underestimated. Therefore, future work should focus on integrating different sources with respect to the use of UGC in urban planning or applications in GIScience such as generalisations, location-based services, or in digital earth.

REFERENCES

Adams, B., and G. McKenzie, Inferring thematic places from spatially referenced natural language descriptions, in *Crowdsourcing geographic knowledge*, pp. 201–221, Springer, 2013.

Agnew, J., Space and place, *Handbook of geographical knowledge*, *2011*, 316–331, 2011.

Alazzawi, A. N., A. I. Abdelmoty, and C. B. Jones, What can I do there? Towards the automatic discovery of place-related services and activities, *International Journal of Geographical Information Science*, *26*(2), 345–364, 2012.

Aletras, N., T. Baldwin, J. H. Lau, and M. Stevenson, Evaluating topic representations for exploring document collections, *Journal of the Association for Information Science and Technology*, *68*(1), 154–167, 2017.

Alivand, M., and H. Hochmair, Extracting scenic routes from vgi data sources, in *Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, pp. 23–30, ACM, 2013.

AlSumait, L., D. Barbará, J. Gentle, and C. Domeniconi, Topic significance ranking of lda generative models, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 67–82, Springer, 2009.

Andrienko, G., N. Andrienko, M. Mladenov, M. Mock, and C. Pölitz, Discovering bits of place histories from people's activity traces, in *2010 IEEE symposium on visual analytics science and technology*, pp. 59–66, IEEE, 2010.

Andrienko, G., N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom, Thematic patterns in georeferenced tweets through space-time visual analytics, *Computing in Science & Engineering*, *15*(3), 72, 2013.

Antoniou, V., J. Morley, and M. Haklay, Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon, *Geomatica*, *64*(1), 99–110, 2010.

Arampatzis, A., M. Van Kreveld, I. Reinbacher, C. B. Jones, S. Vaid, P. Clough, H. Joho, and M. Sanderson, Web-based delineation of imprecise regions, *Computers, Environment and Urban Systems*, *30*(4), 436–459, 2006.

Bahrehdar, A., and R. S. Purves, Linking vgi for place-based map generalization, 2016.

Basiri, A., T. Moore, C. Hill, and P. Bhatia, Challenges of location-based services market analysis: current market description, in *Progress in Location-Based Services 2014*, pp. 273–282, Springer, 2015.

Bennett, B., and P. Agarwal, Semantic categories underlying the meaning of 'place', in *International Conference on Spatial Information Theory*, pp. 78–95, Springer, 2007.

Blei, D. M., and J. D. Lafferty, Dynamic topic models, in *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, ACM, 2006.

Blei, D. M., A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, *Journal of machine Learning research*, *3*, 993–1022, 2003.

Boy, J. D., and J. Uitermark, Reassembling the city through instagram, *Transactions of the Institute of British Geographers*, *42*(4), 612–624, 2017.

Boyd, D., and K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, communication & society*, *15*(5), 662–679, 2012.

Brill, E., and R. C. Moore, An improved error model for noisy channel spelling correction, in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 286–293, Association for Computational Linguistics, 2000.

Canter, D. B., and J. Groat, The psychology of place architectural press, 1977.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang, A density-based method for adaptive lda model selection, *Neurocomputing*, *72*(7-9), 1775–1781, 2009.

Capineri, C., Kilburn high road revisited, *Urban Planning*, *1*(2), 128–140, 2016.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, Reading tea leaves: How humans interpret topic models, in *Advances in neural information processing systems*, pp. 288–296, 2009.

Chesnokova, O., M. Nowak, and R. S. Purves, A crowdsourced model of landscape preference, in *13th International conference on spatial information theory (COSIT 2017)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Chu, Z., S. Gianvecchio, H. Wang, and S. Jajodia, Who is tweeting on twitter: human, bot, or cyborg?, in *Proceedings of the 26th annual computer security applications conference*, pp. 21–30, ACM, 2010.

Compton, R., D. Jurgens, and D. Allen, Geotagging one hundred million twitter accounts with total variation minimization, in *2014 IEEE international conference on big data (big data)*, pp. 393–401, IEEE, 2014.

Cox, A. M., P. D. Clough, and J. Marlow, Flickr: a first look at user behaviour in the context of photography as serious leisure., *Information Research*, *13*(1), 2008.

Craglia, M., et al., Digital earth 2020: towards the vision for the next decade, *International Journal of Digital Earth*, *5*(1), 4–21, 2012.

Crandall, D. J., L. Backstrom, D. Huttenlocher, and J. Kleinberg, Mapping the world's photos, in *Proceedings of the 18th international conference on World wide web*, pp. 761–770, ACM, 2009.

Cresswell, T., *Place: an introduction*, John Wiley & Sons, 2014.

Davies, C., I. Holt, J. Green, J. Harding, and L. Diamond, User needs and the implications for modelling place, in *Proc. of the International Workshop on Computational Models of Place*, vol. 8, pp. 1–14, 2008.

Davies, J., Display, identity and the everyday: Self-presentation through online image sharing, *Discourse: studies in the cultural politics of education*, *28*(4), 549–564, 2007.

Delafontaine, M., T. Neutens, and N. Van de Weghe, A gis toolkit for measuring and mapping space–time accessibility from a place-based perspective, *International Journal of Geographical Information Science*, *26*(6), 1131–1154, 2012.

Derungs, C., and R. S. Purves, From text to landscape: locating, identifying and mapping the use of landscape features in a swiss alpine corpus, *International Journal of Geographical Information Science*, *28*(6), 1272–1293, 2014.

Derungs, C., and R. S. Purves, Characterising landscape variation through spatial folksonomies, *Applied geography*, *75*, 60–70, 2016.

Di Minin, E., H. Tenkanen, and T. Toivonen, Prospects and challenges for social media data in conservation science, *Frontiers in Environmental Science*, *3*, 63, 2015.

Dittrich, A., D. Richter, and C. Lucas, Analysing the usage of spatial prepositions in short messages, in *Progress in Location-Based Services 2014*, pp. 153–169, Springer, 2015.

Dunkel, A., Visualizing the perceived environment using crowdsourced photo geodata, *Landscape and Urban Planning*, *142*, 173–186, 2015.

Edwardes, A. J., Re-placing location: Geographic perspectives in location based services, Ph.D. thesis, Verlag nicht ermittelbar, 2007.

Elsner, J. B., and A. B. Kara, *Hurricanes of the North Atlantic: Climate and society*, Oxford University Press, 1999.

Elwood, S., M. F. Goodchild, and D. Sui, Prospects for vgi research and the emerging fourth paradigm, in *Crowdsourcing geographic knowledge*, pp. 361–375, Springer, 2013.

Fisher, P., and D. Unwin, *Re-presenting GIS*, John Wiley & Sons, 2005.

Fisher, P. F., Extending the applicability of viewsheds in landscape planning, *Photogrammetric engineering and remote sensing*, *62*(11), 1297–1302, 1996.

Fjørtoft, I., The natural environment as a playground for children: The impact of outdoor play activities in pre-primary school children, *Early childhood education journal*, *29*(2), 111–117, 2001.

Fleure, H. J., Human regions, *Scottish Geographical Magazine*, *35*(3), 94–105, 1919.

Gao, S., K. Janowicz, G. McKenzie, and L. Li, Towards platial joins and buffers in place-based gis., in *Comp@ Sigspatial*, pp. 42–49, 2013.

Gao, S., et al., A data-synthesis-driven method for detecting and extracting vague cognitive regions, *International Journal of Geographical Information Science*, *31*(6), 1245–1271, 2017.

Gibson, J. J., The theory of affordances, *Hilldale, USA*, *1*(2), 1977.

Girardin, F., F. Calabrese, F. Dal Fiore, C. Ratti, and J. Blat, Digital footprinting: Uncovering tourists with user-generated content, *IEEE Pervasive computing*, *7*(4), 36–43, 2008.

Gliozzo, G., N. Pettorelli, and M. Haklay, Using crowdsourced imagery to detect cultural ecosystem services: a case study in south wales, uk, *Ecology and Society*, *21*(3), 2016.

Go, A., R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford*, *1*(12), 2009, 2009.

Goodchild, M. F., Citizens as sensors: the world of volunteered geography, *GeoJournal*, *69*(4), 211–221, 2007.

Goodchild, M. F., Formalizing place in geographic information systems, in *Communities, neighborhoods, and health*, pp. 21–33, Springer, 2011.

Graham, M., S. Hale, and M. Stephens, Featured graphic: Digital divide: the geography of internet access, *Environment and Planning A*, *44*(5), 1009–1010, 2012.

Gustafson, P., Meanings of place: Everyday experience and theoretical conceptualizations, *Journal of environmental psychology*, *21*(1), 5–16, 2001.

Hahmann, S., R. Purves, and D. Burghardt, Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes, *Journal of Spatial Information Science*, *2014*(9), 1–36, 2014.

Haklay, M., 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, 2010.

Haklay, M. E., Why is participation inequality important?, Ubiquity Press, 2016.

Harrison, S., and D. Tatar, Places: people, events, loci–the relation of semantic frames in the construction of place, *Computer Supported Cooperative Work (CSCW)*, *17*(2-3), 97–133, 2008.

Harvey, F., and U. Wardenga, Richard hartshorne's adaptation of alfred hettner's system of geography, *Journal of Historical Geography*, *32*(2), 422–440, 2006.

Hausmann, A., R. Slotow, J. K. Burns, and E. Di Minin, The ecosystem service of sense of place: benefits for human well-being and biodiversity conservation, *Environmental conservation*, *43*(2), 117–127, 2016.

Hauthal, E., and D. Burghardt, Mapping space-related emotions out of user-generated photo metadata considering grammatical issues, *The Cartographic Journal*, *53*(1), 78–90, 2016.

Herbertson, A. J., The major natural regions: an essay in systematic geography, *The Geographical Journal*, *25*(3), 300–310, 1905.

Hill, L. L., *Georeferencing: The geographic associations of information*, Mit Press, 2009.

Hobel, H., P. Fogliaroni, and A. U. Frank, Deriving the geographic footprint of cognitive regions, in *Geospatial data in a changing world*, pp. 67–84, Springer, 2016.

Hollenstein, L., Capturing vernacular geography from georeferenced tags, Ph.D. thesis, Geographisches Institut der Universität Zürich, 2008.

Hollenstein, L., and R. Purves, Exploring place through user-generated content: Using flickr tags to describe city cores, *Journal of Spatial Information Science*, *2010*(1), 21–48, 2010.

Hong, L., G. Convertino, and E. H. Chi, Language matters in twitter: A large scale study, in *Fifth international AAAI conference on weblogs and social media*, 2011.

Hu, Y., Geo-text data and data-driven geospatial semantics, *Geography Compass*, *12*(11), e12,404, 2018.

Hu, Y., S. Gao, K. Janowicz, B. Yu, W. Li, and S. Prasad, Extracting and understanding urban areas of interest using geotagged photos, *Computers, Environment and Urban Systems*, *54*, 240–254, 2015.

Huang, H., Context-aware location recommendation using geotagged photos in social media, *ISPRS International Journal of Geo-Information*, *5*(11), 195, 2016.

Janowicz, K., M. Raubal, and W. Kuhn, The semantics of similarity in geographic information retrieval, *Journal of Spatial Information Science*, *2011*(2), 29–57, 2011.

Jenkins, A., A. Croitoru, A. T. Crooks, and A. Stefanidis, Crowdsourcing a collective sense of place, *PloS one*, *11*(4), e0152,932, 2016.

Jones, C. B., H. Alani, and D. Tudhope, Geographical information retrieval with ontologies of place, in *International Conference on Spatial Information Theory*, pp. 322–335, Springer, 2001.

Jones, C. B., R. S. Purves, P. D. Clough, and H. Joho, Modelling vague places with knowledge from the web, *International Journal of Geographical Information Science*, *22*(10), 1045–1065, 2008.

Jordan, T., M. Raubal, B. Gartrell, and M. Egenhofer, An affordance-based model of place in gis, in *8th Int. Symposium on Spatial Data Handling, SDH*, vol. 98, pp. 98–109, 1998.

Jorgensen, B. S., and R. C. Stedman, Sense of place as an attitude: Lakeshore owners attitudes toward their properties, *Journal of environmental psychology*, *21*(3), 233–248, 2001.

Keim, D. A., F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, Visual analytics: Scope and challenges, in *Visual data mining*, pp. 76–90, Springer, 2008.

Kelm, P., V. Murdock, S. Schmiedeke, S. Schockaert, P. Serdyukov, and O. Van Laere, Georeferencing in social networks, in *Social Media Retrieval*, pp. 115–141, Springer, 2013.

Kennedy, L., M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, How flickr helps us make sense of the world: context and content in community-contributed media collections, in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 631–640, ACM, 2007.

Keßler, C., P. Maué, J. T. Heuer, and T. Bartoschek, Bottom-up gazetteers: Learning from the implicit semantics of geotags, in *International Conference on GeoSpatial Sematics*, pp. 83–102, Springer, 2009.

Kıcıman, E., S. Counts, M. Gamon, M. De Choudhury, and B. Thiesson, Discussion graphs: Putting social media analysis in context, in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

Kienast, F., B. Degenhardt, B. Weilenmann, Y. Wäger, and M. Buchecker, Gis-assisted mapping of landscape suitability for nearby recreation, *Landscape and Urban Planning*, *105*(4), 385–399, 2012.

Kim, J., M. Vasardani, and S. Winter, Similarity matching for integrating spatial information extracted from place descriptions, *International Journal of Geographical Information Science*, *31*(1), 56–80, 2017.

Kim, S., and Y. Yoon, Recommendation system for sharing economy based on multidimensional trust model, *Multimedia Tools and Applications*, *75*(23), 15,297–15,310, 2016.

Kisilevich, S., M. Krstajic, D. Keim, N. Andrienko, and G. Andrienko, Event-based analysis of people's activities and behavior using flickr and panoramio geotagged photo collections, in *2010 14th International Conference Information Visualisation*, pp. 289–296, IEEE, 2010.

Kõivumägi, E., M. Vait, A. Hadachi, G. Singer, and E. Vainikko, Real time movement labelling of mobile event data, *Journal of Location Based Services*, *9*(1), 55–76, 2015.

Kuhn, W., Ontologies in support of activities in geographical space, *International Journal of Geographical Information Science*, *15*(7), 613–631, 2001.

Kwak, H., C. Lee, H. Park, and S. Moon, What is twitter, a social network or a news media?, in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, ACM, 2010.

Lansley, G., and P. A. Longley, The geography of twitter topics in london, *Computers, Environment and Urban Systems*, *58*, 85–96, 2016.

Larson, R. R., Geographic information retrieval and spatial browsing, *Geographic information systems and libraries: patrons, maps, and spatial information [papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995]*, 1996.

Leveling, J., and S. Hartrumpf, On metonymy recognition for geographic information retrieval, *International Journal of Geographical Information Science*, *22*(3), 289–299, 2008.

Levenshtein, V. I., Binary codes capable of correcting deletions, insertions, and reversals, in *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.

Liang, H., Y. Xu, Y. Li, R. Nayak, and L.-T. Weng, Personalized recommender systems integrating social tags and item taxonomy, in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 540–547, IEEE Computer Society, 2009.

Lim, K. H., K. E. Lee, D. Kendal, L. Rashidi, E. Naghizade, S. Winter, and M. Vasardani, The grass is greener on the other side: Understanding the effects of green spaces on twitter user sentiments, in *Companion Proceedings of the The Web Conference 2018*, pp. 275–282, International World Wide Web Conferences Steering Committee, 2018.

Lynch, K., *The image of the city*, vol. 11, MIT press, 1960.

Marlow, C., M. Naaman, D. Boyd, and M. Davis, Ht06, tagging paper, taxonomy, flickr, academic article, to read, in *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pp. 31–40, ACM, 2006.

Massey, D., A global sense of place, in *Space, Place, and Gender*, pp. 146–156, University of Minnesota Press, 1994.

Massey, D., Power-geometry and a progressive sense of place, in *Mapping the futures*, pp. 75–85, Routledge, 2012.

McCallum, A. K., Mallet: A machine learning for language toolkit, 2002.

McKenzie, G., and B. Adams, Juxtaposing thematic regions derived from spatial and platial user-generated content, 2017.

McKenzie, G., K. Janowicz, S. Gao, and L. Gong, How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest, *Computers, Environment and Urban Systems*, *54*, 336–346, 2015.

Mei, Q., X. Shen, and C. Zhai, Automatic labeling of multinomial topic models, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 490–499, ACM, 2007.

Merschdorf, H., and T. Blaschke, Revisiting the role of place in geographic information science, *ISPRS International Journal of Geo-Information*, *7*(9), 364, 2018.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, Optimizing semantic coherence in topic models, in *Proceedings of the conference on empirical methods in natural language processing*, pp. 262–272, Association for Computational Linguistics, 2011.

Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl, Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries, *Spatial Cognition & Computation*, *3*(2-3), 185–204, 2003.

Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl, Where's downtown?: Behavioral methods for determining referents of vague spatial queries, in *Spatial Vagueness, Uncertainty, Granularity*, pp. 185–204, Psychology Press, 2017.

Mountain, D., and A. MacFarlane, Geographic information retrieval in a mobile environment: evaluating the needs of mobile individuals, *Journal of Information Science*, *33*(5), 515–530, 2007.

Newman, D., J. H. Lau, K. Grieser, and T. Baldwin, Automatic evaluation of topic coherence, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108, Association for Computational Linguistics, 2010.

Nguyen, D., R. Gravel, D. Trieschnigg, and T. Meder, " how old do you think i am?" a study of language and age in twitter, in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

Nielsen, J., The 90-9-1 rule for participation inequality in social media and online communities, 2006.

Norman, D. A., *The psychology of everyday things.*, Basic books, 1988.

Nussbaum, D., M. T. Omran, and J.-R. Sack, Maintaining anonymity using-privacy, *Journal of Location Based Services*, *11*(1), 1–28, 2017.

Olteanu, A., C. Castillo, F. Diaz, and E. Kiciman, Social data: Biases, methodological pitfalls, and ethical boundaries, *Frontiers in Big Data*, *2*, 13, 2019.

Openshaw, S., The modifiable areal unit problem (vol. 38), *Concepts and Techniques in Modern Geography*, p. 43, 1983.

Prelipcean, A. C., F. Schmid, and T. Shirabe, A space time alarm, in *Progress in Location-Based Services 2014*, pp. 187–198, Springer, 2015.

Purves, R., A. Edwardes, and J. Wood, Describing place through user generated content, *First Monday*, *16*(9), 2011.

Purves, R. S., P. Clough, C. B. Jones, M. H. Hall, V. Murdock, et al., Geographic information retrieval: Progress and challenges in spatial search of text, *Foundations and Trends® in Information Retrieval*, *12*(2-3), 164–318, 2018.

Purves, R. S., S. Winter, and W. Kuhn, Places in information science, *Journal of the Association for Information Science and Technology*, 2019.

Rattenbury, T., and M. Naaman, Methods for extracting place semantics from flickr tags, *ACM Transactions on the Web (TWEB)*, *3*(1), 1, 2009.

Relph, E., *Place and placelessness*, vol. 1, Pion, 1976.

Resch, B., A. Summa, P. Zeile, and M. Strube, Citizen-centric urban planning through extracting emotion information from twitter in an interdisciplinary space-time-linguistics algorithm, *Urban Planning*, *1*(2), 114–127, 2016.

Roche, S., Geographic information science ii: Less space, more places in smart cities, *Progress in Human Geography*, *40*(4), 565–573, 2016.

Rorissa, A., User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory, *Information Processing & Management*, *44*(5), 1741–1753, 2008.

Rosch, E., and B. B. Lloyd, Cognition and categorization, 1978.

Rosen, A., https://blog.twitter.com/, 2017.

Rosenkrans, G., and K. Myers, Optimizing location-based mobile advertising using predictive analytics, *Journal of Interactive Advertising*, *18*(1), 43–54, 2018.

Salvini, M. M., and S. I. Fabrikant, Spatialization of user-generated content to uncover the multirelational world city network, *Environment and Planning B: Planning and Design*, *43*(1), 228–248, 2016.

Samal, A., S. Seth, and K. Cueto, A feature-based approach to conflation of geospatial sources, *International Journal of Geographical Information Science*, *18*(5), 459–489, 2004.

Scheider, S., and K. Janowicz, Places as media of containment, in *Proceedings of the 6th International Conference on Geographic Information Science (extended abstract, forthcoming)*, 2010.

Schneider, B., The people make the place, *Personnel psychology*, *40*(3), 437–453, 1987.

Senaratne, H., A. Bröring, and T. Schreck, Assessing the credibility of vgi contributors based on metadata and reverse viewshed analysis: an experiment with geotagged flickr images, in *16th AGILE International Conference on Geographic Information Science*, 2013a.

Senaratne, H., A. Bröring, and T. Schreck, Using reverse viewshed analysis to assess the location correctness of visually generated vgi, *Transactions in GIS*, *17*(3), 369–386, 2013b.

Shatford, S., Analyzing the subject of a picture: a theoretical approach, *Cataloging & classification quarterly*, *6*(3), 39–62, 1986.

Shelton, T., A. Poorthuis, and M. Zook, Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information, *Landscape and urban planning*, *142*, 198–211, 2015.

Sigurbjörnsson, B., and R. Van Zwol, Flickr tag recommendation based on collective knowledge, in *Proceedings of the 17th international conference on World Wide Web*, pp. 327–336, ACM, 2008.

Stevens, K., P. Kegelmeyer, D. Andrzejewski, and D. Buttler, Exploring topic coherence over many models and many topics, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961, Association for Computational Linguistics, 2012.

Straumann, R. K., A. Cöltekin, and G. Andrienko, Towards (re) constructing narratives from georeferenced photographs through visual analytics, *The Cartographic Journal*, *51*(2), 152–165, 2014.

Tuan, Y.-F., *Space and place: The perspective of experience*, U of Minnesota Press, 1977.

Tuan, Y.-F. T., A study of environmental perception, *Attitudes, and Values*, 1974.

Tversky, B., and K. Hemenway, Categories of environmental scenes, *Cognitive psychology*, *15*(1), 121–149, 1983.

Twaroch, F., R. Purves, and C. Jones, Stability of qualitative spatial relations between vernacular regions mined from web data, in *Proceedings of Workshop on Geographic Information on the Internet, Toulouse, France*, 2009.

Van Mierlo, T., The 1% rule in four digital health social networks: an observational study, *Journal of medical Internet research*, *16*(2), 2014.

van Zanten, B. T., D. B. Van Berkel, R. K. Meentemeyer, J. W. Smith, K. F. Tieskens, and P. H. Verburg, Continental-scale quantification of landscape values using social media data, *Proceedings of the National Academy of Sciences*, *113*(46), 12,974–12,979, 2016.

Vasardani, M., S. Winter, and K.-F. Richter, Locating place names from place descriptions, *International Journal of Geographical Information Science*, *27*(12), 2509–2532, 2013.

Vickery, G., and S. Wunsch-Vincent, *Participative web and user-created content: Web 2.0 wikis and social networking*, Organization for Economic Cooperation and Development (OECD), 2007.

Vögele, T., C. Schlieder, and U. Visser, Intuitive modelling of place name regions for spatial information retrieval, in *International Conference on Spatial Information Theory*, pp. 239–252, Springer, 2003.

Wartmann, F. M., E. Acheson, and R. S. Purves, Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach, *International Journal of Geographical Information Science*, *32*(8), 1572–1592, 2018.

Winter, S., and C. Freksa, Approaching the notion of place by contrast, *Journal of Spatial Information Science*, *2012*(5), 31–50, 2012.

Ye, M., D. Shou, W.-C. Lee, P. Yin, and K. Janowicz, On the semantic annotation of places in location-based social networks, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 520–528, ACM, 2011.

Zheng, X., J. Han, and A. Sun, A survey of location prediction on twitter, *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1652–1671, 2018.

Zielstra, D., and H. H. Hochmair, Positional accuracy analysis of flickr and panoramio images for selected world regions, *Journal of Spatial Science*, *58*(2), 251–273, 2013.

Zook, M., M. Graham, T. Shelton, and S. Gorman, Volunteered geographic information and crowdsourcing disaster relief: a case study of the haitian earthquake, *World Medical & Health Policy*, *2*(2), 7–33, 2010.

Zou, J., and L. Schiebinger, Ai can be sexist and racist—it's time to make it fair, 2018.

Part II

PUBLICATIONS

A

PUBLICATION I: APPROACHING LOCATION-BASED SERVICES FROM A PLACE-BASED PERSPECTIVE: FROM DATA TO SERVICES?

Bahrehdar, A. R., Koblet, O., and Purves, R. S. (2019), Approaching location-based services from a place-based perspective: from data to services?. Journal of Location Based Services, 1-21.

JOURNAL OF LOCATION BASED SERVICES
https://doi.org/10.1080/17489725.2018.1564383

Taylor & Francis
Taylor & Francis Group

Approaching location-based services from a place-based perspective: from data to services?

Azam Raha Bahrehdar, Olga Koblet and Ross S. Purves

Department of Geography, University of Zurich, Zurich, Switzerland

ABSTRACT

Despite the seemingly obvious importance of a link between notions of place and the provision of context in location-based services (LBS), truly place-based LBS remain rare. Place is attractive as a concept for designing services as it focuses on ways in which people, rather than machines, represent and talk about places. We review papers which have extracted place-relevant information from a variety of sources, examining their rationales, the data sources used, the characteristics of the data under study and the ways in which place is represented. Although the data sources used are subject to a wide range of biases, we find that existing methods and data sources are capable of extracting a wide range of place-related information. We suggest categories of LBS which could profit from such information, for example, by using place-related natural language (e.g. vernacular placenames) in tracking and routing services and moving the focus from geometry to place semantics in location-based retrieval. A key future challenge will be to integrate data derived from multiple sources if we are to advance from individual case studies focusing on a single aspect of place to services which can deal with multiple aspects of place.

ARTICLE HISTORY
Received 16 July 2018
Accepted 20 December 2018

KEYWORDS
Place; user generated content; unstructured text; location-based services; context

1. Introduction

Location-based services (LBS) are, we suggest, all about place. Delivering relevant information presupposes that we understand the context of an information need, be that in the form of a need to navigate from one location to another (Kurashima et al. 2010), a desire for information about available services around a users' current and forecasted location (Poslad 2001) or interactions with a dialogue based virtual assistant (Bartie et al. 2018). Treating such context as simply spatial information, for example, as a set of coordinates, flies in the face of what we understand about how people interact with places.

Thus, for example, places have names (Coates 2006), which form an efficient shorthand for communicating about location without a need to resort to complex coordinate systems, and yet allow us to zoom in and out with

CONTACT Azam Raha Bahrehdar azam.bahrehdar@gmail.com

minimal cognitive effort (e.g. Richter et al. 2013). They have properties, in the sense of their physical materiality (Relph 1976), which in turn can reflect affordances and activities (Lansley and Longley 2016) associated with particular places at particular times (Mckenzie and Adams 2017). They are related, in that they may be contained by, overlap with or be distinct from other places (Schlieder, Vögele, and Werner 2001). Furthermore, individuals and groups may associate particular places with experiences and emotions, giving rise to the notion of sense of place (Shelton, Poorthuis, and Zook 2015). Place, in short, represents a shared meaning, and in turn, should be viewed as an indispensable form of context for LBS (c.f. Farrelly 2014).

This importance of place as a component of context is emphasised in Dey's seminal paper, where he defines context as:

> …any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves (Dey 2001).

Implicitly, according to this definition places can both *be* context in the form of information characterising an entity, but also take the role of an entity, and thus *have* context. Despite this obvious importance, attempts to deal with place in LBS, and more broadly geographic information science, are piecemeal and disconnected. They essentially fall into three camps. The first uses place as a shorthand for location and make no distinction between places and other sorts of locations. Perhaps the most obvious example is the definition of place in schema.org as 'Entities that have a somewhat fixed, physical extension[1]'. This definition reduces places to geometric objects, and while not *per se* wrong, it effectively ignores the nuances presented above and treats places as objects represented in some entity-based model of space. For example, Villegas et al. (2018) introduce the idea of location context, where a place is treated simply as a location:

> Location context: Refers to the place associated with an entity's activity (e.g. the city where a user lives). This category is sub-classified as physical (e.g. the coordinates of the user's location, a movie theater's address, or the directions to reach the movie theater from the costumer's current location), and virtual (e.g. the IP address of a computer that is located within a network) (Villegas et al. 2018).

While we do not dispute the utility of this definition, we argue that it ignores the potential richness of place as a source of context. Even where the notion of location as a social and dynamic construct is recognised (e.g. Gasparetti 2017) theories relating to place as a concept appear to be neglected despite their potential utility in better understanding and modelling context.

A second strand of work considering place concentrates on deriving general models of place, most often starting from the literature in human geography, and aiming to describe a conceptual data model suitable for

dealing with place in information systems (e.g. Jordan et al. 1998; Jorgensen and Stedman 2001; Winter and Freksa 2012). These attempts are useful and interesting, but unfortunately, they have typically stopped at the conceptual level, and thus have had limited influence on the third set of approaches.

This third group is fuelled by the opportunities offered by social media and user-generated content as data sources allowing access to a seeming 'avalanche of data'. Here, place is used as a motivation and exemplar attributes are operationalised (e.g. Hauthal and Burghardt 2013; Richter et al. 2012), though typically not further utilised in providing specific services.

Our aim in this paper is to bring together the second and third strands of research identified above and contribute to the first, focussing on place as a form of contextual information in LBS, which we suggest could benefit from considering the concept of place in more detail. We, therefore, analyse existing data-driven research to explore how authors have extracted place-related context. Based on this analysis, we identify ways in which the use of place as context could enhance specific tasks in LBS related to navigation and tracking, marketing and location-based information retrieval.

2. Exploring place in data-driven research

Since our aim was to use existing works exploring aspects of place, we performed a literature review. A major challenge in finding papers related to place is that, as we have demonstrated above, place is often simply used as a synonym for a geometric location on the one hand, and on the other not all papers dealing with place do so explicitly. Therefore, searching for literature using keywords alone is not helpful and might be misleading. To select a broad range of relative literature we used a combination of purposive and snowball sampling (Wohlin 2014).

As a starting point, we selected three papers known to us (Chesnokova, Nowak, and Purves 2017; Jenkins et al. 2016; Shelton, Poorthuis, and Zook 2015), from different research groups, including a variety of aspects of place which we wished to cover in our study. Moreover, we identified four criteria to identify further papers for our list in the next step of 'snowballing'.

- Papers must be data driven and have extracted place properties from some form of web accessible content such as Wikipedia, Twitter, Foursquare, etc. (this criterion excludes purely conceptual papers).
- Papers must capture some form of shared meaning of place. Therefore, place properties have to be generated by identifiable multiple contributors. This criterion enables us to know more about who creates descriptions, that is to say, the social aspect of place.

- Papers aim to derive properties for places, rather than attributing existing point of interest data.
- Finally, we were only interested in papers where place properties varied in space, since otherwise, such information is not useful contextual information for LBS.

Only articles which met all of these criteria were retained, and we did not aim to find an exhaustive, but rather a representative set of papers. Representativity in our study implied diversity in the set of four aspects described below, which we used to analyse our papers. It is thus important to make clear that the process of paper selection was iterative, and necessarily subjective. Thus, for example, the seed set of papers we chose included a paper from our research group, and the papers cited by these works unsurprisingly reflect a particular research network (Skupin 2014).

After carrying out our snowballing process, all selected articles were characterised according to the following four aspects: research rationale, sources of place data, data characteristics, and place dimensions.

To understand the **rationale** behind each article, we looked at the *application domain* (if applicable) targeted by the study and the *motivation* given for exploring place descriptions. In exploring **data sources**, we not only listed data sources, but also analysed the ways in which data were retrieved. The third aspect we studied concerned the **characteristics** of the data collected. For example, the *study area* and the *time span* associated with the dataset, what techniques, if any were used to *localise* data with respect to places and, finally, in what *language* data were created. The final aspect we explored related to the **place dimensions** accounted for in a paper. We compared these with a model from information science (Shatford 1986), and classified papers according to the *where* facet of the Panofsky–Shatford facet matrix. Thus, we categorised papers as addressing one or more of the following dimensions: the *specific of* (related to named places or instances of places), *generic of* (properties or features of places), and the *about* (associated emotions and feelings).

3. Findings

In total, we selected 18 articles for further examination (Table 1). All of the articles were published between 2010 and 2018, reflecting both our initial set of seed articles, and the recent and increasing popularity of such data-driven research.

3.1. Rationale

Of the 18 papers we identified, only two (Huang 2016; Ye et al. 2011) made direct proposals for applications in LBS. The next group of papers made general claims about applications in landscape studies, for example, with

Table 1. The list of selected papers about place and their year of publication.

Year of publication	Selected articles
2010	(Hollenstein and Purves 2010)
2011	(Ye et al. 2011)
2013	(Adams and McKenzie 2013; Hauthal and Burghardt 2013)
2015	(Dunkel 2015; Hobel, Fogliaroni, and Frank 2016; Shelton, Poorthuis, and Zook 2015)
2016	(Capineri 2016; Derungs and Purves 2016; Gliozzo, Pettorelli, and Haklay 2016; Huang 2016; Jenkins et al. 2016; Resch et al. 2016)
2017	(Chesnokova, Nowak, and Purves 2017; Gao et al. 2017; Mckenzie and Adams 2017)
2018	(Chen, Parkins, and Sherren 2018; Lim et al. 2018)

respect to aesthetics, cultural ecosystem services and more generally the perceived environment (Dunkel 2015; Gliozzo, Pettorelli, and Haklay 2016; Derungs and Purves 2016; Chen, Parkins, and Sherren 2018; Chesnokova, Nowak, and Purves 2017). Interestingly, these papers focused mostly on non-urban environments, while a further group was motivated by exploring properties of cities, with reference to both inequality and the need for more nuanced ways of analysing such data (Shelton, Poorthuis, and Zook 2015; Capineri 2016). All of these papers made claims about application domains focussing on understanding specific places and their properties. A related group of papers also focussed on urban areas, but zoomed into the emotions experienced and reported in such places by individuals (Hauthal and Burghardt 2013; Resch et al. 2016; Lim et al. 2018). In contrast to the earlier works, Lim et al. (2018) quantitavely compare the difference between sentiment associated with green and other urban spaces. The potential of user-generated content as a way of finding out more about how places are named, in the sense of the (vague) footprints associated with vernacular usage of placenames is explored by three papers (Hollenstein and Purves 2010; Hobel, Fogliaroni, and Frank 2016; Gao et al. 2017). A final group of three papers essentially focussed on exploring and deriving thematic regions, associated with or forming places, and have a clear methodological rather than application focus (Adams and McKenzie 2013; Jenkins et al. 2016; Mckenzie and Adams 2017).

A number of points are worth making here. Firstly, and contrary to our initial expectations, we found that papers dealing with place focus on both urban and rural landscapes, and thus that contextual information in both settings appears to be available. Secondly, many of the papers made strong arguments as to the availability of new data sources and their potential for allowing contextual information related to subjective experiences of places, be that in the context of their naming, properties or emotions related to them. Thus, a key motivation for such research is clearly pragmatic and data driven. Thirdly, and importantly, direct applications in LBS reflecting more complex conceptualisations of place, or indeed even arguing for its importance, were rare in our sample despite, we would argue, their obvious importance.

3.2. Data sources

Data-driven research requires data – and the choice of data may have implications for the conclusions which can be drawn. In terms of place for LBS, it is important to understand, for example, not only which places are represented, but also by whom and when. Just as important as what we can derive from such data are the gaps – the things which are not said, but which might be equally important in capturing aspects of place. For example, which communities produced the data (and thus who did not participate), which places are mapped (and thus which are ignored), and what objects or emotions are more commonly shared.

We identified four broad categories of data which were used in the papers we explored. The first, images and their associated metadata, have been argued to potentially provide an immediate and direct link to place (Fisher and Unwin 2005). We identified three sources of such data in the papers we analysed: Flickr, which was most common (Capineri 2016; Dunkel 2015; Gao et al. 2017; Gliozzo, Pettorelli, and Haklay 2016; Hauthal and Burghardt 2013; Hollenstein and Purves 2010; Huang 2016), Instagram (Chen, Parkins, and Sherren 2018; Gao et al. 2017; Mckenzie and Adams 2017) and the now defunct Panoramio (Gliozzo, Pettorelli, and Haklay 2016; Hauthal and Burghardt 2013). We note that the popularity of Flickr might be attributable to the relatively straightforward access to data, with all non-private images and their metadata being accessible through the Flickr API, and both spatial (e.g. using a bounding box) and textual (e.g. using a term like *Downtown)* queries being straightforward to implement. By contrast Instagram's API is no longer easily accessible and the terms of use of the data are more complex. However, it has been argued that the Instragram community is broader than that of Flickr, potentially providing access to a wider range of place descriptions (Di Minin, Tenkanen, and Toivonen 2015; Gao et al. 2017). All three sources focus on the use of tags as a way of both indexing content (Mountain and MacFarlane 2007) and improving searchability (and thus visibility). Importantly, images taken, uploaded and tagged on Flickr and other image sharing platforms are not randomly sampled – they represent popular places (Crandall et al. 2009), are often part of a narrative (Davies 2007) and have been argued to be indicators of aesthetics and recreation in a landscape context (Van Zanten et al. 2016). One source of ambiguity concerns the locations associated with image metadata. Typically, these are the location of the photographer, though users may also associate images directly with the location of content. A fourth source of image data was the Geograph platform, used by Chesnokova, Nowak, and Purves (2017) and Gliozzo, Pettorelli, and Haklay (2016). Here, images are related to 1-km grid squares and associated with textual descriptions, which can then be analysed.

The second broad group of data used are microblogs, exclusively in the form of Twitter data (Capineri 2016; Gao et al. 2017; Jenkins et al. 2016; Mckenzie and Adams 2017; Resch et al. 2016; Shelton, Poorthuis, and Zook 2015; Lim et al. 2018). Twitter's popularity, like that of Flickr, is mostly ascribable to its ease of access through an API, though in contrast to Flickr data, historical data are difficult to obtain. Indeed, most researchers only have access to some small proportion of the total volume of Tweets. Twitter messages are short, often with a relatively simple language structure (Dittrich, Richter, and Lucas 2015), covering a wide variety of topics without a focus on specific domain (Go, Bhayani, and Huang 2009; Kwak et al. 2010) and are a popular source for research, despite a wide range of challenges including a high frequency of misspelling and slang (Go, Bhayani, and Huang 2009), the use and mixing of multiple languages (Hong, Convertino, and Chi 2011), the prevalence (especially, it appears, in geocoded Tweets) of bots (Chu et al. 2010; Compton, Jurgens, and Allen 2014) and the fundamental question of whether or not location is strongly correlated with the topic of discussion in a Tweet (Hahmann, Purves, and Burghardt 2014). In terms of LBS this is of crucial importance, since, unlike images, the location of a Tweet is associated with where something was said, rather than the location of the object being described.

The third category of data we identified were reviews and check-ins, for example, in the form of Foursquare, Yelp and the now-defunct Whrrl (Ye et al. 2011; Mckenzie and Adams 2017). Interestingly these data were used not only to describe places, but also as a source of place geometry, where the places were essentially the points of interest stored by the services. It is worth noting that these services are typically already available as LBS.

The fourth, and final category of data were unstructured texts, for example, in the form of travel blogs, TripAdvisor entries, Wikipedia pages and the Text +Berg corpus (i.e. Adams and McKenzie 2013; Hobel, Fogliaroni, and Frank 2016; Gao et al. 2017; Derungs and Purves 2016). In unstructured text, more complex methods are required to both relate content to specific places and to extract information related to place. Importantly many of the sources chosen are already associated with places explicitly, for example, in TripAdvisor entries and Wikipedia pages where content associated with specific locations is extracted. An important issue with such texts relates to their availability and copyright associated with them. While Wikipedia texts are freely available under an open licence, this is not the case for TripAdvisor, where content is copyrighted and only available under specific terms.

A number of comments can be made about the data sources chosen in our papers. Firstly, we once again note a strong dose of pragmatism in the choice of data sources – researchers often chose data which were relatively easily available, and where access was free. Secondly, the nature of the data used is heterogeneous, ranging from content with a more or less immediate link to

place (in the form of images and their metadata and reviews) to much less direct links (in the form of Tweets and some unstructured text, for example, articles in the Text+Berg corpus describing Alpine plants or animals). Furthermore, the range of granularities captured in such data, and thus the scales of the places described is not constant, with in particular unstructured text and microblogs capable of capturing information across a very wide range of scales, with important implications for the nature of the context which can be extracted. In the next section, we, therefore, explore the approaches taken to extracting and analysing data such that place could be characterised.

3.3. Data characteristics

Exploring study areas and the time spans over which data were collected gives us some insight into both the potential, and also the limitations, of the approaches taken especially with respect to their use in LBS. All but one study (Adams and McKenzie 2013), chose to limit their study area to specific places at a variety of scales. Furthermore, studies took two essential approaches to linking datasets to places. Derungs and Purves (2016) and Chesnokova, Nowak, and Purves (2017) both used complete corpora covering Switzerland and Great Britain, respectively, and mapped these corpora onto a continuous, field-based, model of place within these countries. All of the other studies we explored either used some form of bounding box (e.g. Resch et al. 2016; Huang 2016) or keywords to identify data associated with specific locations (e.g. Capineri 2016). The data thus collected can be thought of as being related to entities, either in the form of a geometry or a named place. Importantly though, these entities are not necessarily treated as having properties which are constant (e.g. Gao et al. 2017), and nor were they always handled as having sharp boundaries (e.g. Hollenstein and Purves 2010). At their simplest the entities with which properties were associated were represented as points related to points of interest (Ye et al. 2011, Mckenzie and Adams 2017), while more complex entities represented linear features (i.e. Kilburn High Road or the High Line (Capineri 2016; Dunkel 2015)) or areal features (e.g. Gao et al. 2017). Although the papers we explored generally did not discuss in detail issues of inequality in data production (Graham, Hale, and Stephens 2012), we believe that these issues make global modelling of place challenging and highly subject to bias. Thus, the often implicitly taken decision to concentrate on individual cities or countries, appears to make sense when using individual data sources.

The temporal variation in data used to characterise places varied widely, from a minimum of one day to one week (Resch et al. 2016) to a maximum of 152 years (Derungs and Purves 2016). However, we observe that in general authors appear to have chosen time scales either based around meaningful events (Resch et al. 2016), an implicit requirement to collect sufficient data to

write a paper (e.g. Shelton, Poorthuis, and Zook 2015) or simply by analysing the complete corpus available (e.g. Derungs and Purves 2016). We think all three of these positions are justifiable, but note that the sampling period will influence the nature of the context which can be analysed and used in downstream LBS, since short-sampling periods cannot capture cyclical events, while long sampling periods may capture variation which represents, for example, change in language use over time rather than changes occurring to places (Nguyen et al. 2013).

Another key question with respect to data characteristics concerns the way in which the data themselves are localised, and how this localisation is then linked to places. The majority of data in the selected papers had explicit coordinates, though as discussed above, these coordinates may be associated with places of differing granularities. This issue is actively exploited in work concerned with vernacular places and vague cognitive regions (e.g. Gao et al. 2017). We would argue that even where coordinates are stored as metadata, more consideration should be given to the ways in which these points are then linked to places, and indeed to the challenges of matching datasets collected in such ways.

These issues become more apparent when working with data where location is conveyed indirectly through a placename. In the studies we explored, methods were used to both identify placenames and link these explicitly to locations (e.g. Adams and McKenzie 2013; Derungs and Purves 2016). Both of these papers chose to link the coordinates assigned to placenames to relatively coarse grids, thus explicitly representing some form of uncertainty in the granularity of descriptions of places. However, such coarse grids, though at least addressing the issue of granularity explicitly, will typically represent place as unchanging context for large distances with respect to LBS.

The third characteristic we explored was language. Even though the papers we analysed focused almost exclusively on textual content, only eight articles specified the analysed language(s). Of these, two processed German as well as English (Hauthal and Burghardt 2013; Hollenstein and Purves 2010), and one analysed text only in German (Derungs and Purves 2016). In general, by exploring the results presented, it was clear that English was favoured, even in locations where it is not an everyday language. This dominance of English in the papers we analysed, which despite its popularity is clearly not representative of the population as a whole, has several implications. Firstly, there is a tendency to conduct studies in English-speaking countries, and to subsume place into context related to English-speaking cultures. Secondly, the dominance (and good performance) of natural language processing methods in English may result in an unrealistic homogenisation of place-related concepts, where in reality much more diversity may actually be present. Thirdly, by using English in places where the language is not spoken, unrepresentative sources may be favoured (e.g. those used by tourists) creating a further negative feedback in terms of the representation of such places (c.f. Graham et al. 2014).

4. Place dimensions

We chose to use Shatford's (1986) model, since in previous work (e.g. Edwardes and Purves 2007) this has proved a reliable and powerful way to explore different aspects of spatial descriptions. In the following, we focus not only on exploring where papers belong in this model, but also the ways in which individual facets are represented.

We relate the *specific of* to concrete ways in which places are named, that is to say, the use of placenames, be they related to administrative or vernacular usages. Generic terms such as *downtown* or the *city centre* become specific when they refer to a particular place. The most important papers relating to the *specific of* are thus those motivated by place names (Hollenstein and Purves 2010; Hobel, Fogliaroni, and Frank 2016; Gao et al. 2017). In these papers, a number of contrasting aspects are explored. All three look at delineating regions associated with specific place names, whether through the use of density surfaces (Hollenstein and Purves 2010), machine learning (Hobel, Fogliaroni, and Frank 2016) or clustering and polygon approximation (Gao et al. 2017). Hollenstein and Purves (2010) also explored the use of specific terms in the contiguous USA, thus identifying places more likely to be referred to as Downtown at a range of scales. Gao et al. (2017) provide a bridge to the next facet in Shatford's classification, the *generic of*, by generating thematic characteristics related to the cognitive regions SoCal and NorCal (Southern California and Northern California) using Latent Dirichlet Allocation. The resulting topics are dominated by generic terms such as desert, beach, mountain and road which give some insight into the properties of these regions. This representation of the *generic of* as a bag of words, often associated with a rank is typical of many of the approaches we looked at (c.f. Adams and McKenzie 2013). Thus, Capineri (2016) mapped terms onto a similar place model (Agnew 1987) and counted terms representing different activities and objects. Dunkel (2015) also related tag frequencies to particular places, though he did not discriminate between placenames and other classes. Derungs and Purves (2016) used a filtered list of nouns, which they claim captures landscape variation in German to capture generic properties of landscape, and they show how locations can be compared using vectors of terms representing individual grid cells.

Shatford, in her characterisation of the *about* facet, describes it as a way of symbolising a place through a locale, or communicating abstract thoughts (e.g. paradise) through a place. Despite our initial expectations, we found that in many instances the *about* facet was the most appropriate home for studies we explored. Thus, though Gliozzo, Pettorelli, and Haklay (2016) count pictures and users, they do so to represent the abstract notion of cultural ecosystem services, while Chesnokova, Nowak, and Purves (2017) use image ratings to model landscape preference, which again, we argue can be considered to relate to an abstract concept (beauty) of places.

Indeed, many of the papers we explored sought to both map and interpret, typically through the use of word clouds or other relatively simple techniques, the semantics associated with preferences for particular places (e.g. Adams and McKenzie 2013; Dunkel 2015). Other approaches which clearly are linked to this more abstract notion of place are those which seek to link emotions about urban locations to time or day of the week or season of the year (Hauthal and Burghardt 2013; Resch et al. 2016). Lim et al. (2018) seek to characterise both the *generic of* (in the form of green areas in an urban setting) and their properties with respect to the *about* facet through sentiment analysis. They move beyond simple quantification of negative and positive sentiment to also explore the nature of emotions (e.g. anger or joy) associated with different urban settings, showing that particularly negative sentiments are often associated with transport infrastructure and explore how these sentiments change over time.

Perhaps the most abstract example is the work of Shelton, Poorthuis, and Zook (2015). They argue for understanding places in terms of the ways they are experienced and moved through, and the importance of relating data points to one another. Their analysis though, is typical of many of the papers we explored, where the *about* facet can only be understood through a high level of interpretation and contextual knowledge brought to the data by the authors. This difficulty is expressed succinctly by Capineri (2016):

> …feelings and emotions are not always expressed by single words like happy, unhappy, love or hate but rather with expressions of more than one word that reveal the state of mind. …only a limited number of records contain emotional expressions which can be linked to the categories.

A few points are worthy of note here. Firstly, real data capture and can represent all three facets of where, as modelled by Shatford. Using this model it is possible to show how places can be delineated and assigned membership values in terms of their names, and how they can be compared and represented as thematic regions. Furthermore, even using simple counts, it is possible to make effective links to more abstract concepts such as aesthetics. However, an important note of caution should also be sounded. We observed that in particular for the more abstract shared notions, which might be best mapped onto sense of place, a great deal of subjective interpretation was performed. Calls to, for example, use data capturing perceived safety in routing (such as emotions derived from social media), may reinforce or even generate inequalities in our understanding and use of place (c.f. Andreas and Mazimpaka 2016). This adds weight to Shelton et al.'s caution to not simply analyse social media, but rather 'construct empirically grounded counter-narratives of these inequalities' (Shelton, Poorthuis, and Zook 2015, 210).

5. Implications and discussion: opportunities and challenges for the place-based modelling in LBS

Having explored the ways in which place information has been extracted from a range of data sources, and analysed some key properties thereof, we now return to the use of the extracted information in the context of LBS. We used the list of application categories for LBS proposed by Basiri et al. (2015) to provide a skeleton for this discussion. Basiri et al. define their categories based on the spatial and temporally related positional requirements for the LBS itself – for example, the need for navigational systems to be precise. However, our focus is on how place-related information could be used to enhance such services, either taking the form of context associated with a place, or being context in and of themselves. We do not claim to be comprehensive, but rather select examples from three domains: navigation and tracking; marketing and location-based information retrieval where we see the most potential use for place-related information. In the following, we present what we see as some key opportunities, and discuss some of the potential challenges and limitations in the use of place-related information in LBS.

LBS used in navigation and route finding typically relies on highly precise, complete data and accurate real-time location to provide context and feedback to the user. However, specifying a route requires that a user input a target destination (assuming that they are travelling from their current location). Specifying target locations in terms of coordinates or exact addresses is in many cases challenging because these are not natural ways for humans to communicate about locations. Incorporating representations of place related to the *specific of*, that is to say placenames commonly used in a particular area, would be one potential way of improving and facilitating such interaction. Using hierarchies of such places, based on UGC, would provide a mechanism for zooming in and out (Richter et al. 2013), and adjusting the requirements of a route to the needs of a user. For example, a requirement to take me Downtown could be met by a general direction along main thoroughfares with a resultantly low cognitive load, rather than complex directions navigating individual streets to arrive at a particular address Downtown. Representing such initial destinations as a generalised geometry, for example, in the form of a bounding rectangle or an alpha shape (Twaroch, Purves, and Jones 2009; Keßler, Krzysztof, and Mohamed 2009), is one approach to dealing with the vagueness inherent in such regions.

In terms of tracking, such information representations could provide a more meaningful way to aggregate user information than purely geometric regions, providing a bottom-up model of the places visited by groups of users (c.f. Huang 2016). In both cases, there is a need for such information to be stored in more amenable data structures, such as the place graphs originally proposed

by Vasardani et al. (2013) and used by Kim, Vasardani, and Winter (2017). These are designed to be closely related to the ways that humans (rather than machines) reason qualitatively about places. They are well suited to both capturing some notion of vagueness, and hierarchy, with a key challenge then lying in mapping such data back onto the more precise geometry and network used in typical routing systems.

Information classified as *generic of* can provide important additional context in both navigation and tracking. In the latter, it may help to annotate user behaviours before these are analysed, for example, by identifying all users who visited similar locations (c.f. Adams and McKenzie 2013; Derungs and Purves 2016) characterised not simply as a place-type associated with a point, but, for example, as a vector of terms associated with regions, for which similarity measures can then be calculated (Janowicz, Raubal, and Kuhn 2011).

In routing, arguments have already been made for using such information in, for example, modelling more pleasant routes (as represented by tourist flows or semantics attributed to pictures) (Prelipcean, Schmid, and Shirabe 2015; Alivand and Hochmair 2013). However, incorporating such forms of context in LBS requires that we also think about the potentially deleterious effects of such algorithmic solutions. Beauty (and other abstract notions) are inherently human constructs and as such are biased by the communities creating the data. Thus, they may, even through seemingly innocuous applications, reinforce prejudices by, for example, generating routes which avoid certain parts of a city (c.f. Shelton, Poorthuis, and Zook 2015).

The use of LBS in navigation, where users actively seek information, and tracking, where user position is analysed with respect to context, naturally leads to our next major domain area, the use of LBS in marketing products and services based on current, past and predicted location and associated context. Thus, for example, by using *generic of* place information to describe activities from previous and current visitors, it is possible to generate movement profiles which suggest likely activities (and thus can trigger location-based advertising) (Kõivumägi et al. 2015). Starting from a place-based model has several potential advantages. Firstly, as we have seen, such models need not be linked with individual POIs, but can rather take the form of continuous grids (protecting privacy by allowing obfuscation of position (Nussbaum, Omran, and Sack 2017)). Secondly, such models could potentially allow for geofencing approaches (Rosenkrans and Myers 2018) to the triggering of such adverts based on meaningful places, rather than administrative boundaries which may have little to do with the ways in which places are experienced. Since many of the papers which we analysed not only attributed places, but also identified them, such approaches can also be seen as powerful ways of generating context for recommendation systems in marketing and more general retrieval contexts (Huang 2016; Ye et al. 2011), our third potential application area. Here we see essentially three key advantages. Firstly, as in navigation and tracking,

place-based models allow us to move away from precise geometric information, and to generate query footprints for information related to places as actually experienced. Since different data sources capture information about different user groups (Gao et al. 2017), and since user groups can be filtered based on behavioural patterns (Huang 2016), then it is also possible to generate query footprints appropriate to different groups (e.g. a representation of the city centre which is appropriate from the perspective of a tourist visiting a location, as opposed to a local). Secondly, by building place-based hierarchies it should be possible to make proximity queries which are not purely based on distance buffers, and rather use more natural topological representations of locations (e.g. the notion that a place is contained or adjacent to another). Such hierarchies need not only take account of place geometries, but also place semantics, as proposed by Gao et al. (2013), who demonstrated the use of *platial-buffering* based on semantic relations between places derived from linked data. Such approaches allow us to focus more on the semantics of place, and reduce the importance of geometric representations. Having built an appropriate hierarchy places can be queried for other contained or overlapping places. Thirdly, place-based models can potentially allow for indexing of documents taking into account both properties related to the *generic of* and *about sensu* Shatford. By doing so, it should be possible to move towards LBS for specific contexts such as tourism which return, for example, information about castle like locations which are considered haunting, or beautiful beaches in a location-based context.

Despite the obvious and demonstrable potential of using information related to place in LBS, there are a number of important limitations in going down this road. The first, and most important, is that data-driven approaches, as is increasingly being recognised, will reflect the inaccuracies, biases and gaps present in the data used (Graham, Hale, and Stephens 2012; McKenzie et al. 2015; Zook et al. 2010). This means that any services developed in such ways must, from the beginning, clearly state what limitations arise from the data used. However, these limitations are not specific to LBS developed taking a place-based perspective. Rather, since studies of place are often inherently critical, then these issues are more likely to emerge (c.f. Shelton, Poorthuis, and Zook 2015). The second major limitation also concerns data availability. Different services are more or less popular with different user groups in different locations (Van Zanten et al. 2016) leading to no *one size fits all* solution to modelling any aspect of place. This is reflected through the lack of attempts at modelling place-based properties globally, and explains why so many of our papers focus on specific examples. A third challenge, and possible route towards solving such problems lies in the development of approaches to link data from different sources to fill such gaps. Currently, most authors use either single data sources, or compare data sources, but direct linkages and integration of such heterogenous data are rare and difficult.

6. Conclusions

In 2001, Dey emphasised the importance of place as potential context, and in 2014 Farrelly argued for the irreplacibility of place information in LBS. Despite these prescient statements, we argued in the introduction that place is still largely simplified or neglected in LBS. By performing a targeted literature review we wished to explore what sorts of place information can be extracted from available data, and also suggest some opportunities and challenges for using such information in LBS. Our study is limited to the set of papers we chose, which were purposively sampled to cover a particular set of criteria. However, we believe that the analysis of these papers illustrates some of the opportunities and challenges for the use of place-based information in LBS.

The first key opportunity arises from the volume of work which has already been done. By using Shatford's model we were able to identify papers which explored both *specific of* and *generic of* aspects of where – in other words, which looked not only at how places were named, but also their properties. We were surprised to find so many papers also exploring more abstract notions, related to the *about*. A number of authors explored detailed notions such as aesthetics (Dunkel 2015; Chesnokova, Nowak, and Purves 2017) or segregation (Shelton et al.) typically by choosing one aspect and then carrying out detailed interpretation of ways in which this aspect was captured in user-generated content. Often the semantics related to the content and its location was interpreted by the authors in useful and thought-provoking ways. In terms of generating LBS, this means that methods are available to derive complex place properties, but that these have typically to date been only applied to answer individual questions, rather than characterise places more generally. Our results suggest that a plethora of place properties can already be modelled, and that by exploring existing work much richer and more multi-dimensional place context could be created.

The second opportunity lies in the nature of the data used in this work, and the specific needs of LBS. Systems for use by humans should communicate with humans in ways which reflect human spatial cognition, rather than data models imposed by computers. Our analysis showed clearly that natural language was often analysed, and available, to characterise places. This, in turn, provides a host of opportunities for developing systems which put language, rather than geometry, at the forefront in not only querying, but also presenting information to users. Furthermore, our analysis suggests a number of ways in which data models might be improved beyond simple point-based representations (for example, by using topology, linking place properties to continuous fields, or building place graphs) which could, in turn, allow more imaginative services to be developed.

In parallel with these opportunities, a number of dangers and challenges arise in using place-based information in LBS. The greatest of these lies in the

potential impacts of biased data and algorithms developed to take advantage of such data. Although these concerns are not unique to LBS (Boyd and Crawford 2012), and nor do they only arise when we used place-based methods, we think they are especially important in this context. Paying attention to place as a concept has a long history in human geography and is concerned with better understanding shared and plural ways of thinking about place. Methods which use place should remember these critical beginnings, and ensure that they do not replicate, or even reinforce inequalities.

The final challenge we see for the development of LBS using place-based concepts arises from the nature of the data and studies which we analysed. It is clear that no single dataset, nor a single method, will allow us to characterise place everywhere. Developing generalisable services however requires that the community address the considerable challenge of integrating place information with widely varying semantics and spatial and temporal granularities. Only by approaching this challenge in a systematic way will it be possible to start to put together the pieces of the jigsaw, and develop place-based LBS which better address real-world needs.

Note

1. http://schema.org/Place.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung [200021_149823].

ORCID

Azam Raha Bahrehdar http://orcid.org/0000-0003-1392-474X
Olga Koblet http://orcid.org/0000-0002-4298-1789

References

Adams, B., and G. McKenzie. 2013. "Inferring Thematic Places from Spatially Referenced Natural Language Descriptions." In *Crowdsourcing Geographic Knowledge*, edited by D. Sui, S. Elwood, and M. Goodchild, 201–221. Dordrecht: Springer. doi:10.1007/978-94-007-4587-2_12.

Agnew, J. A. 1987. *Place and Politics: The Geographical Mediation of State and Society*. Boston: Allen & Unwin.

Alivand, M., and H. Hochmair. 2013. "Extracting Scenic Routes from VGI Data Sources." In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '13*, 23–30. New York: ACM Press. doi:10.1145/2534732.2534743.

Andreas, K., and J. D. Mazimpaka. 2016. "Safety-Aware Routing for Motorised Tourists Based on Open Data and VGI." *Journal of Location Based Services* 10 (1): 64–77. doi:10.1080/17489725.2016.1170216.

Bartie, P., W. Mackaness, O. Lemon, T. Dalmas, S. Janarthanam, R. L. Hill, A. Dickinson, and X. Liu. 2018. "A Dialogue Based Mobile Virtual Assistant for Tourists: The SpaceBook Project." *Computers, Environment and Urban Systems* 67 (2018): 110–123. doi: 10.1016/j.compenvurbsys.2017.09.010.

Basiri, A., T. Moore, C. Hill, and P. Bhatia. 2015. "Challenges of Location-Based Services Market Analysis: Current Market Description." In *Progress in Location-Based Services 2014*, edited by G. Gartner and H. Huang, 273–282. Cham: Springer. doi:10.1007/978-3-319-11879-6_19.

Boyd, D., and K. Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication and Society* 15 (5): 662–679. doi:10.1080/1369118X.2012.678878.

Capineri, C. 2016. "Kilburn High Road Revisited." *Urban Planning* 1 (2): 128–140. doi:10.17645/up.v1i2.614.

Chen, Y., J. R. Parkins, and K. Sherren. 2018. "Using Geo-Tagged Instagram Posts to Reveal Landscape Values around Current and Proposed Hydroelectric Dams and Their Reservoirs." *Landscape and Urban Planning* 170 (2018): 283–292. doi: 10.1016/j.landurbplan.2017.07.004.

Chesnokova, O., M. Nowak, and R. S. Purves. 2017. "A Crowdsourced Model of Landscape Preference." *LIPIcs-Leibniz International Proceedings in Informatics* 86 (19): 1–13. doi:10.4230/LIPIcs.COSIT.2017.19.

Chu, Z., S. Gianvecchio, H. Wang, and S. Jajodia. 2010. "Who Is Tweeting on Twitter." In *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC '10*, 21–30. New York: ACM Press. doi:10.1145/1920261.1920265.

Coates, R. A. 2006. "Properhood." *Language* 82 (2): 356–382. doi:10.1353/lan.2006.0084.

Compton, R., D. Jurgens, and D. Allen. 2014. "Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization." In *2014 IEEE International Conference on Big Data (Big Data)*, 393–401. IEEE. doi:10.1109/BigData.2014.7004256.

Crandall, D. J., L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. "Mapping the World's Photos." In *Proceedings of the 18th International Conference on World Wide Web - WWW '09*, 761–771. New York: ACM Press. doi:10.1145/1526709.1526812.

Davies, J. 2007. "Display, Identity and the Everyday: Self-Presentation through Online Image Sharing." *Discourse: Studies in the Cultural Politics of Education* 28 (4): 549–564. doi:10.1080/01596300701625305.

Derungs, C., and R. S. Purves. 2016. "Characterising Landscape Variation through Spatial Folksonomies." *Applied Geography* 75 (2018): 60–70. doi: 10.1016/j.apgeog.2016.08.005.

Dey, A. K. 2001. "Understanding and Using Context." *Personal and Ubiquitous Computing* 5 (1): 4–7. doi:10.1007/s007790170019.

Di Minin, E., H. Tenkanen, and T. Toivonen. 2015. "Prospects and Challenges for Social Media Data in Conservation Science." *Frontiers in Environmental Science* 3: 1–6. doi:10.3389/fenvs.2015.00063.

Dittrich, A., D. Richter, and C. Lucas. 2015. "Analysing the Usage of Spatial Prepositions in Short Messages." In *Progress in Location-Based Services 2014*, edited by G. Gartner and H. Huang, 153–169. Cham: Springer. doi:10.1007/978-3-319-11879-6_11.

Dunkel, A. 2015. "Visualizing the Perceived Environment Using Crowdsourced Photo Geodata." *Landscape and Urban Planning* 142: 173–186. doi:10.1016/j.landurbplan.2015.02.022.

Edwardes, A. J., and R. S. Purves. 2007. "A Theoretical Grounding for Semantic Descriptions of Place." In *International Symposium on Web and Wireless Geographical Information Systems*, 106–120. Berlin: Springer.

Farrelly, G. 2014. "Irreplaceable: The Role of Place Information in a Location Based Service." *Journal of Location Based Services* 8 (2): 123–132. doi:10.1080/17489725.2013.879217.

Fisher, P., and D. Unwin. 2005. "Re-presenting geographical information systems." In *Re–presenting GIS*, edited by Peter Fisher and David Unwin, 1-17. London: John Wiley & Sons.

Gao, S., K. Janowicz, G. McKenzie, and L. Li. 2013. "Towards Platial Joins and Buffers in Place-Based GIS." *Comp@ Sigspatial* (2013): 42–49.

Gao, S., K. Janowicz, D. R. Montello, Y. Hu, J. Yang, G. McKenzie, Y. Ju, L. Gong, B. Adams, and B. Yan. 2017. "A Data-Synthesis-Driven Method for Detecting and Extracting Vague Cognitive Regions." *International Journal of Geographical Information Science* 31 (6): 1–27. doi:10.1080/13658816.2016.1273357.

Gasparetti, F. 2017. "Personalization and Context-Awareness in Social Local Search: State-Of-The-Art and Future Research Challenges." *Pervasive and Mobile Computing* 38 (2): 446–473. doi:10.1016/j.pmcj.2016.04.004.

Gliozzo, G., N. Pettorelli, and M. Haklay. 2016. "Using Crowdsourced Imagery to Detect Cultural Ecosystem Services: A Case Study in South Wales, UK." *Ecology and Society* 21: 3. doi:10.5751/ES-08436-210306.

Go, A., R. Bhayani, and L. Huang. 2009. "Twitter Sentiment Classification Using Distant Supervision." *Processing* 150: 1–6. doi:10.1016/j.sedgeo.2006.07.004.

Graham, M., S. Hale, and M. Stephens. 2012. "Featured Graphic: Digital Divide: The Geography of Internet Access." *Environment and Planning A* 44 (5): 1009–1010. doi:10.1068/a44497.

Graham, M., B. Hogan, R. K. Straumann, and A. Medhat. 2014. "Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty Uneven Geographies of Knowledge." *Annals of the Association of American Geographers* 104 (4): 746–764. doi:10.1080/00045608.2014.910087.

Hahmann, S., R. S. Purves, and D. Burghardt. 2014. "Twitter Location (Sometimes) Matters: Exploring the Relationship between Georeferenced Tweet Content and Nearby Feature Classes." *Journal of Spatial Information Science* 9 (9): 1–36.

Hauthal, E., and D. Burghardt. 2013. "Detection, Analysis and Visualisation of Georeferenced Emotions." In *Proceedings of the 26th International Cartographic Conference* (ICC 2013), edited by M. F. Buchroithner, 47. Dresden: International Cartographic Association.

Hobel, H., P. Fogliaroni, and A. U. Frank. 2016. "Deriving the Geographic Footprint of Cognitive Regions." In *Lecture Notes in Geoinformation and Cartography: Geospatial Data in a Changing World*, edited by T. Sarjakoski, M. Santos, and L. Sarjakoski, 67–84. Cham: Springer. doi:10.1007/978-3-319-33783-8_5.

Hollenstein, L., and R. S. Purves. 2010. "Exploring Place through User-Generated Content: Using Flickr to Describe City Cores." *Journal of Spatial Information Science* 1 (1): 21–48. doi:10.5311/JOSIS.2010.1.3.

Hong, L., G. Convertino, and E. H. Chi. 2011. "Language Matters in Twitter : A Large Scale Study." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 518–521. Barcelona: AAAI Press.

Huang, H. 2016. "Context-Aware Location Recommendation Using Geotagged Photos in Social Media." *ISPRS International Journal of Geo-Information* 5 (11): 195. doi:10.3390/ijgi5110195.

Janowicz, K., M. Raubal, and W. Kuhn. 2011. "The Semantics of Similarity in Geographic Information Retrieval." *Journal of Spatial Information Science* 2: 29–57. doi:10.5311/JOSIS.2011.2.3.

Jenkins, A., A. Croitoru, A. T. Crooks, and A. Stefanidis. 2016. "Crowdsourcing a Collective Sense of Place." *PloS One* 11 (4): 1–20. Edited by T. Preis. doi: 10.1371/journal.pone.0152932.

Jordan, T., M. Raubal, B. Gartrell, and M. Egenhofer. 1998. "An Affordance-Based Model of Place in GIS." *8th International Symposium on Spatial Data Handling, SDH* 98: 98–109.

Jorgensen, B. S., and R. C. Stedman. 2001. "Sense of Place as an Attitude: Lakeshore Owners Attitudes towards Their Properties." *Journal of Environmental Psychology* 21 (3): 233–248. doi:10.1006/jevp.2001.0226.

Keßler, C., J. Krzysztof, and B. Mohamed. 2009. "An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval." In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information* Systems *(GIS '09)*, 91–100. New York: ACM Press. doi: 10.1145/1653771.1653787.

Kim, J., M. Vasardani, and S. Winter. 2017. "Similarity Matching for Integrating Spatial Information Extracted from Place Descriptions." *International Journal of Geographical Information Science* 31 (1): 56–80. doi:10.1080/13658816.2016.1188930.

Kõivumägi, E., M. Vait, A. Hadachi, G. Singer, and E. Vainikko. 2015. "Real Time Movement Labelling of Mobile Event Data." *Journal of Location Based Services* 9 (1): 55–76. doi:10.1080/17489725.2015.1032377.

Kurashima, T., T. Iwata, G. Irie, and K. Fujimura. 2010. "Travel Route Recommendation Using Geotags in Photo Sharing Sites." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 579–588. New York, NY: ACM. doi:10.1145/1871437.1871513.

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. "What Is Twitter, a Social Network or a News Media?" In *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 591–600. New York: ACM Press. doi:10.1145/1772690.1772751.

Lansley, G., and P. A. Longley. 2016. "The Geography of Twitter Topics in London." *Computers, Environment and Urban Systems* 58 (2016): 85–96. doi:10.1016/j.compenvurbsys.2016.04.002.

Lim, K. H., K. E. Lee, D. Kendal, L. Rashidi, E. Naghizade, S. Winter, and M. Vasardani 2018. "The Grass Is Greener on the Other Side: Understanding the Effects of Green Spaces on Twitter User Sentiments." In *Companion of the The Web Conference 2018 on The Web Conference 2018*, 275–282. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee.

Mckenzie, G., and B. Adams. 2017. "Juxtaposing Thematic Regions Derived from Spatial and Platial User-Generated Content." In *Leibniz International Proceedings in Informatics (Lipics)*, edited by E. Clementini, M. Donnelly, M. Yuan, C. Kray, P. Fogliaroni, and A. Ballatore, 1–13. Dagstuhl: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.COSIT.2017.20.

McKenzie, G., K. Janowicz, S. Gao, J. Yang, and Y. Hu. 2015. "POI Pulse: A Multi-Granular, Semantic Signature–Based Information Observatory for the Interactive Visualization of Big Geosocial Data." *Cartographica: the International Journal for Geographic Information and Geovisualization* 50 (2): 71–85. doi:10.3138/cart.50.2.2662.

Mountain, D., and A. MacFarlane. 2007. "Geographic Information Retrieval in a Mobile Environment: Evaluating the Needs of Mobile Individuals." *Journal of Information Science* 33 (5): 515–530. doi:10.1177/0165551506075333.

Nguyen, D., R. Gravel, D. Trieschnigg, and T. Meder. 2013. "How Old Do You Think I Am?: A Study of Language and Age in Twitter." In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 439–448. Palo Alto, CA: AAAI Press.

Nussbaum, D., M. T. Omran, and J. R. Sack. 2017. "Maintaining Anonymity Using -Privacy." *Journal of Location Based Services* 11 (1): 1–28. doi:10.1080/17489725.2017.1363419.

Poslad, S. 2001. "CRUMPET: Creation of User-Friendly Mobile Services Personalised for Tourism." In *Second International Conference on 3G Mobile Communication Technologies (3G 2001)*, 28–32. IEEE. doi:10.1049/cp:20010006.

Prelipcean, A. C., F. Schmid, and T. Shirabe. 2015. "A Space Time Alarm." In *Progress in Location-Based Services 2014. Lecture Notes in Geoinformation and Cartography*, edited by G. Gartner and H. Huang, 187–198. Cham: Springer. doi:10.1007/978-3-319-11879-6_13.

Relph, E. 1976. *Place And Placelessness*. London: Pion.

Resch, B., A. Summa, P. Zeile, and M. Strube. 2016. "Citizen-Centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm." *Urban Planning* 1 (2): 114–127. doi:10.17645/up.v1i2.617.

Richter, D., M. Vasardani, L. Stirlng, K. Richter, and S. Winter. 2013. "Zooming In–Zooming Out Hierarchies in Place Descriptions." In *Progress in Location-Based Services. Lecture Notes in Geoinformation and Cartography*, edited by J. M. Krisp, 339–355. Berlin: Springer. doi:10.1007/978-3-642-34203-5_19.

Richter, D., S. Winter, K. Richter, and L. Stirling. 2012. "How People Describe Their Place: Identifying Predominant Types of Place Descriptions." In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '12*, 30–37. New York: ACM Press. doi:10.1145/2442952.2442959.

Rosenkrans, G., and K. Myers. 2018. "Optimizing Location-Based Mobile Advertising Using Predictive Analytics." *Journal of Interactive Advertising* 18 (1): 43–54. doi:10.1080/15252019.2018.1441080.

Schlieder, C., T. Vögele, and A. Werner. 2001. "Location Modeling for Intentional Behavior in Spatial Partonomies." In *Proceedings of Ubicomp 2001: Workshop on "Location Modeling for Ubiquitous Computing"*, 63–70.

Shatford, S. 1986. "Analyzing the Subject of A Picture: A Theoretical Approach." *Cataloging and Classification Quarterly* 6 (3): 39–62. doi:10.1300/J104v06n03_04.

Shelton, T., A. Poorthuis, and M. Zook. 2015. "Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information." *Landscape and Urban Planning* 142: 198–211. doi: 10.1016/j.landurbplan.2015.02.020.

Skupin, A. 2014. "Making A Mark: A Computational and Visual Analysis of One Researcher's Intellectual Domain." *International Journal of Geographical Information Science* 28 (6): 1209–1232. doi:10.1080/13658816.2014.906040.

Twaroch, F. A., R. Purves, and C. Jones. 2009. "Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data." In *Proceedings of Workshop on Geographic Information on the Internet*. Toulouse: Springer.

Van Zanten, B. T., D. B. Van Berkel, R. K. Meentemeyer, J. W. Smith, K. F. Tieskens, and P. H. Verburg. 2016. "Continental-Scale Quantification of Landscape Values Using Social Media Data." *Proceedings of the National Academy of Sciences* 113 (46): 12974–12979. doi:10.1073/pnas.1614158113.

Vasardani, M., S. Timpf, S. Winter, and M. Tomko. 2013. "From Descriptions to Depictions: A Conceptual Framework." In *International Conference on Spatial Information Theory*, 299–319. Cham: Springer.

Villegas, N. M., C. Sánchez, J. Díaz-Cely, and G. Tamura. 2018. "Characterizing Context-Aware Recommender Systems: A Systematic Literature Review." *Knowledge-Based Systems* 140 (2018): 173–200. doi:10.1016/j.knosys.2017.11.003.

Winter, S., and C. Freksa. 2012. "Approaching the Notion of Place by Contrast." *Journal of Spatial Information Science* 5 (5): 31–50. doi:10.5311/JOSIS.2012.5.90.

Wohlin, C. 2014. "Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering." In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, 1–10. New York: ACM Press. doi:10.1145/2601248.2601268.

Ye, M., D. Shou, W. C. Lee, P. Yin, and K. Janowicz. 2011. "On the Semantic Annotation of Places in Location-Based Social Networks." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*, 520–528. New York, NY: ACM. doi:10.1145/2020408.2020491.

Zook, M., M. Graham, T. Shelton, and S. Gorman. 2010. "Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake." *World Medical and Health Policy* 2 (2): 6–32. doi:10.2202/1948-4682.1069.

B

PUBLICATION II: DESCRIPTION AND CHARACTERISATION OF PLACE PROPERTIES USING TOPIC MODELLING ON GEOREFERENCED TAGS

Bahrehdar, A. R. and Purves, R. S. (2018). Describing and characterising place using topic modelling on georeferenced tags. Journal of Geo-spatial Information Science: Special Issue on Crowdsourcing for Urban Geoinformatics, 21(3):173-184.

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2018.1493238

Taylor & Francis
Taylor & Francis Group

Description and characterization of place properties using topic modeling on georeferenced tags

Azam R. Bahrehdar and Ross S. Purves

Department of Geography, University of Zurich, Zurich, Switzerland

ABSTRACT

User-Generated Content (UGC) provides a potential data source which can help us to better describe and understand how places are conceptualized, and in turn better represent the places in Geographic Information Science (GIScience). In this article, we aim at aggregating the shared meanings associated with places and linking these to a conceptual model of place. Our focus is on the metadata of Flickr images, in the form of locations and tags. We use topic modeling to identify regions associated with shared meanings. We choose a grid approach and generate topics associated with one or more cells using Latent Dirichlet Allocation. We analyze the sensitivity of our results to both grid resolution and the chosen number of topics using a range of measures including corpus distance and the coherence value. Using a resolution of 500 m and with 40 topics, we are able to generate meaningful topics which characterize places in London based on 954 unique tags associated with around 300,000 images and more than 7000 individuals.

ARTICLE HISTORY
Received 28 November 2017
Accepted 4 June 2018

KEYWORDS
Place property; topic modeling; Volunteered Geographic Information (VGI); tagging

1. Introduction and motivation

How can we develop methods which better capture the diversity of ways of experiencing and understanding places, and yet which also allow representation and reasoning in information systems? One possible approach, which has recently gained much attention, is through the use of Volunteered Geographic Information (VGI), or more generally User-Generated Content (UGC), to derive place-relevant information that reflects notion of place as lived and experienced space (Capineri 2016; Hauthal and Burghardt 2016; Jenkins et al. 2016; Lansley and Longley 2016; Shelton, Poorthuis, and Zook 2015). An obvious strength of such data is the potentially large number of contributors, and corresponding potential multiplicity of ways of describing the same location. However, this strength is also a challenge – given such large volumes of data, we need methods which can allow us to identify coherent themes, or topics, if we wish to be able to characterize and compare places in a useful way (Adams and McKenzie 2013).

This need for coherent summaries of place-related data is underlined by the growth in location-based services and associated web-mapping products. Here, we observe a rapid increase in the development of services capable of adapting to individual users and use contexts, for instance by identifying preferences for a particular activity through previous actions or discriminating between tourist and local interests (Huang 2016; Nivala and Sarjakoski 2003). Such approaches, implicitly or explicitly, recognize that we think about and perceive the world in terms of places, rather than as geometric coordinates detached from meaning. Thus, in developing approaches which can adapt content according to shared notions of place, there is a need for data which represent these concepts.

Increasing calls have been made for the need to model and reason using place-based concepts in Geographic Information Science (GIScience). This is reflected by work, first, considering spatial vagueness as an important property of cognitive models of place, and second, a realization that natural language can provide us with access to a multiplicity of ways in which place is conceptualized (Montello et al. 2003). Much of this research is, at least in passing, inspired by ideas developed in human geography. Key to the work described in this article is the notion of place as being a socially produced concept (De Certeau 1984; Dourish 2006) associated with not only locals (Harrison and Tatar 2008) but also having an identity from people connected with places at a global level (Massey 1993). In GIScience, Agnew's model (Agnew 2011), which conceptualizes three dimensions of place related to location, locale, and sense of place, has proved popular. These dimensions are often interpreted as relating to named places (locations), their properties or affordances (locale), and the meanings and emotions that people associate with these places (sense of place) (Capineri 2016; Hollenstein and Purves 2010; MacEachren 2017). It

CONTACT Azam R. Bahrehdar azam.bahrehdar@geo.uzh.ch

is also clear that such notions of place are dynamic, since place can also be considered to emerge as a semantic tangle of people related to activities and events at a locus (Harrison and Tatar 2008).

Natural language data, in the form of texts describing locations, are one way of attempting to build place descriptions. One, often-discussed source of such data is the Flickr photo-sharing platform. There are a number of reasons for this popularity. First, a large number of Flickr images are georeferenced, and their metadata are easily accessible through an application programming interface (Smith et al. 2012). Second, an image is the immediate and straightforward way of capturing our interactions with place, and early research demonstrated that coherent information related to both places and events could be extracted from Flickr tags (Rattenbury, Good, and Naaman 2007). Third, Flickr, has been shown to be used by different sorts of users, allowing for example access to contrasting conceptualizations related to both locals and tourists (Straumann, Çöltekin, and Andrienko 2014). Fourth, tags, given their lack of syntax are relatively simple to process, allowing the rapid implementation of arguably naïve, annotation and co-occurrence studies (Hollenstein and Purves 2010; Purves, Edwardes, and Wood 2011). More generally, increasing access to UGC has led to many claims with respect to the possibilities of characterizing place in a wide variety of ways from essentially bottom-up sources (Dunkel 2015; Shelton, Poorthuis, and Zook 2015). As well as simple studies, focusing on frequency and co-occurrence of tags, other methods include a variety of approaches from natural language processing to, for example, cluster and aggregate content semantically and spatially, and extract and characterize sentiment (Davies 2013; Hauthal and Burghardt 2016; Jenkins et al. 2016; Vasardani et al. 2013).

One very commonly applied family of methods in natural language processing, used to meaningfully group documents in a large corpus, is topic modeling (Blei and Lafferty 2006). The basic idea is relatively simple – given a set of documents, made up of individual words, it should be possible to group these using co-occurrence (i.e. documents in which similar words co-occur are more likely to be related). Perhaps the most common approach to topic modeling is Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). LDA is a probabilistic approach which outputs a user-defined number of topics, each represented as a multinomial distribution over words. This implies, since documents consist of words, that documents can be represented as a mixture of topics. If topics can be assigned to meaningful labels, and if documents belong more to some topics than others, then a collection of documents can be summarized in terms of the topics, the words associated with each topic, and their labels. From this brief explanation, and the plethora of literature associated with LDA, two things should become clear. First, topic modeling appears to offer a beguiling simple way of summarizing large sets of documents. Second, the number of topics and labels attached to topics are chosen, and interpreted, by people. Topic modeling simply returns the number of topics defined as an input parameter and, presumably, if a corpus consists of a set of very similar documents, these topics should in turn be very similar (and thus not capture nonexistent semantic differences). However, the interpreting probabilities is generally known to be hard for humans, and issues such as semantic coherence, topic significance and ranking and the use of topic modeling in exploring data have all been the subject of attention (AlSumait et al. 2009; Chang et al. 2009; Mimno et al. 2011).

Topic models have obvious potential applications to understanding place and have been used in this context (Jenkins et al. 2016). On the one hand, we might expect documents describing the same place, but looking at different aspects of it to be captured in topics related to the place name (or its location). On the other hand, different places, affording similar environments, might be captured in topics focusing on locale. And finally, places which evoke similar emotions, we might imagine, could be captured in topics related to sense of place. Adams and McKenzie (2013) analyzed georeferenced travel blogs using LDA, and indeed observed that four categories of topics emerged: what they called *localities* (specific geographic locations), *activities* and *features* (things to see and do), and *miscellaneous*. They demonstrated that LDA could generate meaningful, place-related topics but focused on understanding individual topics and similarities of locations to these.

In this article we focus on the use of image descriptions as a source of place information, or more specifically the tags associated with Flickr images. Since topic models treat documents as bags of words, documents based around tags (which can be considered to be simple sets of terms) are particularly well-suited to topic modeling since no underlying syntax is discarded in the analysis. Similar to the approach of Adams and McKenzie (2013), in this article we explicitly generate topic models in space, but our starting point are not individually authored documents, but rather all of the tags associated with a grid cell. Since previous work has shown that parameter choices and interpretation of topics models are not trivial, we explicitly set out to explore the extent to which our approach allows us to capture different aspects of place and the sensitivity of our results. By aggregating textual information associated with a cell, we aim to explore the shared meaning and descriptions of places from/for people who either live in or

visit these locations. Finally, we link these descriptions to a model of place to explore different ways in which London is described through Flickr tags. Our contribution is thus threefold:

(1) We use LDA to generate spatially explicit topics in London. Our model is spatially continuous, and thus every location is associated with a set of topics.
(2) Since parameter choice has been shown to be important in LDA, we explore the sensitivity of our results to both the number of topics and grid resolution. Furthermore, we use topic measures to explore the extent to which semantically coherent topics are distinctive.
(3) We interpret and classify individual topics, relating these to place properties derived from the literature.

2. Data

Data were gathered using queries to Flickr's Application Programming Interface (API) for georeferenced images within a given bounding box and taken before July 2013. Metadata included user ids, tags, image coordinates, two timestamps referring to the times a photo was taken and uploaded, and accuracy information provided by Flickr with respect to coordinates. Note that metadata reporting on accuracy in Flickr actually better reflect precision, and are often used to filter imprecisely georeferenced data (Hollenstein and Purves 2010). Our case study region is centered around the River Thames in inner London (Figure 1) and includes very commonly photographed places such as Buckingham Palace, Hyde Park, and Tower Bridge (Crandall et al. 2009) and has a total area of 170 km^2.

2.1. *Data filtering and cleaning*

Our focus was on modeling place by capturing shared notions ascribed to georeferenced images through tagging. Before carrying out topic modeling, we first carried out a range of filtering steps. We first removed images with accuracy values lower than 15 (i.e. georeferences reported as being less precise than street level). Second, bulk uploads, images with identical tags, either a textual tag or geotag from a single user, and tags which were not meaningful (e.g. camera generated titles "DIC 0001") were removed using regular expressions. Furthermore, since tagging is known to be influenced by behavior, we removed users with the following characteristics:

(1) Very inactive users who had a single image in our dataset or less than ten images in total associated with their profiles over a 24-h period (i.e. users experimenting with the system) (Hollenstein and Purves 2010);
(2) Users who had deleted their profiles since our data collection;
(3) Prolific users may introduce large biases in UGC, and in particular can clearly mask more general shared meanings (Nielsen 2006; Hollenstein and Purves 2010). We removed the 1% most prolific users who generated 20% of the whole dataset.

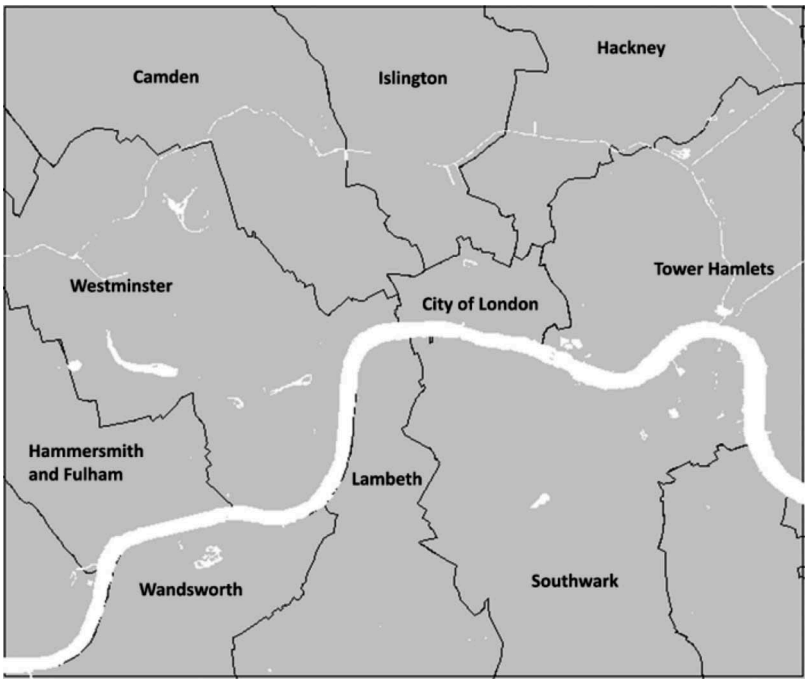


Figure 1. Study area within inner London.

Moreover, the following images were removed:

(1) Images with no tags;
(2) Images with only Flickr machine generated tags which thus do not represent shared notions of place created by an individual user.

The final dataset thus consisted of 7753 unique users who had shared 371,752 images. Table 1 shows the effect of each filtering task over the number of images and users, with a reduction in the original number of images collected, of approximately 90%.

Finally, the remaining 186,632 unique tags associated with the 371,752 images reported in Table 1 were normalized by being converted to lower case. Special characters (such as, @ in "@park"), numbers (e.g. the 2 in "park2"), and stop-words (such as, a, an, the) were removed. In addition, we eliminated all tags consisting only of numbers. We did not control for typographical errors (e.g. match londin to london) or remove duplicate tags associated with a single image.

Even after filtering, it is still possible that an individual user can bias usage of individual tags. We therefore generated tag profiles (Hollenstein and Purves 2010) which for each tag reflect tag usage over the population as a whole. We then used the coefficient of variation of standardized tags contribution to measure whether a tag was used equally among users with different contribution patterns. Tags with high coefficients of variation are only used by a few people and are therefore subject to contribution bias. We eliminated tags with a high coefficient of variation (> 200) (Hollenstein and Purves 2010) from our set of unique tags. The final tag list thus contained 954 unique commonly used tags, which formed the basis for the topic modeling described next.

Table 1. Remaining numbers of Flickr images and users after applying each task of the data filtering.

Function	Images	Contributors
Original dataset	3,105,544	49,130
Accuracy filtering	1,047,003	31,092
Bulk-upload filtering	839,822	31,080
Camera generated contents (either titles or tags)	571,241	30,377
Inactive users	503,536	8143
Prolific users	404,329	8060
Null tags	371,752	7753

2.2. *Spatial distribution of Flickr images and corresponding metadata*

The density of contributors to our dataset after filtering and cleaning our data is shown in Figure 2. The map shows that the concentration of Flickr users, in Central London, particularly to the west, is higher. We assume this is because of tourist and leisure attractions in this area, since some of the most photographed places in the world are located in the western part of London (Crandall et al. 2009).

The correlation between the number of users and corresponding contributed images, using a linear regression, is very high (r^2 = 0.95). Since we expect users at a given location to be spatially autocorrelated (Tobler 1970; Miller 2004), we tested for the influence of spatial autocorrelation using a spatial autoregressive regression (SAR) model including the coordinates of grid cells in the model. The correlation value (r^2 = 0.96) is very similar, suggesting that the influence of spatial autocorrelation on our model is limited, and that the number of images in a grid cell is indeed strongly linked to the

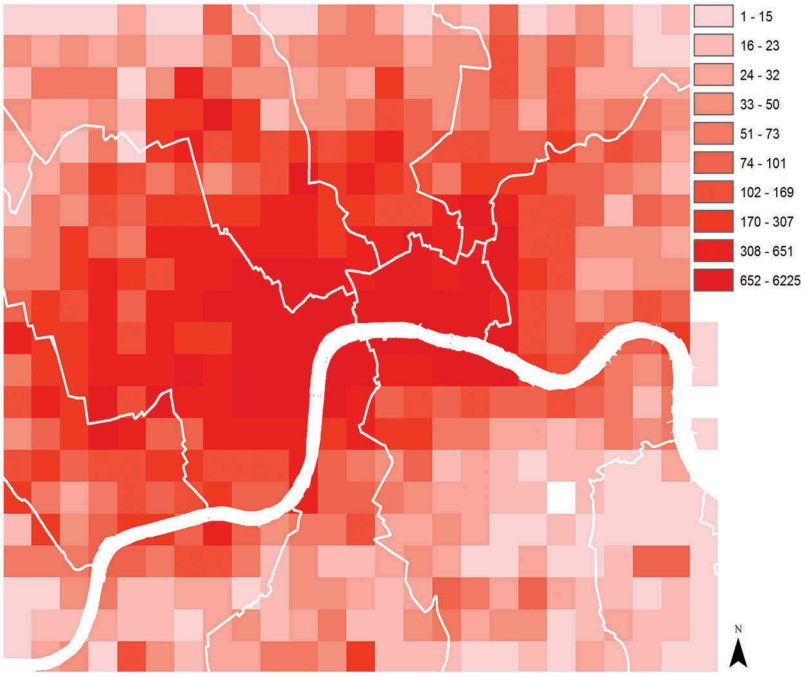


Figure 2. Number of users taking photographs in each grid cell.

number of photographers. The high correlation between the number of users and images demonstrates that individual users do not influence the spatial distribution of images, and that the contribution bias, both in space and semantically, as a result of filtering for tags with low coefficient of variation, is no longer a major influence on our data.

3. Methods

Our focus in this article was on using semantics to group locations which are associated with shared meanings. We chose to do so by overlaying our study area with a grid and treating grid cells as the basic spatial units for analysis. We therefore treated each cell as a textual document, where all tags from all images located within a cell constitute the content of a single document. We then used topic modeling to explore the characteristics of, and in particular to group, similar grid cells. In a first stage we tested the sensitivity of our approach to the spatial resolution of our grid and the parameters used in our topic modeling. Having identified an optimum resolution and set of parameters we then labeled individual topics and finally annotated these labels according to the conceptual models of place introduced by Agnew (2011) and Harrison and Tatar (2008).

3.1. *Topic modeling*

In the introduction we described the basic principles of topic modeling. We used the Machine Learning for Language Toolkit (MALLET) to carry out LDA (McCallum 2002). Here, we explain how we generated topics for our data.

(1) Documents for input to LDA were grid cells, each associated with a vector of all occurrences of each of the 954 unique tags identified after filtering;
(2) These documents were input to the MALLET LDA toolkit and optimized hyper parameters for a given number n of topics (Cao et al. 2009) calculated;
(3) The following outputs were produced:
    a. For each of n topics, a list of all tokens (tags) and their probabilities of belonging to that topic;
    b. For each grid cell (document), a vector of n topics and the probability of the grid cell belonging to each topic;
    c. For each topic, a set of measures describing topic quality, which we introduce later.
(4) We then assigned the most probable topic to each grid cell. Tags associated with low probabilities are not useful in characterizing an individual topic (Aletras et al. 2017) and we therefore chose representative tags by sorting tags associated with a topic according to probability, and then exploring the resulting cumulative probability curves.

Since our approach is based on a grid, the result is a spatially continuous model characterizing locations in terms of the tags which best describe each cell. An obvious limitation of this approach is the Modifiable Areal Unit Problem (MAUP) (Openshaw 1983). In addressing MAUP, we focused on the *scale effect* – the influence of the size of the units over which data are aggregated. We explored the influence of MAUP by testing our results for four different resolutions: 50 m, 250 m, 500 m, and 1000 m resolution cells.

MALLET also outputs a range of measures which aim to characterize the quality or meaningfulness of the output topics. We selected three of these *corpus distance, number of tokens*, and *coherence value* to investigate first the sensitivity of the model to resolution and number of topics (using *corpus distance* and *number of tokens*) and second, the semantic qualities of our topics (using *coherence value*).

Corpus distance characterizes how similar a topic is to the corpus as a whole. Small corpus distances imply that topics are similar to the corpus, and thus have limited power to distinguish documents from the corpus, or in our case, to differentiate between places with different characters (AlSumait et al. 2009). *Number of tokens* gives some indication of the number of words associated with each topic. As the number of topics increases (or the resolution decreases), the number of tokens associated with topics might be expected to decrease (since the need to generalize over locations and topics is less). An optimum number of tokens is therefore both sufficient to characterize individual topics, but small enough to allow topics to be distinguished from one another (c.f. corpus distance) (Mimno et al. 2011). These two measures were thus used in our sensitivity study to optimize grid resolution and number of topics.

The *coherence value* is based on the probability of words in a topic co-occurring in the grid cells belonging to that topic. It is calculated by taking the log of the sum of the probabilities of co-occurrence as a function of higher ranked words belong to a topic:

$$coherence = \sum_{i} \sum_{j<i} \log \frac{D(w_j, w_i) + \beta}{D(w_i)} \qquad (1)$$

where β is a parameter to prevent log zero errors, $D(w_j, w_i)$ is the number of co-occurrences of two terms in a document, and $D(w_i)$ is the number of occurrences of the more probable terms.

Very negative (since the value is a log) coherence values indicate that the tokens in a topic rarely

co-occur in grid cells, while values of coherence close to zero suggest semantically coherent topics and associated tokens (Stevens et al. 2012).

3.2. *Topic labeling and annotation*

The final step in our methods moved away from computational methods to identify coherent topics using LDA and focused on the interpretation of these topics. Our aim here was twofold: first, we wished to assign a label to each topic, and second, to characterize topics according to notions of place introduced earlier. Our underlying hypothesis was that by using UGC, in the form of Flickr tags, we could extract semantics characterizing locations relating to similar places that might otherwise go unnoticed (Goodchild 2007). Having a list of most probable words based on the topic modeling, we set out to interpret these topics. Crucially, the local knowledge was central to interpreting topics, since individual tokens are often ambiguous and need to be interpreted in terms of London's geography and the other tokens with which they co-occur. Thus, for example, the tokens *Kings Cross, railway* and *Paddington* would suggest a railway-related cluster (since these are the names of two nearby London railway stations). Since labeling topics varied in its difficulty, we only labeled those where we were reasonably confident of our interpretation. We hypothesized that these topics would also have higher coherence values, since the previous works have suggested that the quality of topics can also be expressed in terms of their interpretability by humans (Mei, Shen, and Zhai 2007; Newman et al. 2010).

In the final step, we annotated our labels with respect to place descriptions based around conceptual models of place focusing on first, the nature of place itself (Agnew 2011) and second the importance of the actors in a given place (Harrison and Tatar 2008). We used the following categories and combinations thereof: *location* (labels related to named places), *locale* (labels describing affordances of a place, either in terms of explicit *activities* or the objects characterizing a place), *sense of place* (labels associated with emotions and feelings), and finally, *people* (labels describing characteristics of the individuals or groups associated with a place).

4. Results and interpretation

4.1. *Sensitivity tests*

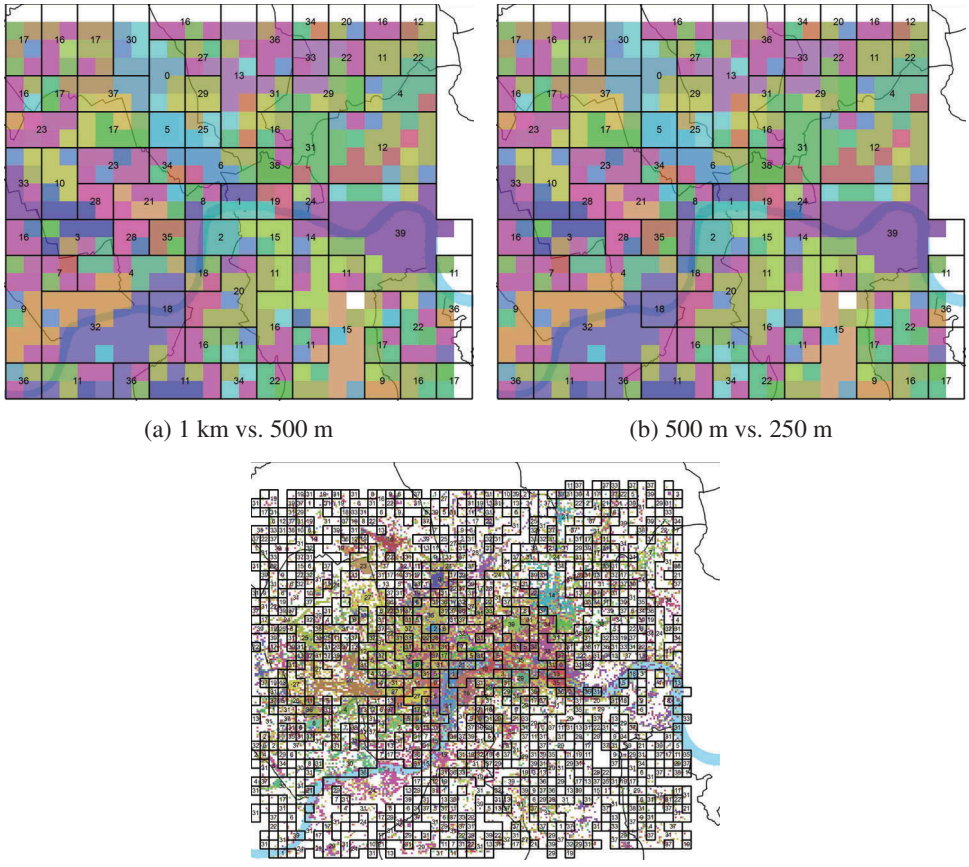
The first set of results we present concern sensitivity tests used to identify optimum grid resolutions and numbers of topics for further analysis. Table 2 summarizes key statistics for the measures we introduced earlier for four grid resolutions and five different values for the number of topics. Median number of tokens showed no correlation with resolution, and we therefore report only on corpus distance.

Table 2. Median corpus distance and number of cells per topic as a function of the number of topics for different grid resolutions.

Resolution (m)	No. of topics	Median no. of cells	Median corpus distance
50	20	678.0	2.32
	40	349.0	3.07
	60	228.5	3.36
	80	166.5	3.59
	100	131.0	3.81
250	20	64.0	2.09
	40	32.5	2.80
	60	18.5	3.13
	80	14.0	3.30
	100	11.0	3.44
500	20	16.0	1.90
	40	10.0	2.41
	60	5.0	2.69
	80	4.0	2.96
	100	3.0	3.27
1000	20	5.5	1.52
	40	2.0	2.21
	60	2.0	2.46
	80	1.0	2.75
	100	1.0	2.93

Mean median corpus distance is strongly correlated with resolution (Pearson correlation: $r^2 = 0.95$) suggesting that the most distinctive topics would be obtained by simply having high resolutions. However, as resolution becomes finer, so too does the number of grid cells not allocated to any topic, because increasingly large numbers of grid cells are not associated with tags. This effect is illustrated in Figure 3 for clusters of 40 topics for four different grid resolutions. The colors reflect the clusters of higher resolution and the black wireframes delineate the low-resolution clusters. The white grid cells could not be allocated to a topic at the higher resolution, because no tags were present in these cells. To balance between very coarse resolutions (where meaningful places are not delineated) and fine resolutions (where for many cells we have insufficient data to describe places), 500 m was identified as an optimum grid resolution – the colored patches in Figure 3(a) and the black outlines in Figure 3(b).

Having identified a suitable resolution, we then explored the sensitivity of our results to the number of topics. Figure 4 shows an inflection point in corpus distance, irrespective of resolution, at 40 topics, suggesting that the biggest change in the distinctiveness of our topics is likely to occur if we increase the number of topics from 20 to 40. As with resolution, simply increasing the number of topics results in higher corpus distances and thus more distinct topics. However, we also explored the sensitivity of the number of topics to two further parameters, both of which are important to our overall aim of delineating meaningful places.



(a) 1 km vs. 500 m

(b) 500 m vs. 250 m

(c) 250 m vs. 50 m

Figure 3. Comparison between clusters of 40 topics with respect to the grid resolution. (a) 1 km vs. 500 m; (b) 500 m vs. 250 m; (c) 250 m vs. 50 m.

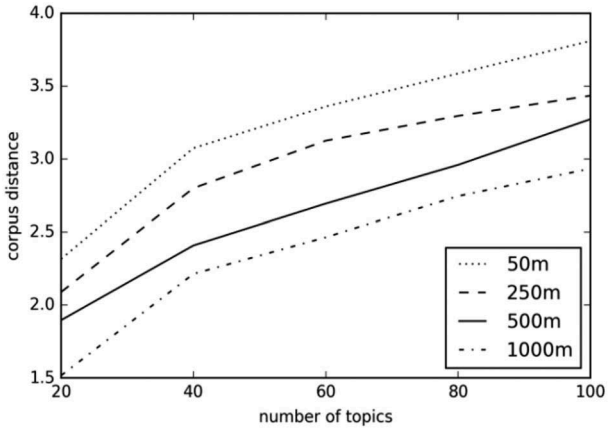


Figure 4. Change in median corpus distance for different number of topics with respect to grid resolution.

First, in Figure 5 we plot corpus distance as a function of the number of cells associated with each topic. Here we observe that corpus distance not only varies as a function of the number of topics, but also the number of cells, or area, associated with a topic. A desirable property of our results is that the distinctiveness of our topics does not strongly vary as a function of area – in other words that topics associated with single cells are not much more distinctive than those associated with large areas or vice versa. We observe that 40 topics seems to have the most stable corpus distance as a function of the number of cells associated with a topic.

Second, we explored the relationship between the number of cells assigned to each topic and the number of topics (as shown in Figure 6). Once again, we observe that the most stable behavior appears to be for 40 topics – in other words, we have a roughly equal distribution of topics with the areas in range of 0.25–1 km^2, 1–2 km^2, and 2–3 km^2.

In summary, based on our detailed sensitivity tests we found a resolution of 500 m best suited to capturing the whole area of interest, while maximizing corpus distance. Selecting 40 topics allowed us generate topics with a roughly constant corpus distance as a function of area. This in turn means that our results are not biased to either topics covering only very large or small areas.

4.2. *Labeling and exploring topics*

Having identified an optimum resolution and number of topics, we then set about analyzing the meaning of the topics created. Based on cumulative

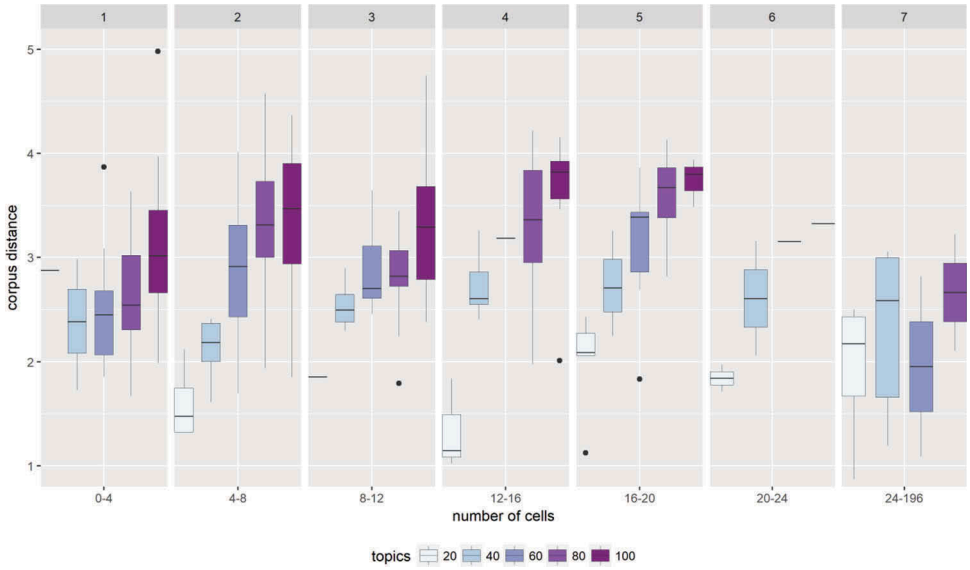


Figure 5. Corpus distance for topics associated with different numbers of cells at a resolution of 500m.

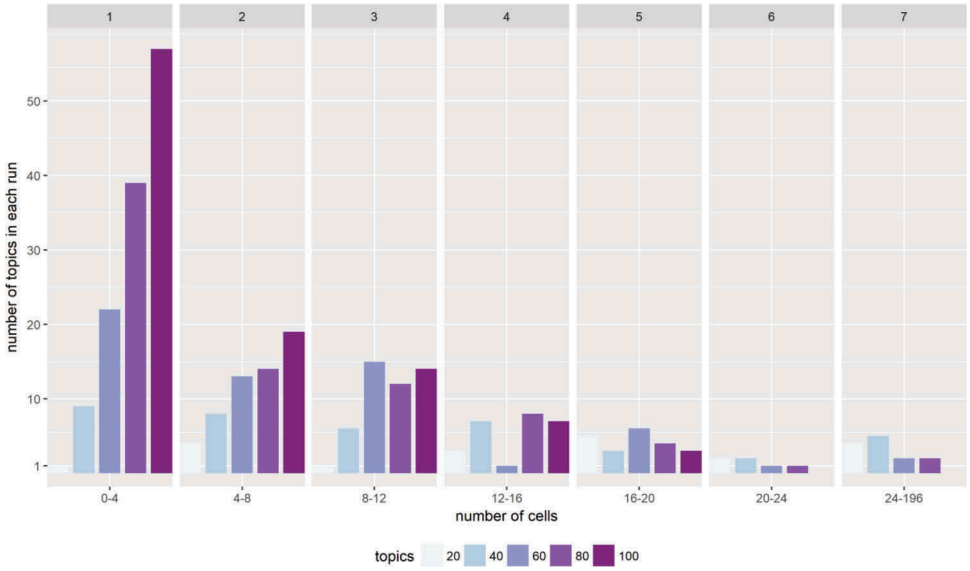


Figure 6. Number of topics associated with the 500m cells for each implementation of the model.

probabilities with respect to tags associated with individual topics, we selected lists of representative tags for each topic. These lists typically contained 15–25 tags. We then attempted to label topics based on these tags and our local knowledge of London. However, it is important to note that we could label only 30 out of 40 topics. We had previously hypothesized that, based on literature, topics which we could label were more likely to have low coherence values. In Figure 7, we plot coherence values for the 30 labeled topics and 10 unlabeled topics and observe that coherence value does indeed appear to be a good potential indicator of the likelihood of topics being interpretable by humans.

Figure 8 allows us to explore the different ways in which the semantics and properties of place are captured by our topic modeling. Note that we removed the two most probable tags from topics 19 (zoo) and 34 (natural) to increase clarity. The first topic, Topic 1 (*views*) is distributed over a range of locations (Figure 8(a)), and mostly includes terms describing general features of scenes (e.g. sunset, clouds, skyline in Figure 8(b)) which are photographed, thus indicating generic views of London. Interestingly, this topic is scattered around the edge of our study area, indicating locations from which London is seen. These places are thus characterized not only by what is found in these locations, but also by what can be seen from them.

The other three examples all capture specific locations, either as a single cell (Topic 34: *South Kensington Museums*) or a cluster of cells (Topic 19: *London Zoo* and Topic 32: *Along the Thames*) (Figure 8(a)). Examining the tag clouds, we observe a mixture of mostly proper nouns in the form of toponyms and building names (e.g. southbank,

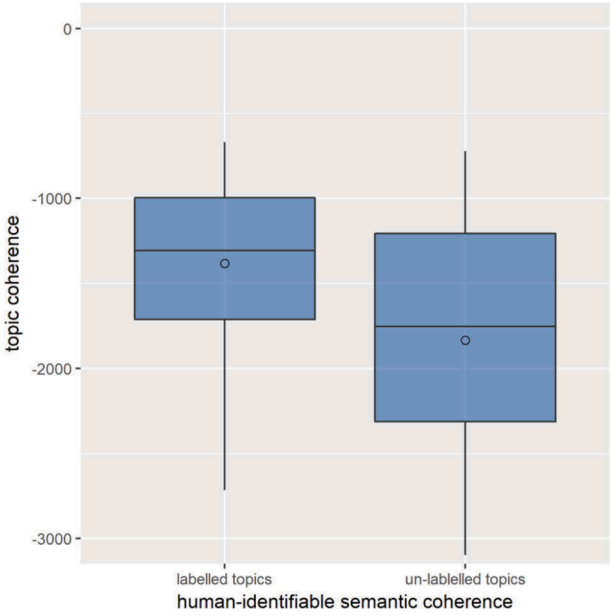


Figure 7. Topic coherence value for labeled and unlabeled topics.

londra, gherkin), nouns (e.g. butterfly, cloud, skyline, family), and more abstract terms (e.g. assembly, authority) (Figure 8(b)).

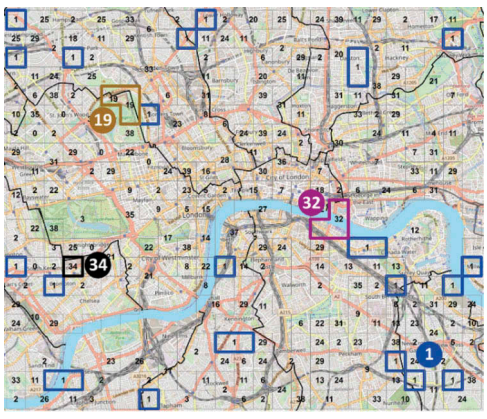
To better understand the nature of terms used in each topic, we associated our labels with a simple taxonomy based on previous work on place. In Figure 9, we show both the 30 labeled topics and a classification of these labels in terms of the five dimensions of place we introduced earlier (location, locale and activity, sense of place and people). The map illustrates well the contrast between, for instance, topics based around locations (e.g. Barbican, Piccadilly), locales (e.g. canals, trains, and stations), and combinations of locations and locales (e.g. Hyde Park, which contains both location and locale information). We found no topics which were clearly related to sense of place, which we interpreted as emotions and feelings, but otherwise a mix of the types proposed.

As has been shown in previous research (Sigurbjörnsson and Van Zwol 2008), toponyms are an important way of describing images, and thus can be effectively used as labels for topics. However, the map also allows us to see that such topics can extend beyond the actual location associated with a toponym (e.g. as occurs for Piccadilly), thus suggesting that such topics actually describe both the place Piccadilly and other similar 2We suggest that some of the classes also seem likely to reflect different sorts of users: views, canals, trains, and stations are distributed across London and seem likely to be indicative of locals interested in certain sorts of views and narratives about the city, rather than visitors characterizing tourist attractions (e.g. London Zoo or the museums in South Kensington). However, this visualization also illustrates some of the challenges of extracting semantics from tags, where we can only assign labels by interpreting and making assumptions about associations between tags. In general, we also note that most of the activities are leisure activities, suggesting that Flickr is typically used to document a mixture of tourist and leisure activities, and also hinting at what might be missing in such characterizations (e.g. more mundane activities and those with less positive associations).

5. Concluding discussion

We are not the first authors to use LDA as approach to describe space, or indeed, to link these notions to place (Adams and McKenzie 2013; Jenkins et al. 2016). Rather, our most important contribution is carrying out a detailed sensitivity study with respect to both resolution and number of topics, and assessing the utility of a range of out of the box measures in describing the quality of our results. Based on our experiences, we make the following suggestions:

(1) Assuming that a spatially continuous model is the aim of a study, then the optimum grid



(a) Labeled topics (numbered cells)

(b) Example topics, labels, and tags

Figure 8. Labeled topics (numbered cells) and example topics, labels, and tags (size as a function of probability).

Figure 9. Map of London describing users' perception of the space as places.

resolution is that which allows most (or all) grid cells to be allocated to topics;

(2) Increasing the number of topics will on average lead to more distinct topics. However, these topics will become increasingly associated with single grid cells, and thus fail to identify similar (not necessarily contiguous) regions. An optimal number of topics is, we would argue, one which allows for a range of topic areas (i.e. numbers of grid cells) and where corpus distance is not strongly influenced by the area associated with a topic;

(3) Topic coherence value is a good predictor of the likelihood of humans being able to interpret and label topics.

Our labeled topics and their classification demonstrate both some strengths, and key limitations of our method. Firstly, after filtering (an important step which is often only cursorily described), we are still left with sufficient semantic variation to generate meaningful semantic topics which both describe specific locations (instances of places) and generic locations (types of places, or groups of similar places). However, since we labeled clusters only according to their semantics and not the locations of grid cells belonging to each cluster, label names alone are not indicative of membership in one of these groups. Thus, our Hyde Park cluster appears to actually encompass not only Hyde Park (an instance of a place) but also Hyde Park-like places. Using tags describing Flickr images obviously biases us toward the visual, and this is particularly well illustrated in our views cluster, where many generic salient, aesthetically pleasing, features of a cityscape are prominent (Dunkel 2015). On the other hand, as has been shown by other authors, we find little direct evidence for terms relating to sense of place (Hauthal and Burghardt 2016) in the sense of emotions and feelings. Indeed, our approach, though it captures shared meanings which relate to coherent places, is data-driven, and since Flickr images are dominated by more positive experiences (Cox, Clough, and Marlow 2008), does not reflect more negative aspects of place. Identifying and integrating data containing such notions would be an important extension to this work, but this is nontrivial, since many other sources also have a less direct relation to the space being described (Hahmann, Purves, and Burghardt 2014). Although we address the MAUP by exploring the sensitivity of our results to scale, we assume that our results are relatively insensitive to the shape and origin of our grid. One possible way of exploring this issue further would be to use an adaptive grid, and also to explore sensitivity to the grid's origin. In future work we will therefore concentrate on methods to effectively integrate data from multiple sources, across a range of scales, and link these data to places either in the form of bona fide objects (e.g. Tower Bridge) or fiat locations (such as, the east end of London).

Acknowledgments

The authors would like to thank Olga Chesnokova for many useful comments and suggestions.

Funding

This research was funded by the Swiss National Science Foundation Project PlaceGen [grant number 200021_149823].

Notes on contributors

Azam R. Bahrehdar is a Ph.D. student in the Geocomputation Unit at the Department of Geography at the University of Zurich. Currently she is working on understanding spatial and platial context of User-Generated Content (UGC) through text.

Ross Purves heads the Geocomputation Unit at the Department of Geography at the University of Zurich. His research interests include Geographic Information Retrieval, uncertainty modeling and characterizing place and landscapes using unstructured text and social media.

References

Adams, B., and G. McKenzie. 2013. "Inferring Thematic Places from Spatially Referenced Natural Language Descriptions." In *Crowdsourcing Geographic Knowledge*, edited by D. Sui, S. Elwood, and M. Goodchild, 201–221. Dordrecht, Netherlands: Springer.

Agnew, J. A. 2011. "Space and Place." In *The SAGE Handbook of Geographical Knowledge*, edited by J. Agnew and D. Livingstone, 316–330. London, United Kingdom: SAGE Publications .

Aletras, N., T. Baldwin, J. H. Lau, and M. Stevenson. 2017. "Evaluating Topic Representations for Exploring Document Collections." *Journal of the Association for Information Science and Technology* 68 (1): 154–167. doi:10.1002/asi.23574.

AlSumait, L., D. Barbará, J. Gentle, and C. Domeniconi. 2009. "Topic Significance Ranking of LDA Generative Models." In *Machine Learning and Knowledge Discovery in Databases*, edited by W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, ECML PKDD 2009. Lecture Notes in Computer Science. Vol. 5781. 67–82. Berlin, Heidelberg: Springer-Verlag.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Blei, D. M., and J. D. Lafferty. 2006. "Dynamic Topic Models." *Proceedings of the 23rd International Conference on Machine Learning – ICML '06*, 113–120. Pennsylvania, USA, June 25–29.

Cao, J., T. Xia, J. T. Li, Y. D. Zhang, and S. Tang. 2009. "A Density-Based Method for Adaptive LDA Model Selection." *Neurocomputing* 72 (7–9): 1775–1781. doi:10.1016/j.neucom.2008.06.011.

Capineri, C. 2016. "Kilburn High Road Revisited." *Urban Planning* 1 (2): 128–140. doi:10.17645/up.v1i2.614.

Chang, J., S. Gerrish, C. Wang, and D. M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Journal of Physics A: Mathematical and Theoretical* 44 (8): 085201.

Cox, A., P. Clough, and J. Marlow. 2008. "Flickr: A First Look at User-Behavior in the Context of Photography as Serious Behavior." *Information Research* 13 (1): 1–20.

Crandall, D. J., L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. "Mapping the World's Photos." *Proceedings of the 18th International Conference on World Wide Web – WWW '09*, 761–770. Madrid, Spain, April 20–24.

Davies, C. 2013. "Reading Geography between the Lines: Extracting Local Place Knowledge from Text." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, edited by T. Tenbrink, J. Stell, A. Galton, and Z. Wood, 320–337. Vol. 8116.Berlin, Heidelberg: Springer.

De Certeau, M. 1984. *The Practice of Everyday Life*. Berkeley, USA: University of California Press.

Dourish, P. 2006. "Re-Space-Ing Place." *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work – CSCW '06*, 299. Alberta, Canada, November 4–8.

Dunkel, A. 2015. "Visualizing the Perceived Environment Using Crowdsourced Photo Geodata." *Landscape and Urban Planning* 142: 173–186. doi:10.1016/j.landurbplan.2015.02.022.

Goodchild, M. F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–221. doi:10.1007/s10708-007-9111-y.

Hahmann, S., R. Purves, and D. Burghardt. 2014. "Twitter Location (Sometimes) Matters: Exploring the Relationship between Georeferenced Tweet Content and Nearby Feature Classes." *Journal of Spatial Information Science* 9 (9): 1–36.

Harrison, S., and D. Tatar. 2008. "Places: People, Events, Loci – The Relation of Semantic Frames in the Construction of Place." *Computer Supported Cooperative Work* 17 (2–3): 97–133. doi:10.1007/s10606-007-9073-0.

Hauthal, E., and D. Burghardt. 2016. "Mapping Space-Related Emotions Out of User-Generated Photo Metadata considering Grammatical Issues." *The Cartographic Journal* 53 (1): 78–90. doi:10.1179/1743277414Y.0000000094.

Hollenstein, L., and R. Purves. 2010. "Exploring Place through User-Generated Content: Using Flickr to Describe City Cores." *Journal of Spatial Information Science* 1 (1): 21–48.

Huang, H. S. 2016. "Context-Aware Location Recommendation Using Geotagged Photos in Social Media." *ISPRS International Journal of Geo-Information* 5 (11): 195. doi:10.3390/ijgi5110195.

Jenkins, A., A. Croitoru, A. T. Crooks, and A. Stefanidis. 2016. "Crowdsourcing a Collective Sense of Place." *PLOS ONE* 11 (4): 1–20. doi:10.1371/journal.pone.0152932.

Lansley, G., and P. A. Longley. 2016. "The Geography of Twitter Topics in London." *Computers, Environment and Urban Systems* 58: 85–96. doi:10.1016/j.compenvurbsys.2016.04.002.

MacEachren, A. M. 2017. "Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier." In *Advances in Geographic Information Science*, edited by C. Zhou, F. Su, F. Harvey, and J. Xu, 139–155. Berlin, Heidelberg: Springer-Verlag.

Massey, D. 1993. "Power-Geometry and a Progressive Sense of Place." In *Mapping the Futures*, edited by J. Bird, B. Curtis, T. Putnam, and L. Tickner. Vol. 11. London, UK: Routledge.

McCallum, A. K. 2002. MALLET: A Machine Learning for Language Toolkit. *http://mallet.cs.umass.edu*. [last accessed on June 14, 2018].

Mei, Q., X. Shen, and C. X. Zhai. 2007. "Automatic Labeling Of Multinomial Topic Models. "Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 490–499. San Jose, CA: Agues 12–15.

Miller, H. J. 2004. "Tobler's First Law and Spatial Analysis." *Annals of the Association of American Geographers* 94 (2): 284–289. doi:10.1111/j.1467-8306.2004.09402005.x.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." *Proceedings of the 2011 Conference on*

Empirical Methods in Natural Language Processing, 262–272. Edinburgh, United Kingdom, July 27–31.

Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl. 2003. "Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries." *Spatial Cognition & Computation* 3 (2–3): 185–204.

Newman, D., J. Lau, K. Grieser, and T. Baldwin. 2010. "Automatic Evaluation of Topic Coherence." In *HLT'10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. Los Angeles, California, June 2–4.

Nielsen, J. 2006. "The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities." *https://www.nngroup.com/articles/participation-inequality/*[last accessed on June 14, 2018].

Nivala, A., and L. T. Sarjakoski. 2003. "Need for Context-Aware Topographic Maps in Mobile Devices." *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science (ScanGIS)*, 15–29. Espoo, Finland, June 4–6.

Openshaw, S. 1983. "The Modifiable Area Unit Problem." *Concepts and Techniques in Modern Geography* 38: 1–41.

Purves, R., A. Edwardes, and J. Wood. 2011. "Describing Place through User Generated Content." *First Monday* 16 (9): 1–17. doi:10.5210/fm.v16i9.3710.

Rattenbury, T., N. Good, and M. Naaman. 2007. "Towards Automatic Extraction of Event and Place Semantics from Flickr Tags." *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information* Retrieval - *SIGIR'07*, 103–110. Amsterdam, Netherlands, July 23–27.

Shelton, T., A. Poorthuis, and M. Zook. 2015. "Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information." *Landscape and Urban Planning* 142: 198–211. doi:10.1016/j.landurbplan.2015.02.020.

Sigurbjörnsson, B., and R. Van Zwol. 2008. "Flickr Tag Recommendation Based on Collective Knowledge." *Proceeding of the 17th International Conference on World Wide Web –WWW'08*, 327–336. Beijing, China, April 21–25.

Smith, M., C. Szongott, B. Henne, and G. Von Voigt. 2012. "Big Data Privacy Issues in Public Social Media." The *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 1–6. Campione d'Italia, Italy, June 18–20.

Stevens, K., P. Kegelmeyer, D. Andrzejewski, and D. Buttler. 2012. "Exploring Topic Coherence over Many Models and Many Topics." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language* Processing *and* Computational Natural Language Learning, 952–961. Jeju Island, Korea, July 12–14.

Straumann, R. K., A. Çöltekin, and G. Andrienko. 2014. "Towards (Re)Constructing Narratives from Georeferenced Photographs through Visual Analytics." *The Cartographic Journal* 51 (2): 152–165. doi:10.1179/1743277414Y.0000000079.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (Sup1): 234–240. doi:10.2307/143141.

Vasardani, M., S. Timpf, S. Winter, and M. Tomko. 2013. "From Descriptions to Depictions: A Conceptual Framework." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, edited by T. Tenbrink, J. Stell, A. Galton, and Z. Wood, 8116 LNCS. 299–319. Berlin, Heidelberg: Springer-Verlag.

C

PUBLICATION III: STREETS OF LONDON: USING FLICKR AND OPENSTREETMAP TO BUILD AN INTERACTIVE IMAGE OF THE CITY

Bahrehdar, A. R., Adams, B., and Purves, R. S. (2019). Streets of London: Using Flickr and OpenStreetMap to build an interactive image of the city. Computers, Environment and Urban Systems (under review).

Streets of London: Using Flickr and OpenStreetMap to build an interactive image of the city

Azam Raha Bahrehdar[a], Benjamin Adams[b] and Ross S. Purves[a]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]Department of Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand

ARTICLE HISTORY
Compiled December 13, 2019

ABSTRACT
In his classic book "The Image of the City" Kevin Lynch used empirical work to show how different elements of the city were perceived: such as paths, landmarks, districts, edges, and nodes. Streets, by providing paths from which cities can be experienced, were argued to be one of the key elements of cities. Despite this long standing empirical basis, and the importance of Lynch's model in policy associated areas such as planning, work with user generated content has largely ignored these ideas. In this paper, we address this gap, using streets to aggregate filtered user generated content related to more than 1 million images and 60,000 individuals and explore similarity between more than 3,000 streets in London across three dimensions: user behaviour, time and semantics. To do our study we used two different sources of user generated content: (1) a collection of metadata attached to Flickr images and (2) street network of London from OpenStreetMap. Our approach allowed us to interactively explore patterns of similarity across multiple dimensions through an implemented Processing tool which allowed us to interactively explore these four dimensions simultaneously. Before drilling into the data to interpret in more detail, the identified patterns demonstrate that streets are natural units capturing perception of cities not only as paths but also through the emergence of other elements of the city proposed by Lynch including districts, landmarks and edges. Our approach also demonstrates how user generated content can be captured, allowing bottom-up perception from citizens to flow into a representation.

KEYWORDS
streets; Lynch; similarity; user generated content; perception

1. Introduction

The tale of Dick Whittington tells the story of a poor country boy who, enticed by rumours of streets paved with gold, makes his way to London. On his arrival he finds a busy, dirty city where his senses are assailed by sounds, smells and sights which are very different from those he had imagined. If the fictional Dick Whittington were alive today, he might use social media to take pictures of London and document some of the things he saw. By analysing not only what he photographed, but also comparing it to what others described, we could perhaps have a way of characterising London. But presumably Dick Whittington's descriptions would be rather different to those of

CONTACT Azam Raha Bahrehdar. Email: azam.bahrehdar@gmail.com

London's inhabitants, for whom the noise and bustle experienced by the country boy are simply background noise. And perhaps some locations, say a city market, would have distinct temporal signatures, reflecting how use of space varies over time. Other spaces might be preferred by locals, and not visited by Dick Whittington and other recent arrivals to the city. How we might extract and use such information to better understand how cities are perceived, by whom, and when is the subject of this paper.

The first question that we must answer in such an endeavour is what are the parts which come together in our perception of a city? Lynch, in his seminal book "The Image of the City" argued that cities are perceived through five elements: paths, nodes, districts, edges and landmarks. Paths, he claimed, are "channels along which the observer ... moves" and included, importantly for our work, streets which were for many people the "predominant elements" in their image of the city (Lynch 1960). This importance of paths or streets is widely recognised in urban planning – both in terms of their function in enabling mobility and as experienced public spaces (von Schnfeld and Bertolini 2017). Districts were described by Lynch as "the relatively large city areas which the observer can mentally go inside of, and which have some common character." Districts contain not only paths, but also salient landmarks in mental maps of the city. Nodes include locations linking paths (such as squares) which may ease orientation. Edges are linear physical or cultural divisions which are often borders between districts, or barriers to paths. Lynch's model is only one possible way of partitioning a city, but its relative simplicity, its empirical grounding, and its prominence in urban planning make it attractive (e.g., Hospers (2010) and Carmona et al. (2012)). The potential of streets as a way of immersing oneself in a virtual city—think of, for example, Google's Street View—is a further indicator that Lynch's paths are a logical group of elements through which to partition a city.

Given a set of objects to describe, a second key question is, with what? User generated content (UGC), in the form of images and their metadata have proven to be a tractable way of collecting perceptual information about locations which was previously the domain of empirical work (such as the questionnaires and interviews used by Lynch). The possibility of using such data at scale has spawned a vast literature exploring the utility of UGC as a data source. Particularly predominant with respect to the characterisation of cities has been research using Flickr, most likely due to relatively stable access, and the ability to query using a range of different dimensions including location. Early research demonstrated that clustering image locations and their descriptions could provide tractable ways of describing space (Rattenbury and Naaman 2009), and showed how perceptual theory (for example, with respect to the elements tagged) was replicated in such data (Tversky and Hemenway 1983; Rorissa 2008). The predominance of toponyms as tags led to a wide range of work focussing on the delineation of vague regions, analogous to the districts described by Lynch (Hollenstein and Purves 2010; Hobel, Fogliaroni, and Frank 2016; Gao et al. 2017). Using Flickr tags as ways of characterising cities and popular landmarks was quickly exploited (Crandall et al. 2009), essentially capturing cities at two contrasting granularities—as aggregate entities and through individual, highly popular cultural attractions analogous to Lynch's landmarks. Further analysis of image tags revealed that they captured not only visually perceived elements, but also allowed inferences to be made about sounds and smells (Quercia et al. 2015) in the city. Through these, and other qualities, it was possible to map potential preferences throughout a city, and thus recommend, for example, beautiful paths (Quercia, Schifanella, and Aiello 2014). However, despite the emphasis Lynch places on paths as key contributors to a city's image, very few UGC studies have focused on data aggregated at the granularity of

path-like features such as streets.

Identifying and characterising similar streets in a city has numerous applications. For example, it can be used, as suggested above, in route recommendation, or also more generally in recommender systems (Huang 2016; Quercia, Schifanella, and Aiello 2014). Similar streets may help to identify meaningful units at different scales, such as the districts proposed by Lynch, and gaps in similarity may suggest potential edges, or barriers of relevance in planning and tourism. By developing computational methods which capture such properties, we can develop tools which might help bridge gaps between quantitative and qualitative methods, by allowing researchers to explore a large space through methods akin to what is known in the digital humanities as *macroreading* before zooming in to apply more qualitative methods to, for instance, compare streets which appear semantically similar based on UGC. In this paper we take a first step towards these aims, demonstrating how, using streets as a proxy for Lynch's paths, we can use perceptive data in the form of image metadata to characterise and compare within a city, using London as an example.

We extend existing work by linking efforts to characterise cities using UGC with emerging computational approaches capturing Lynch's ideas. We focus on comparing paths, one of the most important elements identified by Lynch, with a further clear need coming from urban studies, and yet strangely neglected in many works focusing on UGC. We do so by considering three dimensions of paths: the users who visit them, the ways in which they describe them and the times at which they are visited.

Our contribution is thus threefold:

- We demonstrate how UGC can be linked to paths allowing us to create a computational version of how a city is perceived after appropriate data filtering.
- We show how paths can be compared and ranked according to three contrasting dimensions: their descriptions, the users who visit them and their temporal signatures.
- We explore how and why contrasting dimensions capture similarity by comparing and interpreting signatures.

2. Background

Describing cities, and capturing the properties which make particular places within cities more or less distinctive, is a key task if we are to effectively digitally represent cities (Miller and Small 1999). Rallying calls to consider place in geographic information science (e.g., Goodchild (2011) and Elwood, Goodchild, and Sui (2013)) have focused on the need to better capture shared, bottom-up representations of place, which go beyond categorisations of space derived from traditional, authoritative sources of spatial data. In particular, arguments for place-based representations often advance the idea of better representing varied human experiences of a location (Adams and McKenzie 2013; Jenkins et al. 2016), moving from the purely physical (e.g., park benches and bus stops) to, for example, emotions and behaviours associated with places (Shelton et al. 2014; Hauthal and Burghardt 2013; van Weerdenburg et al. 2019). The advent of social media, particularly user generated content, has provided an opportunity to capture human cognitive notion of place. While work like that of Montello (2003) focussed on capturing areas of interest through interviews, (Hu et al. 2015, p. 1) argued that such approaches are labor-intensive, time-consuming, and do not scale well.

A key reason for the emergence of research on computationally representing place

can therefore be linked to the data avalanche referred to by Miller (Miller 2010) with respect to the production of fine-grained data on urban spaces, and in particular rich UGC contributed by many individuals containing not only spatial information but also related temporal and semantic content. UGC, in different forms, has been used by many authors to characterise different dimensions of cities. One of the most prominent examples of such data are georeferenced microblog entries in Twitter. However, we note that though these data are suitable for exploring broad scale patterns of, for example language use or segregation in cities (Shelton et al. 2014), they have shortcomings with respect to fine-grained analysis (Lansley and Longley 2016). On the one hand, attempts to georeference the large proportion of Tweets not explicitly furnished with coordinates typically fail at fine resolutions except when matching to select sets of commercial points of interest (Zheng, Han, and Sun 2018), and, on the other hand, the content of a georeferenced Tweets was not always relevant to location (Hahmann, Purves, and Burghardt 2014). By contrast, image descriptions, uploaded to image sharing platforms, have a number of desirable properties. Firstly, early work demonstrated that image tags contained sufficient semantics to allow meaningful descriptions to be generated for locations (e.g., Rattenbury and Naaman (2009) and Crandall et al. (2009)). Secondly, image tags capture not only visually perceived elements, but also inherent qualities of places including affordances and perceptual properties (Dunkel 2015). Thirdly, since one reason why users tag images is to make them findable, image tags often reflect basic levels (Rorissa 2008)—they use shared terms which are both informative and succinct. Fourthly, data quality is good, such that image metadata containing coordinates are both accurate and precise, with caveats as to whether the location of the photographer or the subject are captured (Zielstra and Hochmair 2013; Hollenstein and Purves 2010), allowing extraction of spatial properties of individual landmarks (Crandall et al. 2009).

These data properties have led to multiple studies based around Flickr images, their locations and associated metadata including tags, timestamps and unique user identifiers. Early work transferred concepts from traditional information retrieval, such as term frequency-inverse document frequency weighting, to derive salient and distinctive descriptions for spatial regions (Kennedy et al. 2007). By analysing the locations of Flickr images and their tags, Crandall et al. (2009) showed that landmarks from different global cities could be extracted, and also demonstrated how the importance of salient landmarks in characterising different cities varied. Flickr also quickly proved to be an excellent source of information allowing vague places and vernacular names to be mapped at the city scale (e.g., Hollenstein and Purves (2010)). Understanding which parts of cities were visited, and in which order, requires that trajectories be built from images taken by individual users (Girardin et al. 2008). In large urban areas, simply distinguishing between 'locals' and 'tourists', based on the length of time an individual is present, proved to be a very effective way of describing use of space as shown by Eric Fischer (2013) in a set of impressive visualisations.[1] Straumann, Cöltekin, and Andrienko (2014) use temporal and user information to build trajectories and compare group behaviours and thus, they argue, explore narratives in the city. Feick and Robertson (2015) make two important observations—firstly, the semantics derived from georeferenced images is dependent on the scale of the analysis unit and, secondly, the distribution of images is strongly influenced by the street network (and open spaces). This second observation makes it all the more surprising in our view that most studies to date have linked urban properties derived from UGC with

[1] http://www.sightsmap.com/

space ignoring the underlying street network. Indeed, even work focusing on deriving "beautiful, quiet, and happy routes in the city" used a grid to characterise locations based on terms extracted from Flickr data and associated with, for example positive and negative emotions (Quercia, Schifanella, and Aiello 2014). Finally, we note that though many studies have characterised and compared regions or grid cells using UGC (e.g., Derungs and Purves (2016) and Gao et al. (2017)) a detailed exploration of the reasons for particular characterisations, or explanation of patterns of similarity is often lacking.

Any work using UGC should consider ways in which data quality can impact on interpretations of results. These include properties such as participation inequality, where a small proportion of users contribute a large volume of content (Van Mierlo 2014), uncertainties in positions or their interpretation (Stvilia and Jörgensen 2009), factors influencing semantics including ambiguity and automation (Varol et al. 2017) and underlying behavioural patterns (Sagl et al. 2012).

Returning to our starting point and aim—capturing the properties of cities in meaningful ways—we note that many authors have used Lynch's initial work to explain and justify the choice of UGC. Somewhat surprisingly, perhaps the most complete attempt to replicate the image of the city to date (Filomena, Verstegen, and Manley 2019) does so based mainly on administrative data (in the form of the road network and building footprints), replicating the potential to perceive through predominantly visual and structural indices. They did however use land use, as determined by OpenStreetMap contributors to capture some place-like properties mainly relating to affordances. Another example is work by Zhang et al. (2018) that used a collection of images annotated with outdoor objects. They used street network as "a major place for human mobility and activity" to capture and represent one aspect of a place: physical appearance. Others have used the street as a fundamental unit to explain place, for example in Massey's (1994) seminal work where Kilburn High Street served as an example for the complexity of place, and a more recent study by Capineri (2016) to explore the same street using Flickr photos.

3. Data and Methods

3.1. *Overview*

To characterise and compare street level similarity patterns we used two datasets: firstly, a selection of elements from the OpenStreetMap roads layer to characterise paths, and secondly, Flickr metadata capturing the locations, unique user identifiers (UUID), tags and times at which pictures were taken. Before calculating similarities we filtered data to remove biases, and identified relevant salient tags. We calculated similarity between street segments for three dimensions: semantics (based on patterns of tag usage), user behaviour (based on unique user identifiers) and temporal (based on times at which images were taken). We then mapped the most similar street segments and identified a range of characteristic similarity patterns, which we interpret based on the data contributing to the patterns of similarity.

3.2. *Modelling paths*

To model paths we downloaded the complete OpenStreetMap roads layer provided by Geofabrik[2] within 33 boroughs of Greater London. Geofabrik provides up-to-date packages of OpenStreetMap data for countries and regions. The initial network consisted of a set of ways (ordered sets of nodes) annotated with names, types and references to UK national road classes. We selected only major roads, using the classes primary, trunk and secondary to reduced the density of the overall network and retain important paths. We then removed pseudo-nodes from individual segments with the same name, type and class to form continuous segments where not split by road junctions. Finally, we retained all segments with lengths of more than 200m, resulting in the street network shown in Figure 1a. Note that this network is not topologically complete, since short segments were removed. Furthermore, some segments represent individual carriageways of the same street, where these have been digitised as separate segments (e.g., as is the case for expressways with separated lanes). After this process we were left with 3,406 unique segments with a median length of 519m.

3.3. *Path attributes*

We downloaded an initial Flickr dataset consisting of all georeferenced images available through the Flickr API for the bounding box of Greater London. We then selected only the images found within the polygon representing Greater London, associated with Flickr accuracy [sic] greater than 14. For each image we stored UUIDs, tags, image coordinates and the timestamp at which a photo was taken. Figure 1b shows the initial Flickr dataset described here.

Before associating images with street segments, we performed several filtering steps to remove biases typical to UGC and retain salient information. Firstly, we removed images (and users) associated with typical forms of *participation inequality*. We did so by a) removing all users who contributed only a single image (typically not representative tags), b) removing a single very prolific user who contributed some 5% of all images and c) retaining only one image in the case of bulk uploads (i.e., multiple images from one user with identical tags and coordinates). Doing so reduced our initial collection 5,119,629 images to 2,537,941 images, and the initial 105,021 users to 72,407 users. This filtered dataset, associated with 100m buffers around street segments, then formed the basis for calculating similarity according to users and time. After extracting only images and users found within the 100m buffers, we were left with a total of 1,250,205 images and 61,184 users.

Since we wished to calculate and interpret semantic similarity, we not only filtered noise from tags, but also selected semantically relevant terms which capture perceived properties of the city. To do so, we removed images with no tags, and tags using non-ASCII characters, duplicate tags in the same list and machine generated tags. We also removed images and tags shared through Instagram links, since we noted that the subjects of such images often had limited relationship to location. Furthermore, many tags used in Instagram relate to memes and filters which are highly ambiguous and biased results (e.g., a popular Instagram filter is called `earlybird`). In previous work Hollenstein and Purves (2010) showed that bias in the use of individual tags could be accounted for by the use of tag profile histograms. These allow us to remove tags with uneven patterns of use (e.g., those used only by a few prolific users). We removed tags

[2] https://www.geofabrik.de/

with a high coefficient of variation (>200) from our dataset.

Having filtered tags using these steps, we were left with a total vocabulary of unique 4,744 tags and 8,967,337 tags describing 1,726,670 images taken by 51,282 users. We then again filtered images to retain only those found within 100m buffers around street segments. To select the most representative tags from the remaining images, we used Latent Dirichlet Allocation (LDA) to perform topic modelling on a 500m grid (Blei, Ng, and Jordan 2003; Bahrehdar and Purves 2018). Briefly, topic modelling outputs for each tag the probability of it belonging to a particular topic (here a group of grid cells). Furthermore, for each topic tags are assigned in ranked order of probability. We retained all tags predicting 80% of the cumulative probability per topic, thus removing tags which provide limited information about specific locations. Having performed topic modelling, we found a mix of generic terms and proper nouns, in the form of places names, as would be expected from typical tagging behaviour (Sigurbjörnsson and van Zwol 2008). Since we did not wish to measure semantic similarity based on place names, but rather properties, we filtered place names from our tags using fuzzy matching on a set of place names extracted from GeoNames. Finally, we treated the remaining list of tags associated with grid cells as an *allow list* for segments passing through that grid cell. The final filtered dataset created by 36,486 users of 671,207 images described by 1,605 unique tags and 4,268,980 tags in total was then used to calculate semantic similarities.

Note that thus two datasets were used in our similarity calculations: one for temporal and user similarities where we did not filter based on tagging behaviour, and a more strongly filtered dataset for the calculation of semantic similarity.

3.4. *Measuring Similarities*

3.4.1. Semantic Similarity

We calculated semantic similarity by comparing tags used to describe segments. Each segment (S) is represented as a vector $V_S = [t_1^s, t_2^s, ..., t_n^s]$ where each element of the vector t_1^s corresponds to a tag's frequency in that segment, and n is the number of unique tags. Since raw counts are biased towards tags which are frequent across London as a whole, we calculated a normalised spatial TF-IDF to increase the weight of tags common in particular segments, but rarer as a whole as follows:

$$tf.idf_{(t_i,s_j)} = ntf_{(t_i,s_j)}.idf_{t_i} \tag{1}$$

where ntf is the number of times a term (t_i) was used associated with a segment (s_j) and was normalised based on the number of total terms associated with the segment, and idf was calculated as follows:

$$idf_{t_i} = 1 + Log_e(\frac{N}{sf_{t_i}}) \tag{2}$$

N is the total number of segments and sf_{t_i} or $segmentfrequency$ is the number of segments with term (t_i) in it.

Similarity between segment pairs was calculated using cosine similarity for the weighted TF-IDF vectors as the dot product of two vectors:

$$\text{Sim}(v_{s_1}, v_{s_2}) = \cos(\theta) = \frac{v_{s_1} \cdot v_{s_2}}{\|v_{s_1}\| \, \|v_{s_2}\|} \tag{3}$$

Similarity values of 1 indicate that the semantics of two segments are identical, while values of 0 indicate complete dissimilarity.

3.4.2. User Similarity

To compare how similar two segments are in term of unique users (who have photos associated with segments), we again used cosine similarity. Here, however, we represented each segment as a binary vector, containing either a) all users found in London or b) only those who took images within the segments over two weeks or less. We treated this second group as tourists (c.f. Girardin et al. (2008); Straumann, Cöltekin, and Andrienko (2014)).

3.4.3. Temporal Similarity

Our third similarity dimension was based on the temporal distribution of images associated with a segment. We chose to compare segments according to the proportion of visits on different days of the week (c.f. McKenzie et al. (2015)) after experimenting with hours of the day and months of the year. We calculated temporal similarity as the Euclidean distance between a seven-dimensional vector, where we treated the proportion of images taken on each day of the week as an independent dimension.

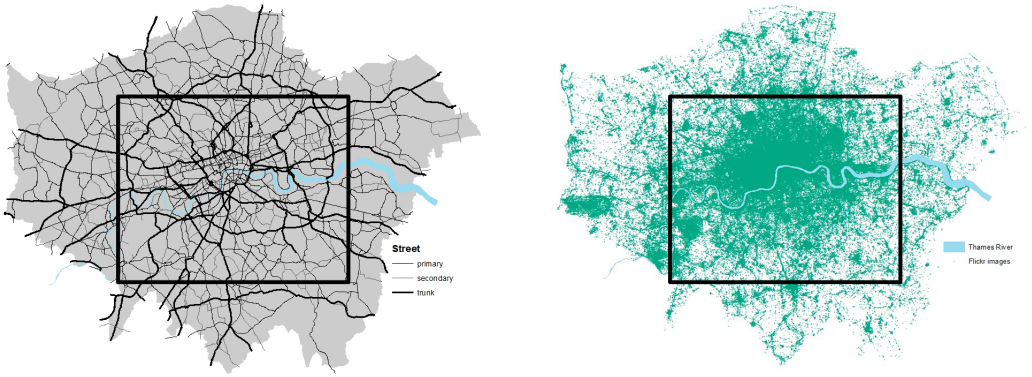
4. Results

Our analysis was based on the metadata derived from 5,119,629 geo-tagged Flickr images from Greater London downloaded in September 2018 using the Flickr API. Figure 1.a and Figure 1.b show the study area and the segments with which images were associated and the locations of the images analysed. They demonstrate that Flickr images are commonly associated with the network structure of the street network, as represented by our model, but also show concentrations in open spaces (which are not captured). The two filtered datasets are of different sizes: more extreme filtering is required to capture semantics, and the initial 5,119,629 images are reduced to around 1,250,205 images and 61,184 for calculations of temporal and user similarity, and 671,207 images and 36,486 users for semantic similarity.

To explore patterns of similarity, we implemented a Processing tool which allowed us to interactively explore four dimensions simultaneously: semantic, users (all and tourists) and temporal. This tool is available online[3] and it is important to note that the following examples were identified through its use. We describe and interpret the properties of four locations, selected because of their contrasting properties and efficacy in illustrating differing aspects of our approach.

The first example is that of a very well-known London location, Tower Bridge (Figure 2). As for each of the following examples, we present four maps of correlations between segments, a tag cloud illustrating the segments shared by at least twelve of the thirty most similar segments to Tower Bridge, and a histogram showing the ten most similar segments in terms of proportions of images taken on different days of the week. The most semantically similar segments to Tower Bridge form a sinuous path along the banks of the Thames, linked by many of its bridges. These give the appearance of forming a path through the city *sensu* Lynch, and exploring the tag cloud reveals that the Thames Path (and Thames River) are indeed tags shared by

[3]Download a zip file https://www.dropbox.com/s/q2mpr3iczkx1x3i/users_cosineSimilarity_binary.zip?dl=0



(a) Street network of major roads longer than 200m

(b) UGC footprints follow street network and open spaces

Figure 1. Study area within 33 boroughs of Greater London

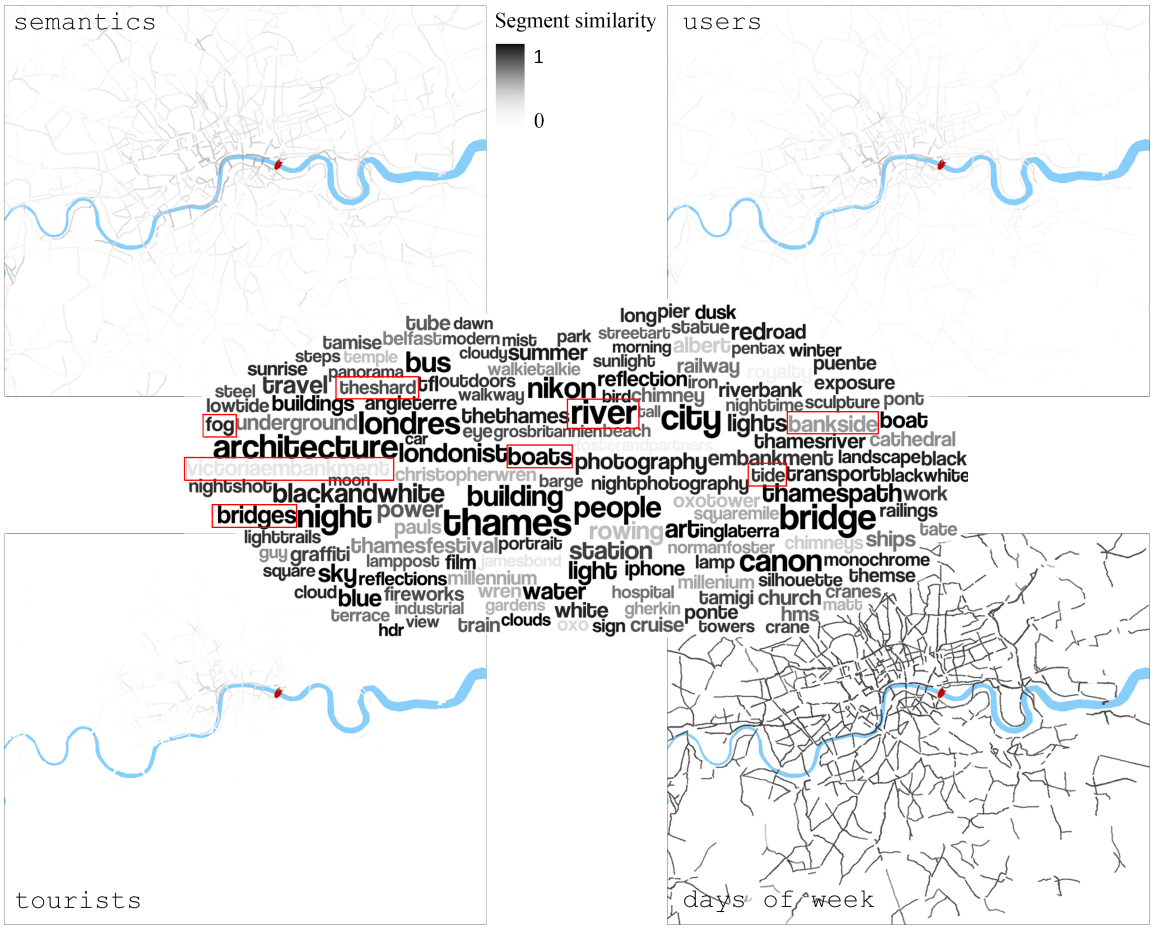


Figure 2. Signature similarities for Tower Bridge: Each of four maps represents the similarity between the queried street in red and all other streets in London. Darker segments are the more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Larger tags are more important, and darker tags are shared by more of the top 30 segments. Tags highlighted in red are discussed in the text.

many of the most similar segments. Many other tags reveal different aspects of this location such as its `bridges`, `boats`, `tides` and the `river` itself. Some more specific tags, for example, `victoriaembankment`, `theshard`, and `bankside` refer to named locations found along the Thames which were not removed by our toponym filtering. Various image properties, some more likely to be related with water (e.g., `reflection` and `fog`) are found, together with a host of photography related terms which could arguably have been filtered (e.g., `canon`, `blackandwhite`, `nightshot`). Nonetheless, our semantic similarity measure both reveals a district, which can also be interpreted as a path, and allows us to interpret it in a meaningful way. Maps of users and tourists show (and consistently in all examples) weaker correlations. Users in general are clustered around Tower Bridge, with a bias to the west, and north of the river, though some users do cross to the south of the Thames. By contrast, tourists are found in a smaller region, almost only in Central London either near, or to the north of the river. These maps indicate the effectiveness of the Thames as a barrier to people, with users much less likely to visit seemingly similar regions (as defined through our semantics). Temporally, we note that correlations for many segments are high, and see little if any spatial pattern. The associated histogram (Figure 6) reveals that, on average, more pictures are taken on Saturdays and Sundays than other days of the week at Tower Bridge and similar segments. However, this behaviour is in fact typical of Flickr usage in our study area as a whole, which in turn explains the limited spatial pattern revealed by this dimension.

Our second example, Chepstow Road (Figure 3) reveals a similarly strong spatial pattern of correlated segments, picking out a very small district around Notting Hill. This is in fact the location of the annual Notting Hill Carnival (tagged as `nottinghillcarnival`), and rather than identifying a district through an affordance (e.g., the banks of the Thames and the Thames Path), here semantic similarity reveals an event. The semantics of the tag cloud confirm this, with shared tags including `carnival`, `dancing`, `parade`, `party` and so on. The pattern of user correlations is more spatially extensive than that for Tower Bridge, revealing that the community visiting this location roams further than that photographing the tourist site of Tower Bridge. Tourists however, appear to share almost no segments in common. Temporal correlations are in general low, and as revealed by the histograms Figure 6 this relates to the taking of photographs on Sundays and Mondays as opposed to other days of the week. Sunday and Monday are in fact the two days of the Notting Hill Carnival, and an inspection of related images revealed that these do indeed reflect the dates of the parade itself.

The third example, Whitehall, lies in the heart of London, and is associated with both political and ceremonial events (Figure 4). Semantically, we can pick out a region around Central London, spanning both sides of the Thames. We note, as was the case for Tower Bridge, a range of tags related to photography and named locations in this region (e.g., `oxfordst`, `stjamespark` and `hydeparkcorner`). Many tags reflect the usage of this part of London, reflecting recurring and rare events (e.g., `celebration`, `royalwedding`, `parade`, `protest`) and their participants (e.g., `soldier`, `queen`, `guards`). The users photographing this segment again capture larger areas than those visiting Tower Bridge, with tourists once more focusing on locations north of the river, and the Thames acting as barrier to movement south. Temporally, Whitehall follows the general pattern of all locations except for Chepstow Road, with most pictures taken at the weekend on Saturday and Sunday.

Our final example (Figure 5), Crystal Palace Parade, reveals a very different pattern to the previous three, all of which allowed us to identify coherent regions associated

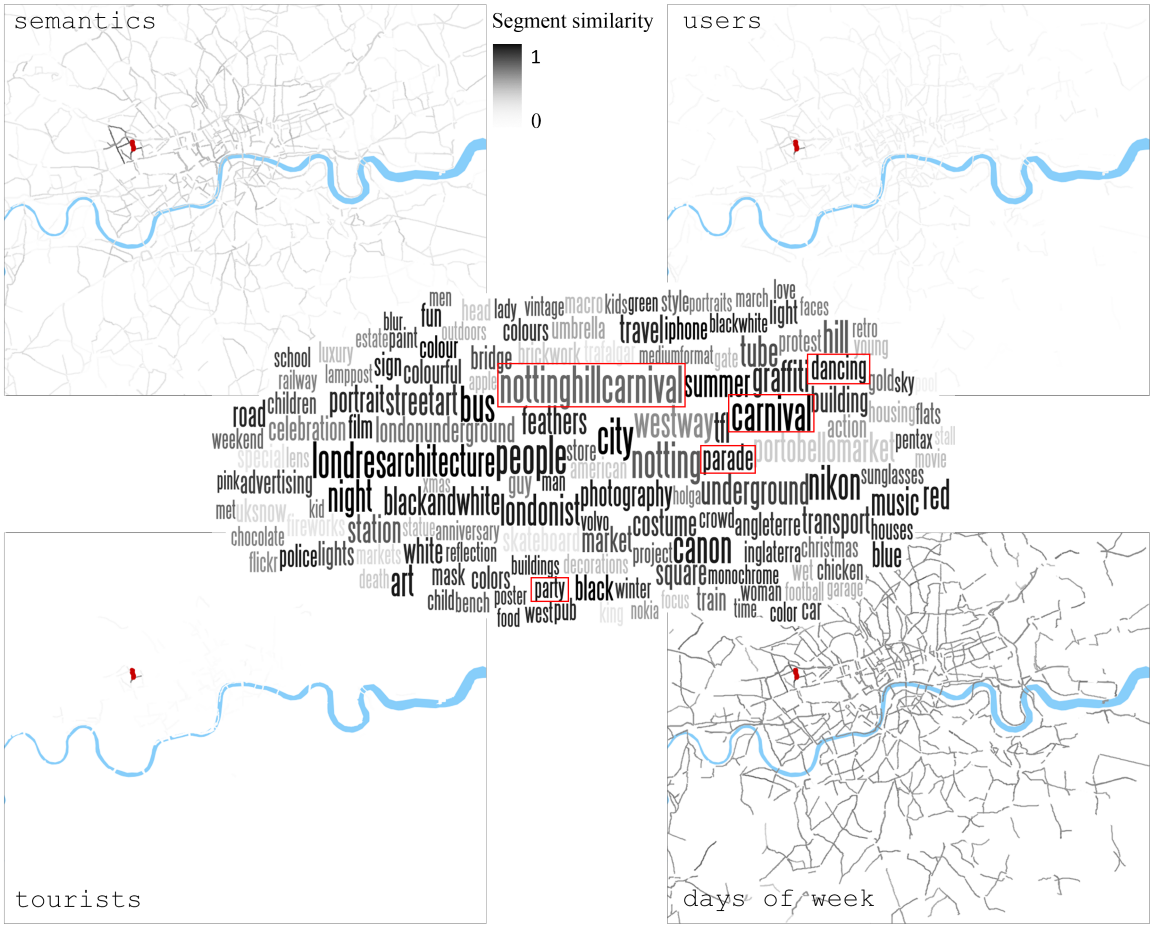


Figure 3. Signature similarities for Chepstow Road: Each of four maps represents the similarity between the queried street in red and all other streets in London. Darker segments are the more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Larger tags are more important, and darker tags are shared by more of the top 30 segments. Tags highlighted in red are discussed in the text.

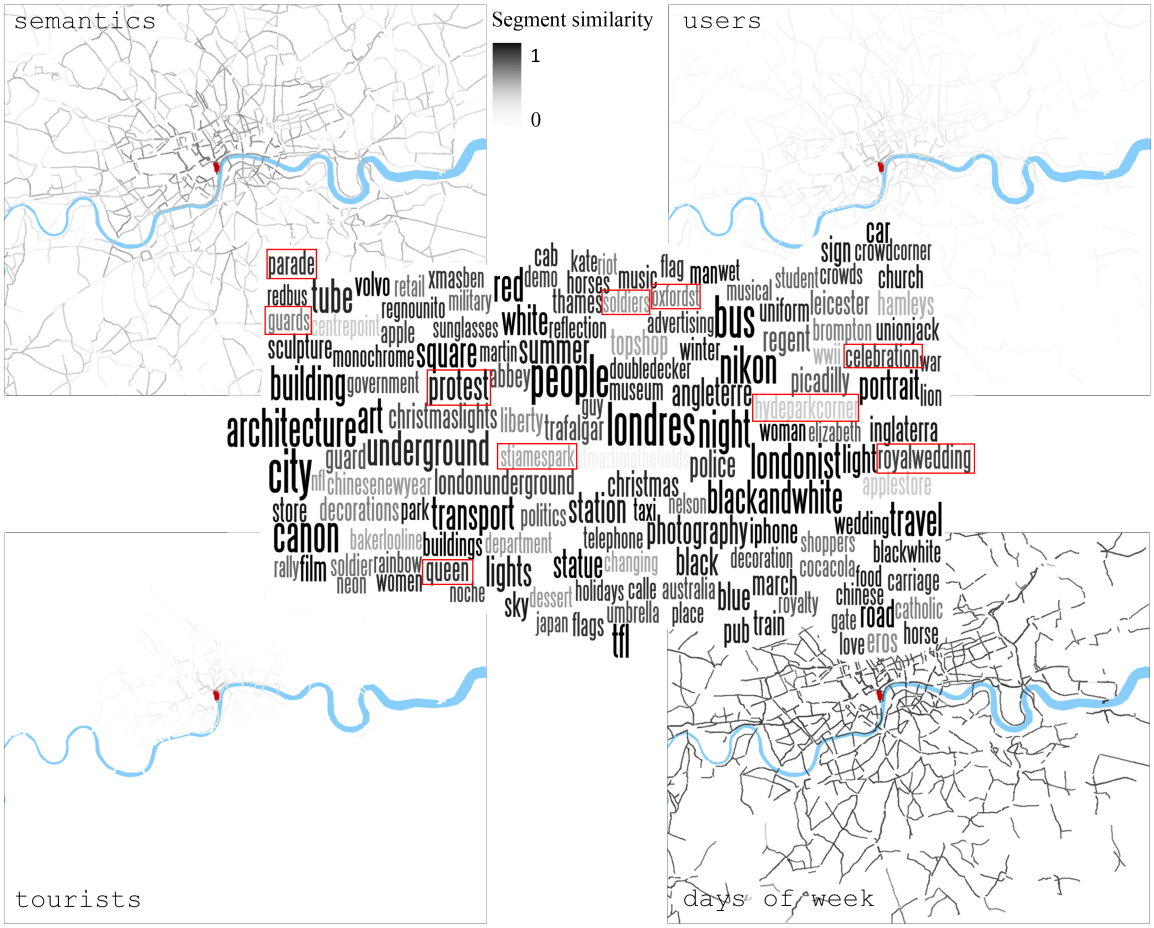


Figure 4. Signature similarities for Whitehall: Each of four maps represents the similarity between the queried street in red and all other streets in London. Darker segments are the more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Larger tags are more important, and darker tags are shared by more of the top 30 segments. Tags highlighted in red are discussed in the text.

Figure 5. Signature similarities for Crystal Palace Parade: Each of four maps represents the similarity between the queried street in red and all other streets in London. Darker segments are the more similar. The word cloud presents more details on shared semantics of the 30 most similar segments related to the first map. Larger tags are more important, and darker tags are shared by more of the top 30 segments. Tags highlighted in red are discussed in the text.

with semantically similar segments. In this case, these segments are distributed, seemingly randomly, across all of London. However, the tag cloud associated with the most similar segments reveals the reason for this pattern. Other than common tags related to photography, we find here many tags related to transport including `bus`, types of bus (e.g., `scania`, `plaxton`, `routemaster`, `mercedes`, `volvo`) and providers of public transport (e.g., `arriva`, `londontransport`, `stagecoach`, `abellio`). Semantic similarity with this location is thus defined by photographs of a particular type, taken by a specialist group interested in public transport. Users present at this location spread not only over south London, but into north London as well, demonstrating an asymmetry in the barrier effect of the Thames apparently limiting movement from north to south more than south to north. Note that since semantic similarity and user similarity have very different patterns, that our method implicitly shows that different photographers are interested in the same subject matter. Tourists taking pictures at this location appear to be rare, and thus have a very limited local spatial spread. Temporally, we note similar patterns to Tower Bridge and Whitehall, though with a noticeable secondary peak midweek.

Having explored these individual examples, the obvious question which arises is, how can we interpret these results more generally, and can the results be linked to the ideas posed by Lynch? With respect to the former question, we note that by using three distinct dimensions (semantics, users and time) different patterns are revealed. The patterns associated with users form regions or districts *sensu* Lynch clustered around the query segment in all cases, though the forms of these regions are not always symmetrical. Thus, for Tower Bridge we note a general tendency to locations north of the river, revealing the Thames influence as a barrier, or in Lynch's terms an edge. However, in the one location south of the river in our examples (Crystal Palace) this barrier is less influential, revealing a different pattern of user behaviour—users south of the river appear less influenced by the Thames as a barrier or edge than those to the north. When selecting out tourists alone, based on their length of stay in London, we find meaningful signatures (which largely replicate the pattern of users in general) only at very popular sites (e.g., Tower Bridge and Whitehall). Our semantic signatures are interesting in a number of different ways. Firstly, they reveal not only where similar aspects of a scene were annotated, but also what was of interest. These include named locations (landmarks *sensu* Lynch) as well as objects commonly found in scenes and properties of scenes. Each example has quite different semantic properties, and the form of the districts associated with similar semantics vary from the linear path through London generated by the Thames and the Thames Path for Tower Bridge, through Central London as a whole associated with Whitehall, to the very small region related to the Notting Hill Carnival for Chepstow Road, and finally the dispersed locations associated with public transport for Crystal Palace, where no meaningful district emerges. In this last case, we suggest these are locations frequented by enthusiasts where this type of photography dominates. Temporally, our method struggles to identify similar regions since the overall distribution of Flickr images (Figure 6) is very similar to that of three of our four exemplars, and thus temporal correlations are generally very high. Only for Cheptstow Road, where an annual event dominates, do we find a meaningful difference from this general pattern of picture taking, with lower average temporal correlations, and a similar set of segments emerging around the location of the event itself.

In Lynch's terms, exploring different dimensions of our data allows districts to emerge, which represent both either coherent areas of semantic similarity, or dispersed locations with a shared identity captured through semantics. These districts may be

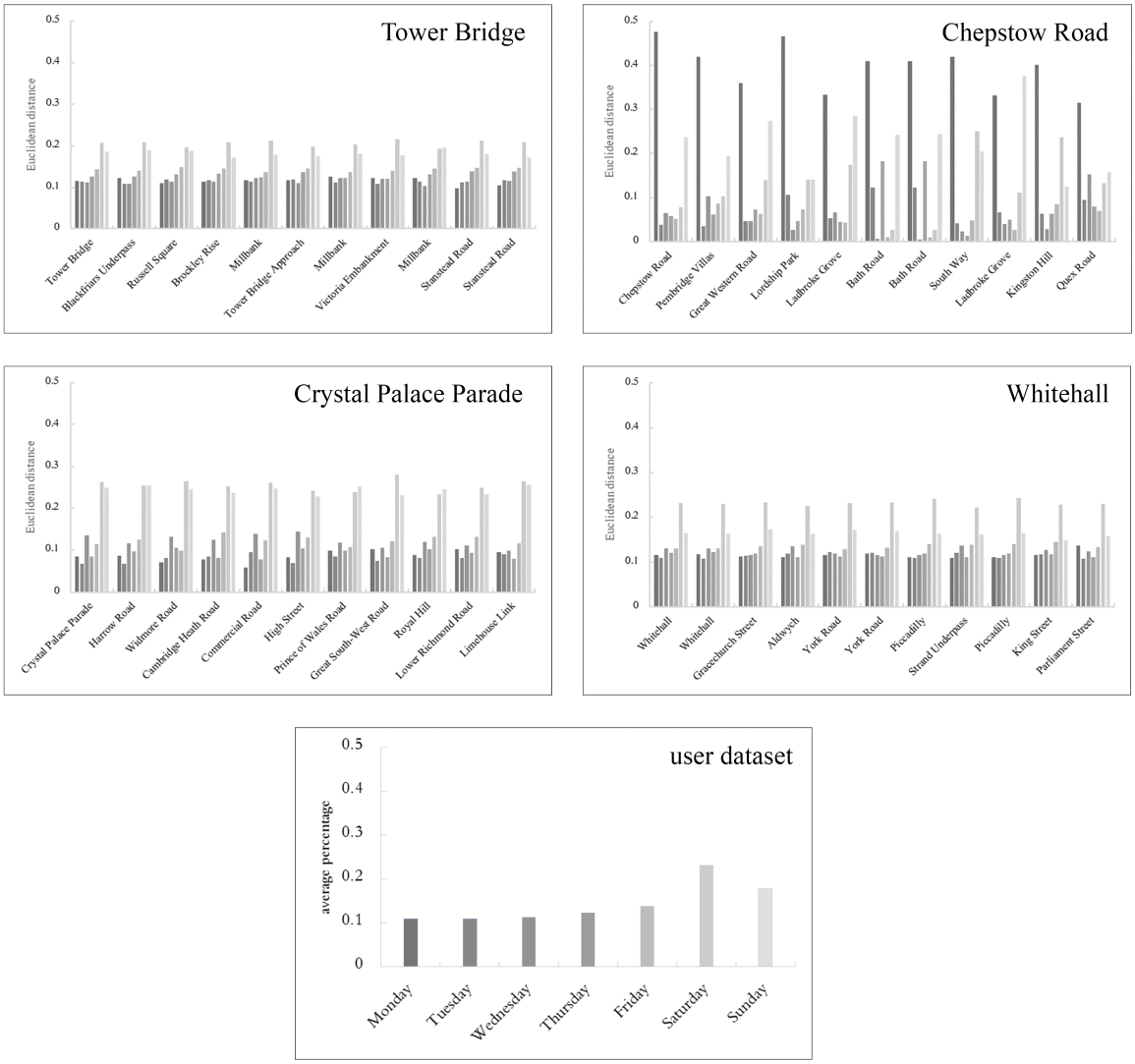


Figure 6. Histogram of daily images taken for the ten most similar segments to query segment as captured by temporal similarity for each of the four examples, and the overall temporal distribution of images in the collection

similar for the same segment across differing dimensions (c.f. Whitehall) or differ significantly (c.f. Tower Bridge or Crystal Palace). Within the word clouds capturing semantic similarity we can identify named places, which can be seen as proxies for landmarks, and interestingly, concepts which appear to give otherwise less salient locations some identity (e.g., the buses of Crystal Palace). The Thames emerges both as a path (in the case of Tower Bridge) and an edge (again for Tower Bridge, but also Whitehall) dividing the city in two.

5. Discussion

We set out to develop a tool which would allow us to capture information about how a city was perceived through the properties of an element identified by Lynch as central to our understanding—paths through the city. By associating UGC, in the form of tags, timestamps and UUIDs with street segments, we were able to interactively explore similarity between segments across three, contrasting, dimensions. In the following, we firstly discuss key influences on, and limitations of, our approach, and discuss it with respect to previous work, before setting out our contribution in the broader context of practice.

The first, and most important influence concerns the filtering of our data. We chose to filter based both on behaviour (e.g., taking account of participation inequality, bulk uploads and so on), semantic biases (primarily seeking to retain only tags used by a broad group of users) and identify semantically distinctive terms (through topic modelling). These choices mean that we explore the temporal and user dimensions with different data sets to the semantic one; however, we argue that knowingly making these choices is a valid approach. Although filtering is often left implicit, or only briefly discussed, in our case these choices reduced the original dataset five-fold. We think the implications and importance of filtering have been neglected in the gold-rush mentality of analysis of UGC, and believe that the attention now being paid to bias in data in artificial intelligence tasks (Zou and Schiebinger 2018) is equally important here (Shelton et al. 2014).

In linking tags to segments we chose a buffer width of 100m; changing this width would also reduce or increase the number of image locations associated with segments. Increasing buffer size would however reduce distinctiveness of tags, since they would be associated with multiple segments, while smaller buffers would lead to a very limited set of tags for less well-covered regions outside of Central London. We explored the overlap of tags between very similar segments, and found that, for example, for the ten most temporally similar segments to Tower Bridge only one shared images, and for Chepstow Road only two from the ten most temporally similar were shared. As well as choices in the filtering of image metadata, we also filtered geometry. By removing short segments (<200m), we removed nodes from the network that were potentially densely photographed (e.g., Trafalgar Square), which were found by others to be some of the most photographed locations in London (Crandall et al. 2009). Furthermore, by only using main roads, we removed paths through some important areas of open space, such as Hyde Park, again limiting our sample of image metadata.

We note that the behaviour of individuals taking photographs is an important source of bias in our work. This manifests itself in multiple ways. For example, temporal signatures are dominated by increased activity at the weekend, related to leisure activity, but showing little variation in space. The use of platforms and hashtags through platforms like Instagram can result in biases both in terms of what is photographed and

the semantics used to describe images. Furthermore, as pointed out by Boy and Uitermark (2017) in their study of Instagram, we run the risk of capturing "an image of the city that is sanitised and nearly devoid of negativity."

Having made these choices, we then calculated similarity values for a total of 3,406 segments resulting in a matrix of about 5,800,000 unique correlation pairs. To explore these correlations, we implemented an interactive visualisation tool, which allowed us to explore patterns of similarity in space and across our three dimensions. This tool allowed us to quickly and easily identify potentially interesting patterns, but still required us to interpret these patterns. We did so in three distinct ways. Firstly, to interpret patterns of similarity we drilled down into data (following the ideas of the visual analytics mantra introduced by Keim et al. (2008)) to show detail, by either showing tags related to the most similar segments (c.f. Rattenbury and Naaman (2009) or histograms of temporal behaviour (c.f. Lansley and Longley (2016)). Only by exploring these details could we meaningfully interpret our results. Secondly, general knowledge about London, for example the relationship between Whitehall and ceremonial events was important in interpreting semantics and suggesting potential themes for exploration such as the effect of the Thames as a barrier to north-south movement. Thirdly, we supplemented this knowledge with research, to for example identify the potential relationship between Cheptstow Road and the Notting Hill Carnival.

Our approach allows us to go further than previous work in capturing ways in which individuals characterise, and thus we assert, perceive, the city. By using three complementary dimensions linked to paths, we can not only find similar regions, but describe their properties and link these to behaviour in the city itself. However, it is important to note that the missing parts of the city, where we find no data, are potentially just as important in understanding how the city is perceived by its inhabitants, and our approach, and other focussing on passively crowdsourced data cannot address this gap alone. Active approaches such as those proposed by the mappiness app (Seresinhe et al. 2019) may go some way to filling this hole, but the importance of such data gaps cannot be overstated (Graham et al. 2014). Nonetheless, our approach starts to suggest how Lynch's ideas can be empirically implemented at scale.

According to Lynch ([p. 8]1960), a workable image of a city requires three important elements such as identity (in the sense of identification of urban elements), structure (indicating spatial or pattern relation among urban elements, for example, in a street network), and meaning (either practical or emotional meaning for an observer). By capturing multiple dimensions of similarity, and linking these to paths through the city, analysis not only of space, but place is enabled, and in doing so important relationships between locations are revealed. Our approach allows, in principle, exploration across time steps, and thus is temporally dynamic, and synthesises heterogeneous data. The tool is easy (and we think fun!) to use, and interactivity enhances exploration. We note that our dimensions could also be combined, exploring for example semantic similarity at particular times, or for particular user groups, though doing so would require that the same filtering approach was taken with all data.

6. Conclusion and Future Work

Starting with Dick Whittington's confusion when confronted with a London very different to the one he had heard off, we set out to model the characteristics and thus similarity between streets in London using UGC. Streets are a natural unit, since they capture the paths described by Lynch, and our results demonstrate how they allow us

to explore perceptions in terms of not only paths through the city, but through the emergence of districts, landmarks and even edges. These elements emerge because we combine different dimensions capturing semantic similarity, user behaviour and temporal patterns. Street segments are a more natural way of organising our data, and reduce the issues caused by aggregating across administrative boundaries or arbitrarily imposed tessellations such as grids. We demonstrate that the data found in London are sufficiently rich, despite numerous filtering steps, to reveal interesting and meaningful patterns, though interpretation of these requires us to both drill down into the data and use external knowledge.

We suggest that future work aiming to use UGC in planning or applications such as location based services consider how such data can be effectively integrated, while not forgetting the implications of data bias and gaps.

Acknowledgement(s)

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Schweizerischer Nationalfonds zur Frderung der Wissenschaftlichen Forschung [200021_149823] and the New Zealand Building Better Homes, Towns and Cities National Science Challenge.

Notes on contributor(s)

References

Adams, Benjamin, and Grant McKenzie. 2013. "Inferring thematic places from spatially referenced natural language descriptions." In *Crowdsourcing geographic knowledge*, 201–221. Springer.

Bahrehdar, Azam R, and Ross S Purves. 2018. "Description and characterization of place properties using topic modeling on georeferenced tags." *Geo-spatial Information Science* 21 (3): 173–184.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet allocation." *Journal of machine Learning research* 3: 993–1022.

Boy, John D, and Justus Uitermark. 2017. "Reassembling the city through Instagram." *Transactions of the Institute of British Geographers* 42 (4): 612–624.

Capineri, Cristina. 2016. "Kilburn high road revisited." *Urban Planning* 1 (2): 128–140.

Carmona, Matthew, Tim Heath, Taner Oc, and Steve Tiesdell. 2012. *Public places-Urban spaces*. Routledge.

Crandall, David J., Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. 2009. "Mapping the World's Photos." In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, New York, NY, USA, 761–770. ACM.

Derungs, Curdin, and Ross S Purves. 2016. "Characterising landscape variation through spatial folksonomies." *Applied Geography* 75: 60–70.

Dunkel, Alexander. 2015. "Visualizing the perceived environment using crowdsourced photo geodata." *Landscape and Urban Planning* 142: 173–186.

Elwood, Sarah, Michael F Goodchild, and Daniel Sui. 2013. "Prospects for VGI research and the emerging fourth paradigm." In *Crowdsourcing geographic knowledge*, 361–375. Springer.

Feick, Rob, and Colin Robertson. 2015. "A multi-scale approach to exploring urban places in geotagged photographs." *Computers, Environment and Urban Systems* 53: 96–109. Special Issue on Volunteered Geographic Information.

Filomena, Gabriele, Judith A Verstegen, and Ed Manley. 2019. "A computational approach to The Image of the City." *Cities* 89: 14–25.

Fischer, Eric. 2013. "The geotaggers world atlas." .

Gao, Song, Krzysztof Janowicz, Daniel R Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. 2017. "A data-synthesis-driven method for detecting and extracting vague cognitive regions." *International Journal of Geographical Information Science* 31 (6): 1245–1271.

Girardin, Fabien, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. 2008. "Digital footprinting: Uncovering tourists with user-generated content." *IEEE Pervasive computing* 7 (4): 36–43.

Goodchild, Michael F. 2011. "Formalizing place in geographic information systems." In *Communities, neighborhoods, and health*, 21–33. Springer.

Graham, Mark, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. 2014. "Uneven geographies of user-generated information: patterns of increasing informational poverty." *Annals of the Association of American Geographers* 104 (4): 746–764.

Hahmann, Stefan, Ross Purves, and Dirk Burghardt. 2014. "Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes." *Journal of Spatial Information Science* 2014 (9): 1–36.

Hauthal, Eva, and Dirk Burghardt. 2013. "Detection, analysis and visualisation of georeferenced emotions." In *Proceedings of the 26th International Cartographic Conference*, .

Hobel, Heidelinde, Paolo Fogliaroni, and Andrew U Frank. 2016. "Deriving the geographic footprint of cognitive regions." In *Geospatial Data in a Changing World*, 67–84. Springer.

Hollenstein, Livia, and Ross Purves. 2010. "Exploring place through user-generated content: Using Flickr tags to describe city cores." *Journal of Spatial Information Science* 2010 (1): 21–48.

Hospers, Gert-Jan. 2010. "Lynch's The Image of the City after 50 Years: City Marketing Lessons from an Urban Planning Classic." *European Planning Studies* 18 (12): 2073–2081.

Hu, Yingjie, Song Gao, Krzysztof Janowicz, Bailang Yu, Wenwen Li, and Sathya Prasad. 2015. "Extracting and understanding urban areas of interest using geotagged photos." *Computers, Environment and Urban Systems* 54: 240–254.

Huang, Haosheng. 2016. "Context-aware location recommendation using geotagged photos in social media." *ISPRS International Journal of Geo-Information* 5 (11): 195.

Jenkins, Andrew, Arie Croitoru, Andrew T Crooks, and Anthony Stefanidis. 2016. "Crowdsourcing a collective sense of place." *PloS one* 11 (4): e0152932.

Keim, Daniel A, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. 2008. "Visual analytics: Scope and challenges." In *Visual data mining*, 76–90. Springer.

Kennedy, Lyndon, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. 2007. "How flickr helps us make sense of the world: context and content in community-contributed media collections." In *Proceedings of the 15th ACM international conference on Multimedia*, 631–640. ACM.

Lansley, Guy, and Paul A Longley. 2016. "The geography of Twitter topics in London." *Computers, Environment and Urban Systems* 58: 85–96.

Lynch, Kevin. 1960. *The image of the city*. Vol. 11. MIT press.

Massey, Doreen. 1994. "A Global Sense of Place." In *Space, Place, and Gender*, 146–156. University of Minnesota Press.

McKenzie, Grant, Krzysztof Janowicz, Song Gao, and Li Gong. 2015. "How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest." *Computers, Environment and Urban Systems* 54: 336–346.

Miller, Harvey J. 2010. "The data avalanche is here. Shouldnt we be digging?" *Journal of*

Regional Science 50 (1): 181–201.

Miller, RB, and C Small. 1999. "Digital cities. I. Integrating data and information resources, towards digital Earth." In *Proceedings of the International Symposium on Digital Earth. Science Press, Beijing*, 217–222.

Montello, Daniel R. 2003. "Regions in geography: Process and content." *Foundations of geographic information science* 173–189.

Quercia, Daniele, Rossano Schifanella, and Luca Maria Aiello. 2014. "The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City." In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, New York, NY, USA, 116–125. ACM.

Quercia, Daniele, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. 2015. "Smelly maps: the digital life of urban smellscapes." *arXiv preprint arXiv:1505.06851* .

Rattenbury, Tye, and Mor Naaman. 2009. "Methods for Extracting Place Semantics from Flickr Tags." *ACM Trans. Web* 3 (1): 1–30.

Rorissa, Abebe. 2008. "User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory." *Information Processing & Management* 44 (5): 1741–1753.

Sagl, Günther, Bernd Resch, Bartosz Hawelka, and Euro Beinat. 2012. "From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments." In *Proceedings of the GI-Forum*, 54–63. Herbert Wichmann Verlag Berlin.

Seresinhe, Chanuki Illushka, Tobias Preis, George MacKerron, and Helen Susannah Moat. 2019. "Happiness is Greater in More Scenic Locations." *Scientific reports* 9 (1): 4498.

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. 2014. "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of big data." *Geoforum* 52: 167–179.

Sigurbjörnsson, Börkur, and Roelof van Zwol. 2008. "Flickr Tag Recommendation Based on Collective Knowledge." In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, New York, NY, USA, 327–336. ACM. `http://doi.acm.org/10.1145/1367497.1367542`.

Straumann, Ralph K, Arzu Cöltekin, and Gennady Andrienko. 2014. "Towards (re) constructing narratives from georeferenced photographs through visual analytics." *The Cartographic Journal* 51 (2): 152–165.

Stvilia, Besiki, and Corinne Jörgensen. 2009. "User-generated collection-level metadata in an online photo-sharing system." *Library & Information Science Research* 31 (1): 54–65.

Tversky, Barbara, and Kathleen Hemenway. 1983. "Categories of environmental scenes." *Cognitive Psychology* 15 (1): 121–149.

Van Mierlo, Trevor. 2014. "The 1% rule in four digital health social networks: an observational study." *Journal of medical Internet research* 16 (2).

van Weerdenburg, Demi, Simon Scheider, Benjamin Adams, Bas Spierings, and Egbert van der Zee. 2019. "Where to go and what to do: Extracting leisure activity potentials from Web data on urban space." *Computers, Environment and Urban Systems* 73: 143–156.

Varol, Onur, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. "Online Human-Bot Interactions: Detection, Estimation, and Characterization." 280–289.

von Schnfeld, Kim Carlotta, and Luca Bertolini. 2017. "Urban streets: Epitomes of planning challenges and opportunities at the interface of public space and mobility." *Cities* 68: 48–55.

Zhang, Fan, Ding Zhang, Yu Liu, and Hui Lin. 2018. "Representing place locales using scene elements." *Computers, Environment and Urban Systems* 71: 153–164.

Zheng, Xin, Jialong Han, and Aixin Sun. 2018. "A survey of location prediction on twitter." *IEEE Transactions on Knowledge and Data Engineering* 30 (9): 1652–1671.

Zielstra, Dennis, and Hartwig H. Hochmair. 2013. "Positional accuracy analysis of Flickr and Panoramio images for selected world regions." *Journal of Spatial Science* 58 (2): 251–273.

Zou, J, and L Schiebinger. 2018. "AI can be sexist and racist-it's time to make it fair." *Nature* 559 (7714): 324.