



# Crowdsourced Data Mining for Urban Activity: Review of Data Sources, Applications, and Methods

Haifeng Niu<sup>1</sup> and Elisabete A. Silva<sup>2</sup>

**Abstract:** The penetration of devices integrated with location-based services and internet services has generated massive data about the everyday life of citizens and tracked their activities happening in cities. Crowdsourced data, such as social media data, points of interest (POIs) data, and collaborative websites, generated by the crowd, have become fine-grained proxy data of urban activity and widely used in research in urban studies. However, due to the heterogeneity of data types of crowdsourced data and the limitation of previous studies mainly focusing on a specific application, a systematic review of crowdsourced data mining for urban activity is still lacking. In order to fill the gap, this paper conducts a literature search in the Web of Science database, selecting 226 highly related papers published between 2013 and 2019. Based on these papers, the review first conducts a bibliometric analysis identifying underpinning domains, pivot scholars, and papers around this topic. The review also synthesizes previous research into three parts: main applications of different data sources and data fusion; application of spatial analysis in mobility patterns, functional areas, and event detection; and application of sociodemographic and perception analysis in city attractiveness, demographic characteristics, and sentiment analysis. The challenges of this type of data are also discussed in the end. This study provides a systematic and current review for both researchers and practitioners interested in the applications of crowdsourced data mining for urban activity. **DOI:** [10.1061/\(ASCE\)UP.1943-5444.0000566](https://doi.org/10.1061/(ASCE)UP.1943-5444.0000566). This work is made available under the terms of the Creative Commons Attribution 4.0 International license, <http://creativecommons.org/licenses/by/4.0/>.

**Author keywords:** Crowdsourced data; Data mining; Urban activity; Review of methods.

## Introduction

The development of technologies such as Information and Communications Technology (ICT) and Web 2.0 technology has brought a *data revolution* to the world (Kitchin 2014, p. 26). As an emerging type of big data, the interest in crowdsourced data has grown in many disciplines (Gray et al. 2015; Garcia-Molina et al. 2016). Two core technologies supporting crowdsourced data have emerged from the multitude of approaches and clustered around two main themes: device/platform-captured data; and user/system-interaction data. The former is the current wave of ICTs, such as digital devices, mobile phones, and the Internet of Things which have penetrated into almost every aspect of daily activities such as work, residency, commuting, communication, consumption, leisure, travel, and so forth, which has been captured with explicit or implicit content at unprecedented spatial and temporal resolutions (Kitchin 2014, p. xv). The second one is the emergence of Web 2.0 technology, which encourages internet users to generate and interact with, rather than only consume, online content (Batty 2012). This allows internet users to create, modify, and supply content to websites, boosting the production of user-generated content related to activities of the public. The penetration of these technologies undoubtedly has led to the explosion of crowdsourced data that are highly related to people's everyday life behavior (Kitchin 2014, p. 80). Consequently,

crowdsourced data have been used in a large body of research, which quickly become an essential source of data-driven analysis in geography and urban studies (Miller and Goodchild 2015).

## Background of VGI and Crowdsourced Data

In the field of geography and urban studies, several scholars have added different perspectives to the basic concept of crowdsourced data, and therefore, it is essential to place crowdsourced data in their context. For example, one perspective adds to the discussion by Crooks et al. (2015) stating that the term *crowdsourcing*, coined by Howe (2006), implied a coordinated bottom-up grassroots effort to contribute information, which is not necessarily limited to geographical information. Adopting this principle, Goodchild (2007) introduced the term *volunteered geographical information* (VGI) to refer to the geographical content generated by nonexpert users. However, Harvey (2013, p. 34) questioned the misuse by researchers who use VGI to refer to data sets that are contributed rather than volunteered by people. He argued that both volunteered and contributed data should be aggregated into the concept of crowdsourced data. Sui et al. (2013, p. 2) also pointed out that VGI is referred to as a type of crowdsourced data for geographic knowledge production. In the book of *The Data Revolution*, Kitchin (2014, p. 96) reviews concepts of various data types in the context of humanities and social sciences and mentions that data that are sourced from a large group of people could be recognized as being crowdsourced, for example, social media data. When applied to urban studies, Crooks et al. (2015) argued that crowdsourced data include explicit sources of collaborative, user-generated mapping and an implicit source such as social media. Over time, given the spread and depth of the type of data that have been generated from devices and platforms, the definition has broadened. Supporting this, See et al. (2016) reviewed the abstracts of 25,338 scientific papers about citizen-derived geographic information published between 1990 and 2015. The literature described this phenomenon

<sup>1</sup>Ph.D. Candidate, Lab of Interdisciplinary Spatial Analysis, Dept. of Land Economy, Univ. of Cambridge, CB3 9EP, Cambridge, UK (corresponding author). ORCID: <https://orcid.org/0000-0002-0182-3573>. Email: [hn303@cam.ac.uk](mailto:hn303@cam.ac.uk)

<sup>2</sup>Reader, Lab of Interdisciplinary Spatial Analysis, Dept. of Land Economy, Univ. of Cambridge, CB3 9EP, Cambridge, UK. Email: [es424@cam.ac.uk](mailto:es424@cam.ac.uk)

Note. This manuscript was submitted on October 23, 2018; approved on September 16, 2019; published online on March 26, 2020. Discussion period open until August 26, 2020; separate discussions must be submitted for individual papers. This paper is part of the *Journal of Urban Planning and Development*, © ASCE, ISSN 0733-9488.

using a multitude of terms, which have emerged from different disciplines; some focused on the spatial nature of the data such as volunteered geographic information (VGI) and neogeography, while other terms have much broader applicability, e.g., crowdsourcing, citizen science, and user-generated content, to name but a few. After identifying the sharp rise of the term crowdsourcing among other 27 relevant terms in academia, See et al. used the term *crowdsourced geographic information* as an umbrella term to represent different types of terms mentioned previously. Building on research by See et al. (2016), the concept of crowdsourced data in this paper refers to data both volunteered and contributed by individuals through ICT-integrated devices and user/system interaction with Web 2.0 technology. The term *crowdsourced* emphasizes the process of data collection, which refers to data sourced by the crowd, rather than the process of data generation. In this context, the main types of crowdsourced data in this paper cover social networking data, points of interest (POIs), and collaborative websites.

### Crowdsourced Data and Urban Activity

In the age of big data, digital data and cities have formed a wide-ranging, diverse, and complex relationship (Kitchin et al. 2017a, p. 44). Crowdsourced data have shown potential in understanding urban activity and its underlying patterns and have been used to solve complex problems or fill important gaps in data analysis that traditional data sets could not cover in urban analysis (Long and Liu 2016; Thakuriah et al. 2016). First of all, since crowdsourced data emerged with location-based services, they are able to provide geographic information such as geotag or geolocation, which is the most rudimentary and vital attribute for urban spatial analysis (Kitchin et al. 2017b, p. 6; Thatcher et al. 2018, p. 123). Second, crowdsourced data are characterized as high-frequency, which updates information that reflects what is happening at present. Furthermore, crowdsourced data are far more cost-effective than traditional data such as surveys or government censuses. Most importantly, this type of data has been collected from volunteered individuals, and their content includes rich information related to urban activity. It should be noted that although the aforementioned advantages of crowdsourced data have been widely perceived by scholars, they are still far from the point of view, abandoning traditional data sets such as traditional census and questionnaire-based data for understanding urban activity. When considering the total number of users and producers of crowdsourced data, they represent only a small fraction of the population; therefore, it would be erroneous to even consider replacing robust census data collection methods with crowdsourced data harvest as a solution for all data problems.

Although the advantages of crowdsourced data have been widely recognized and applied widely, it is apparent that a systematic understanding of how crowdsourced data contribute to urban activity analysis is still lacking. Previous studies either examined crowdsourced data in a general context or focused on specific application of crowdsourced data in urban studies. It is still difficult for researchers in urban studies to have an overall understanding of crowdsourced data in terms of data types, metrics, and methodologies, and furthermore to apply the data in their studies (Shelton et al. 2015; Chen et al. 2017; Xu et al. 2017). Particularly in the current context of big data, how to engage powerful techniques from computer science in terms of data mining is also an obstacle for the majority of researchers in urban dynamics (French et al. 2017). Therefore, this paper aims to investigate how crowdsourced data mining helps understand urban activity and understand how the established perception of crowdsourced data will replace other types of data collection. In order to achieve these goals,

this paper not only focuses on the types and characteristics of crowdsourced data but also critically presents how the methods are applied to data processing. Therefore, it is anticipated that this paper will offer urban researchers the opportunity to develop more robust applications while analyzing urban activity. This study reviews the literature of crowdsourced data applications in the domain of urban activity since 2013. It first introduces review methods, especially for literature inclusion and bibliometric analysis. Based on the cocitation analysis of included papers, it then identifies the fundamental domains, key researchers, and papers on the topic of crowdsourced data. This is followed by a qualitative review of synthesizing data sources, applications, and methods engaged in spatial analysis and sociodemographic and perception analysis. This review also summarizes the potential challenges of crowdsourced data mining.

### Review Method

#### Literature Search

This study first conducted a literature search on the Web of Science database to include papers in the review process. The search query covers two key concepts: crowdsourced data; and urban activity analysis. Each concept is an umbrella of the search terms. Concept I includes terms regarding crowdsourced geographic information, i.e., crowdsourced data, social media, and geotagged. Concept II refers to terms such as urban, city, space, planning, and so forth to select papers focusing on urban analysis. In this way, papers retrieved by this search query are highly relevant to the topic: urban activity analysis with crowdsourced data. In order to retrieve targeted pieces of literature more accurately, this paper adjusts search items after multiple searches. The list of search terms eventually included in the search query is given in Table 1. In the search query, the Boolean AND is used to combine the two main concepts, while OR is used to include research papers. The search terms are expected to appear in field TS which refers to the fields of title, abstract, or keywords. Also, the search query refines publication year in the time span from 2013 to 2019. The final search query is: TS=(crowd\*sourc\* OR social media OR social networks data OR microblog\* OR POI\$ OR point\*of\*interest\* OR VGI OR location-based OR LBS OR LBSM or LBSN OR volunteered geographic information OR user\*generated content OR

**Table 1.** Main concepts and search terms for the search query

Concepts	Search terms
Crowdsourced data	Crowdsourced data Social media LBSN/location-based social network Points of interest/POIs Volunteered geographic information/VGI Geotagged Twitter/tweets Foursquare Flickr Check-ins
Urban activity analysis	Urban City Space Spatial Planning Neighborhood Mobility Spatiotemporal

geo\*tagged OR *geo-big data* OR Twitter OR tweets OR Four-square OR Flickr OR geodata OR check-ins) AND TS=(urban OR city OR cities OR space OR spatial OR planning OR spatio? temporal OR mobility) AND PY=(2013–2019).

## Literature Inclusion

To exclude irrelevant studies, this paper refines the results with Web of Science Categories where urban studies and regional urban planning are chosen. Then, this paper screens the results with research question-based criteria to further narrow down the data. The inclusion criteria are: (1) Does the paper apply crowdsourced data to conducting urban spatial analysis? (2) Does the paper clearly state the method of processing crowdsourced data? (3) Does the paper discuss the trends or challenges of crowdsourced data?

Finally, after reviewing papers with the aforementioned criteria, 226 papers are selected and subjected to review. The sources of papers are from journals and edited books such as *Computers, Environment and Urban Systems*; *Cities*; *Landscape and Urban Planning*; *International Journal of Geographical Information Science*; *ISPRS International Journal of Geo-Information*; *Urban Planning*; *Journal of Urban Technology*; *Seeing Cities Through Big Data Research Methods and Applications in Urban Informatics*; *Springer Geography*; and *Applied Geography*. Almost 19% of articles are from *Computers, Environment and Urban Systems* because most applications of crowdsourced data are multidisciplinary research involving computer science and urban studies (see Table 2).

## Bibliometric Analysis

Before reviewing the selected papers for qualitative synthesis, a bibliometric analysis based on the metadata of literature was conducted. Bibliometric analysis is used to explore the relationships among publications in terms of citation information, bibliographic information, abstract, keywords, funding details, and other metadata. As a method of bibliometric analysis, cocitation analysis is used to measure the frequency of which two documents/authors are together cited by others. The more cocitation two documents/authors get, the higher cocitation strength between them, and the more likely they are related semantically (van Eck and Waltman 2011). Author cocitation was proposed by White and Griffith (1981), which was described as a measure of the relatedness of authors' works. Through author cocitation,

this analysis shows from which disciplines/domains a topic is derived and who are the pivotal researchers/scholars in each domain, and how they connect. This analysis not only gives a broad conception of the background but also specifically answers the questions: What are the fundamental domains of crowdsourced data mining in urban activity? Who are the impactful researchers in these domains? Additionally, through cocitation analysis based on reference lists of all the selected papers, this review identifies key studies which are highly cocited by included papers, answering the third question: What are the key papers in this topic? In addition, bibliometric analysis helps to cover those high-cited, valuable, and fundamental papers that are excluded in the previous process due to the limitation of time span. In this paper, these cocitation analyses are conducted and visualized on VOSviewer (van Eck and Waltman 2011).

## Fundamental Domains, Key Researchers, and Papers

### Domains and Key Researchers

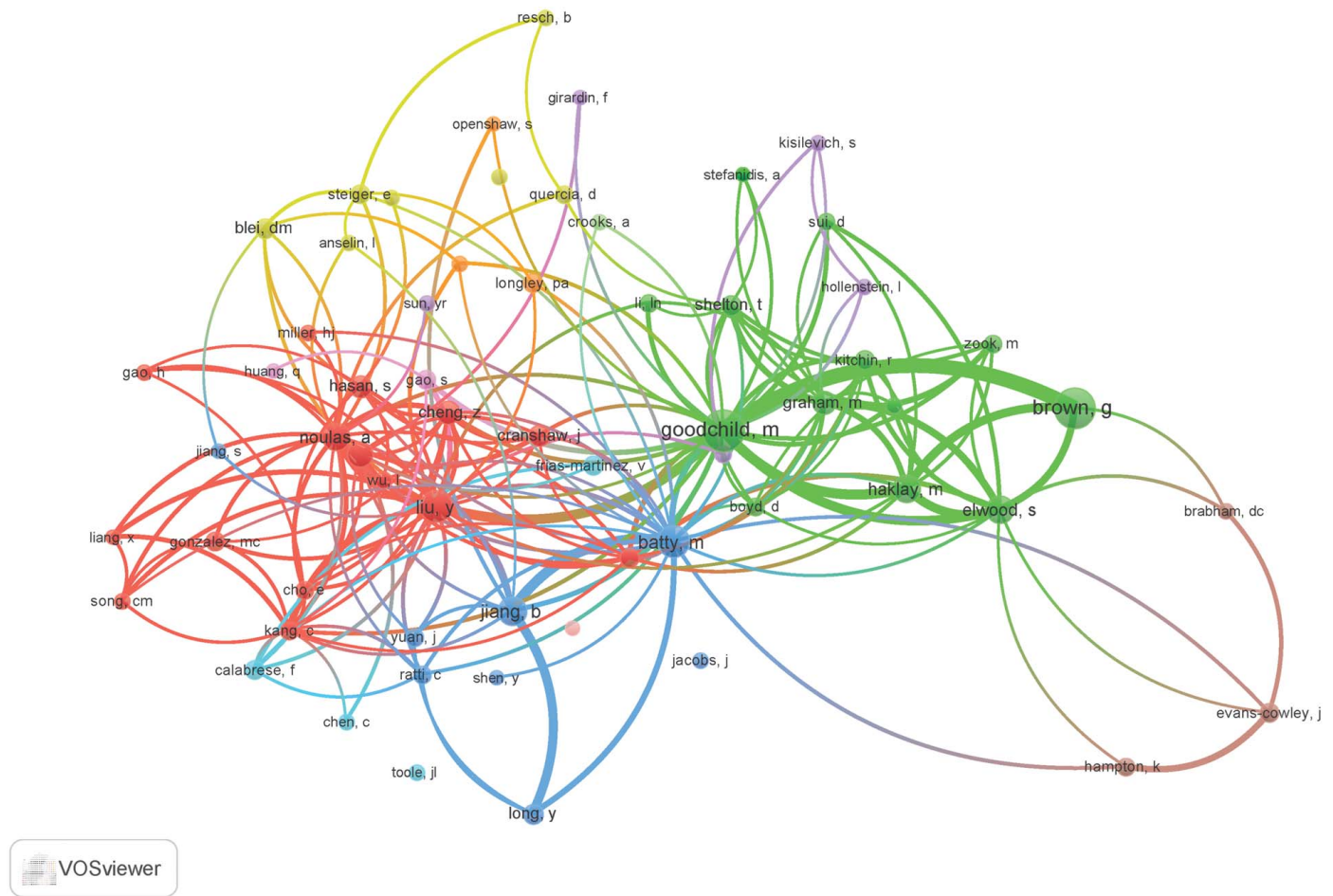
According to the results of the cocitation analysis, 6,867 authors are cited in 226 selected papers while 62 authors received more 15 citations. The network between cocited authors is visualized in Fig. 1. Each author is represented as a node in the figure whose size refers to the frequency with which two authors are together cited by others. The distance between nodes indicates the relatedness between two authors, and the thickness of lines indicates the total link strength. Here, total link strength indicates the sum of link strength of an author with other authors. Four clusters, assigned with different colors, are identified in the cocitation map. Although cocitation analysis can well illustrate the disciplinary structure well, it does not identify the topic of each cluster. To interpret the semantic clusters shown in Table 3, we searched the expertise and research interests of key researchers from their university profile page, Google Scholar, ResearchGate, LinkedIn, personal websites, and other sources. By synthesizing their research interests and expertise, we identified clusters as four domains, which are GIScience, Data Science, Urban Studies, and Human Geography.

GIScience plays a vital role in this topic, which is led by Michael Goodchild, Sarah Elwood, and Greg Brown, who clearly have a background in geography and geographic information systems. Their studies dominate the whole network of authors. Another big group led by Zheng Yu, Justin Cranshaw, and Daniele Quercia with data science background also contributes findings to this research topic. Most of them work with internet companies such as Microsoft Research, Foursquare and Google, which have access to big location-based data. With massive location-based social media data, they focus on exploring spatiotemporal patterns of human mobility to solve practical problems such as transportation, location recommendation, route planning, and emergency management. Researchers from Urban Studies also provide building blocks through applying crowdsourcing not only on understanding urban structure and urban form but also supporting urban models and planning support systems. Michael Batty is the pivot for proposing the nature of complexity of urban systems and emphasizing the importance of data derived from citizens. Another domain with a geographic background is led by Liu Yu, focusing on time geography and human geography. Their studies emphasize on the relationship between mobility and demographic characteristics and also pay attention to explore the patterns beyond spatiotemporal attributes.

**Table 2.** Top 10 sources title of 226 papers

Order	Source title	Number	Ratio (%)
1	<i>Computers, Environment and Urban Systems</i>	42	18.58
2	<i>Cities</i>	19	8.41
3	<i>Landscape and Urban Planning</i>	12	5.31
4	<i>International Journal of Geographical Information Science</i>	11	4.87
5	<i>ISPRS International Journal of Geo-Information</i>	11	4.87
6	<i>Urban Planning</i>	11	4.87
7	<i>Journal of Urban Technology</i>	10	4.43
8	<i>Seeing Cities Through Big Data Research Methods and Applications in Urban Informatics</i>	6	2.66
9	<i>Springer Geography</i>	6	2.66
10	<i>Applied Geography</i>	5	2.21





**Fig. 1.** Network map of cocited authors by 226 included papers.

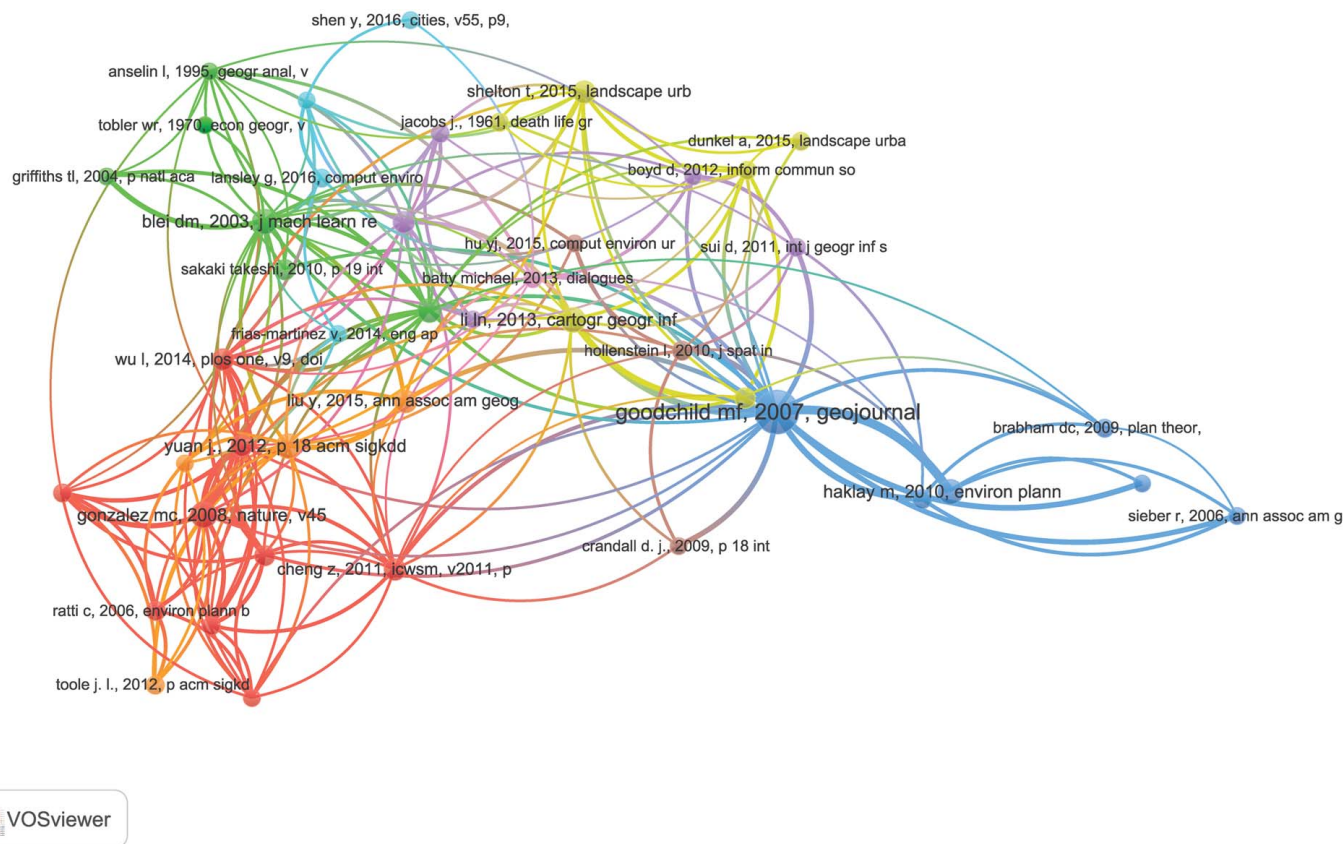
**Table 3.** Fundamental domains, key researchers, their research interests, and expertise; names with \* are the leading researchers

Clusters	Domains	Key researchers	Research interests and expertise
Cluster1	GIScience	Michael Frank Goodchild*; Mark Graham; Sarah Elwood*; Muki Haklay; Gregory G Brown*; Taylor Shelton; Rob Kitchen; Matthew Zook; and Danah Boyd	Volunteered Geographic Information, Crowdsourcing, Citizen Science, Digital Economy, DigitalGeo, Geoweb, Open Data, Participation Geographic Information, Social Media Data, Social Network Analysis, and Social Science
Cluster2	Urban Studies	Michael Batty*; Jiang Bin; Ying Long; and Nicholas Jing Yuan	Urban Modeling, Spatial Analysis, Complex System, and Urban Computing
Cluster3	Human Geography	Liu Yu*; Chaogui Kang; Marta C. Gonzalez; and Francesco Calabrese	Human Geography, Time Geography, Urban Dynamics, Human Mobility, Travel Pattern, Social Media Data, and Social Sensing
Cluster4	Data Science	Anastasios Noulas*; Zheng Yu*; Derek Zhiyuan Cheng; Justin Cranshaw; Samiul Hasan; Eunjoon Cho; and Daniele Quercia	Big Data, Data Mining, Spatiotemporal Data Mining, Artificial Intelligence, Machine Learning, Transportation, Human Mobility, Disaster Management, Mobile Phone Data, Social Media Data, Location-based Services, Recommendation Systems, and Social Network Analysis

### Core Papers

According to the results of cocitation based on references, there are 44 references most cited (more than 10 times) among all the 9,499 references cited in 226 papers. The visualization of cocitation networks is shown in Fig. 2, where each dot represents one reference item. Those most cocited papers are the fundamental and core findings/arguments for studies in crowdsourced data mining within an urban context (see Table 4). In order to understand the main arguments of those core papers and how they link with others, we analyzed the relationship between them alongside the scatterplot shown in Fig. 3, in association with the key domains and researchers

identified in the previous section. In general, based on the chronological order of these core papers, we found that this topic was first driven by GIScience and then fueled by the application in research from data science and human geography. Concerning the topic of crowdsourced data, Goodchild (2007) sharply noticed the potential from the blooming of techniques such as Web 2.0 and Global Position Services in the context of GIScience and coined the concept VGI. In this most highly cited paper, Goodchild also highlighted the vast potential of citizens who carry sensors and voluntarily contribute geographic information. Since then, this concept has been broadly accepted by many researchers. Haklay (2010) examined



**Fig. 2.** Network map of cocited references by 226 included papers.

the quality of crowdsourced data and stated its robustness. After this, there is a rapid increase in applications of crowdsourced data.

Before crowdsourced data became a hot topic, another type of data, called mobile phone data, had shown the potential of a massive and fine-granularity location-based data set. In the level of application, Ratti et al. (2006) creatively used mobile phone data to understand human mobility. However, this topic began to attract considerable attention only from 2008 when González et al. (2008) published a paper in *Nature* about the law of human trajectories at the individual level which stated that “human trajectories show a high degree of temporal and spatial regularity, each individual being characterized by a time-independent characteristic travel distance and a significant probability to return a few highly frequented location.” This robust finding based on complex system and statistics has attracted massive attention from different disciplines and has contributed to countless research about human mobility and located-based data. It led a trend in the domains of data science (some of which are from computer science), exploring human mobility and urban dynamics integrated with different algorithms. The most cited algorithm is latent Dirichlet allocation proposed by Blei et al. (2003) and since then, the topic model has been widely applied to cluster human behavior or their mobility pattern from crowdsourced data. At the same time, the trend of data mining expanded from mobile phone data to crowdsourced data such as social media data (Cheng et al. 2011; Noulas et al. 2012) and POIs data (Yuan et al. 2012). One interpretation is that social media data are more accessible compared with mobile phone data and taxi trajectory data which are exclusively owned by telecommunication carriers and private companies, respectively.

Benefiting from data mining techniques in data science, the use of crowdsourced data starts to appear in Geography, especially

Human/Time Geography. The overlapped clusters shown in Fig. 1 reveal the relatedness between these two domains. This obvious shift starts approximately in 2013 with two papers that (Crampton et al. 2013; Li et al. 2013) summarize the potential of crowdsourced data in finding patterns of urban dynamics, especially the socioeconomic patterns that have been neglected by previous research. Their argument is later echoed by almost two simultaneous papers: Liu et al. (2015) who propose the concept of *social sensing* that denotes capturing socioeconomic features from big data on the individual level and Shelton et al. (2015).

## Sources of Crowdsourced Data

By reviewing the 226 papers included in the search result and 20 most cited references, the sources of crowdsourced data can be divided into (a) social media, published by individuals for sharing or social networking; (b) points of interest, specific location of functional buildings and facilities; and (c) collaborative websites, volunteered information, or contributed information with collaboration through online platforms. From the aspect of analysis, applications of crowdsourced data include activity patterns, patterns of mobility, social behavior, land use, semantic analysis, event detection, location inference, disaster management, traffic management, and so forth. Both attributes of each type of data and the method of each facet of analysis will be argued subsequently.

## Social Media

As a technology built on Web 2.0, social media, for example, Facebook, Twitter, Instagram, and Sina Weibo, provide internet

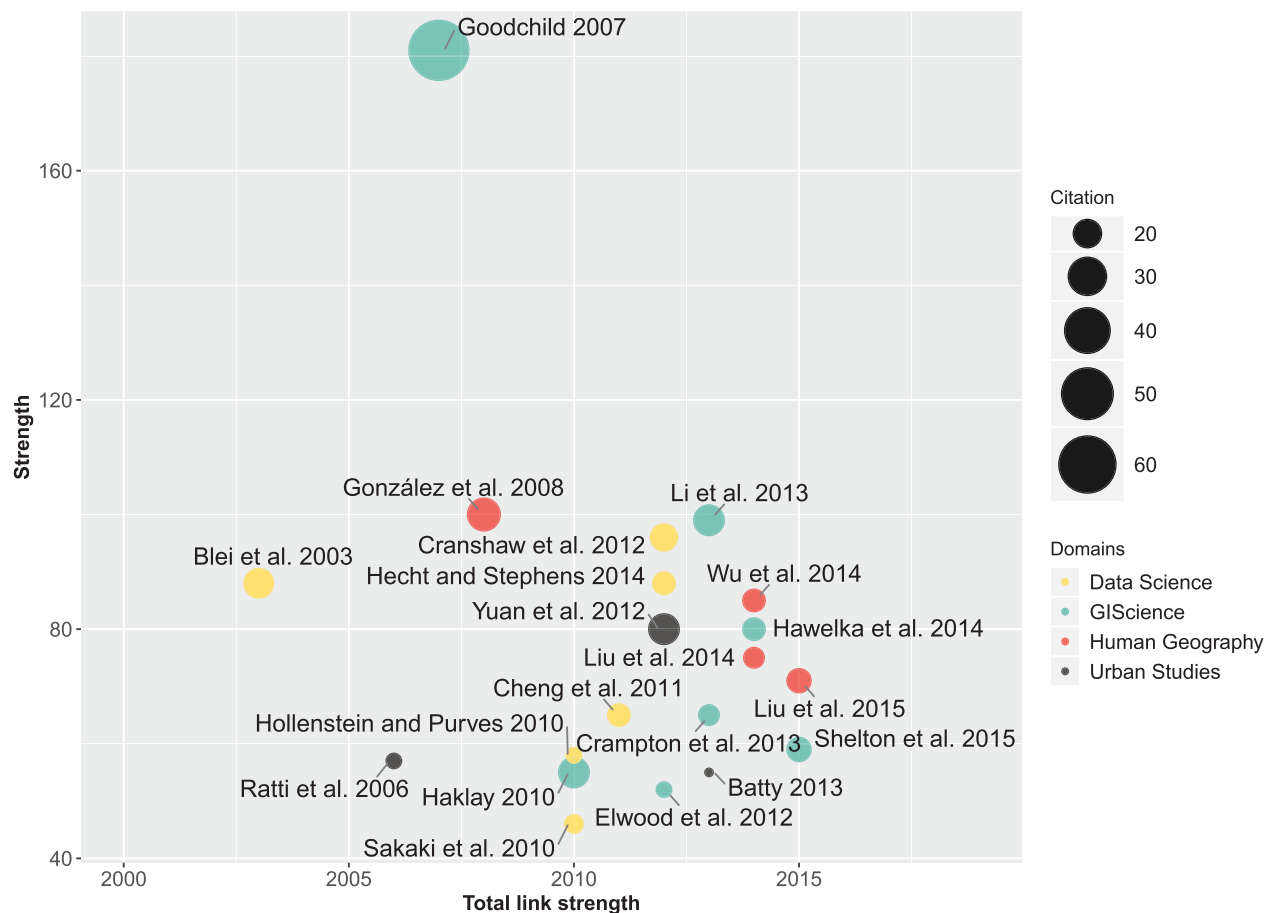
**Table 4.** 20 Most cited references by 226 included papers

Studies	Title	Year	Publication title	Citation	Total link strength
Goodchild (2007)	Citizens as Sensors: The World of Volunteered Geography	2007	<i>GeoJournal</i>	66	181
González et al. (2008)	Understanding Individual Human Mobility Patterns	2008	<i>Nature</i>	24	100
Haklay (2010)	How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets	2010	<i>Environment and Planning B: Planning and Design</i>	22	55
Li et al. (2013)	Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr	2013	<i>Cartography and Geographic Information Science</i>	22	99
Yuan et al. (2012)	Discovering Regions of Different Functions in a City Using Human Mobility and POIs	2012	<i>Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining</i>	22	80
Blei et al. (2003)	Latent Dirichlet Allocation	2003	<i>Journal of Machine Learning Research</i>	21	88
Cranshaw et al. (2012)	The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City	2012	<i>Sixth International AAAI Conference on Weblogs and Social Media</i>	19	96
Liu et al. (2015)	Social Sensing: A New Approach to Understanding Our Socioeconomic Environments	2015	<i>Annals of the Association of American Geographers</i>	17	71
Shelton et al. (2015)	Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information	2015	<i>Landscape and Urban Planning</i>	17	59
Cheng et al. (2011)	Exploring Millions of Footprints in Location Sharing Services.	2011	<i>ICWSM</i>	16	65
Hawelka et al. (2014)	Geo-Located Twitter as Proxy for Global Mobility Patterns	2014	<i>Cartography and Geographic Information Science</i>	16	80
Noulas et al. (2012)	A Tale of Many Cities: Universal Patterns in Human Urban Mobility	2012	<i>PloS One</i>	16	88
Wu et al. (2014)	Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-in Data	2014	<i>PloS One</i>	16	85
Crampton et al. (2013)	Beyond the Geotag: Situating “big Data” and Leveraging the Potential of the Geoweb	2013	<i>Cartography and Geographic Information Science</i>	15	65
Liu et al. (2014)	Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-in Data	2014	<i>PloS One</i>	15	75
Sakaki et al. (2010)	Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors	2010	<i>Proceedings of the 19th International Conference on World Wide Web</i>	14	46
Elwood et al. (2012)	Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice	2012	<i>Annals of the Association of American Geographers</i>	13	52
Hollenstein and Purves (2010)	Exploring Place Through User-Generated Content: Using Flickr Tags to Describe City Cores	2010	<i>Journal of Spatial Information Science</i>	13	58
Ratti et al. (2006)	Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis	2006	<i>Environment and Planning B: Planning and Design</i>	13	57
Batty (2013)	Big Data, Smart Cities and City Planning	2013	<i>Dialogues in Human Geography</i>	12	55

users with a platform to actively generate or contribute content (text, places, images, and videos) through sharing and interacting with others in the digital community (Kitchin 2014). With the proliferation of ICTs, social media data (stream/post) have become ubiquitous. Only for Twitter, there are approximately 500 million tweets sent by millions of users every day (Internet Live Statistics 2016). Sina Weibo, an equivalent of Twitter in China, has up to 376 million monthly active users according to the data in the third quarter in 2017. Most social media platforms provide sample data sets or application programming interface (API) to access historical or real-time social media streaming. Data harvested from the database record social media posts by row with metadata such as content, created time, location, and tags and also the information of users or so-called data subjects such as user id, user location, and profile image. Built on these high-dimensional features, social media data, in particular, work well for mining the patterns of human mobility, activity patterns, and social networks and sentiment detection, sociodemographic characteristics analysis, event detection, and disaster management (see Table 5).

Characterized as high spatiotemporal resolution, social media data are commonly used in the identification of human mobility pattern at intraurban, interurban, and global levels (Gabrielli et al. 2014; Hawelka et al. 2014; Liu et al. 2014; Wu et al. 2014; Abbasi et al. 2017), activity patterns (Noulas et al. 2013; Hasan and Ukkusuri 2014; Steiger et al. 2015; Martín et al. 2019) including specific patterns of tourists in urban areas (Salas-Olmedo et al. 2018), and event detection and emergency management (De Albuquerque et al. 2015; Granell and Ostermann 2016; Kim and Hastak 2018). Focusing on contextual information, user-generated content of social media provides a source for describing actual activities of digital footprints (Lansley and Longley 2016) and becomes the indicator for sentiment detection with natural language processing techniques (Mitchell et al. 2013; Hollander and Hartt 2018). In recent years, the application of social media in urban dynamics has expanded to urban social problems by integrating with census data to assign socioeconomic/demographic features to users (Li et al. 2013; Shelton et al. 2015) or by extracting demographic features (gender, age, and ethnicity) from profile information through methods such as face detection and name analysis





**Fig. 3.** Scatter plot of 20 most cited references.

**Table 5.** Main applications and representative studies by social media

Main applications	Representative studies
Human mobility pattern	Hawelka et al. (2014), Liu et al. (2014), Wu et al. (2014), Gabrielli et al. (2014), and Abbasi et al. (2017)
Human activity pattern	Noulas et al. (2013), Hasan and Ukkusuri (2014), Steiger et al. (2015), Martin et al. (2019), Salas-Olmedo et al. (2018), and Li et al. (2018)
Event detection and emergency management	De Albuquerque et al. (2015), Granell and Ostermann (2016), and Kim and Hastak (2018)
Sentiment detection	Mitchell et al. (2013), Steiger et al. (2016), Hollander and Hartt (2018), Roberts et al. (2018), and Ashkezari-Toussi et al. (2019)
Social dynamics	Li et al. (2013), Shelton et al. (2015), Huang and Wong (2016), and Luo et al. (2016)
Urban perception	Hollenstein and Purves (2010) and Feick and Robertson (2015)

(Bocconi et al. 2015; Luo et al. 2016). Moreover, social media sources such as Instagram and Flickr contain massive image content examined for urban perception (Liu et al. 2016b) and patterns of tourists (Li et al. 2018).

### POIs Data

POIs refer to location points associated with commercial facilities, public areas, and transportation facilities (see Table 6). POIs data

are normally obtained from open-sourced databases such as OpenStreetMap POIs, business map service POIs database (Google Places and Gaode map), and POIs-based social networking data, known as check-in data, i.e., Foursquare or Yelp. POIs-related data contain location name, function, postcode, and address which reflects the distribution of facilities. Recently, there is noticeable integration between common social media platforms and location-based social networking services. For example, Twitter partnered with Foursquare and Yelp in 2015 and 2016, respectively, whereas Sina Weibo developed their POIs creator. This paper differentiates POI check-ins from social media data because POIs data are basically location/venues-oriented data gathering users' visit log or reviews, while social media steaming is content-based data whereby location is alternatively included from individuals.

POIs data including POI check-ins are applied by researchers in exploring urban dynamics in terms of urban structure, urban functional use detection, patterns of urban activity, urban vibrancy, and population mapping. As location-based data, POIs have a strong connection with the urban-built environment. They have been commonly used to explore the physical features of urban areas such as urban form/structure (Pan et al. 2018; Song et al. 2018; Deng et al. 2019), urban boundary, and urban growth (Long et al. 2015; Daggitt et al. 2016). Based on the subdivided categories of places, a group of researchers use POIs to identify the land use at the store level and detect urban function (Yuan et al. 2012; Zhan et al. 2014; Frias-Martinez and Frias-Martinez 2014; Liu et al. 2017).

To explore patterns of activity, Foursquare Labs presented the global distribution of urban activity and allowed activity distribution analysis to visualize 500 million POIs data (Foursquare Labs

2013). Since the increasing integration between social media data and location-based services, POIs data have become more prevalent in exploring the purpose of human activity (Hasan and Ukkusuri 2015; Pouke et al. 2016; Yu et al. 2016) and urban dynamics related to human activity such as urban vibrancy (Jin et al. 2017; Yue et al. 2017; Wu et al. 2018), urban sociospatial inequality (Shelton et al. 2015), and urban deprivation (Quercia and Saez 2014). Overall, POIs data link human activities with the built environment, which provides researchers with an opportunity to understand purposeful urban activity and the function of land use.

### Collaborative Websites

As a primary source of crowdsourced data, collaborative websites refer to the web services contributed by individuals by uploading and editing exclusively thematic content such as geographic maps and geotagged images complying with specific acceptance policy. OpenStreetMap is the wiki-like mapping service focusing on geospatial features such as land, transportation infrastructure, and building. This volunteered geographic information has been applied to urban land parcel identification (Liu and Long 2016), land cover assessment (Estima and Painho 2015), and urban boundary identification (Schlesinger 2015). Geotagged photosharing websites such as Geograph Worldwide and Panoramio provide platforms for photo sharing according to geolocation, content, and categories of photos. This type of data has been applied to the detection of location preference, patterns of travel routes, perception of the built environment, and city attractiveness (Paldino et al. 2015; García-Palomares et al. 2015; Dubey et al. 2016). Apart from these sources, there is an increasing number of

crowdsourcing projects which are focusing on specific aspects of urban dynamics. Recently, Rae (2016) built a crowdsourcing website for citizens to digitalize the boundary of several cities on the interactive map. CASA (2018) from UCL released the Colouring London project to invite residents of London to assign color according to the features of buildings with which they are familiar.

Among the three main sources, social media is the dominant one because of its extensive coverage, deep usage penetration, massive size, and relatively open accessibility. Social media will further show its value in analyses of urban dynamics because of its growing integration with location-based services. For collaborative websites, it shows more flexibility of collecting knowledge from citizens by designing platforms focusing on different aspects of activities. The crowdsourcing mode in those projects with online websites will also be adopted by researchers in this area. Although this paper separately discusses the three primary sources and their main applications, what should be mentioned is that most of the studies require multiple sources of crowdsourced data. Data fusion (within crowdsourced data, other data sets including census data, smart card records, bank transactions, among others) is broadly used in the big data-driven analysis (Hasan et al. 2013; Lenormand et al. 2015; Jin et al. 2017).

In the following section, this paper synthesizes all the applications about urban dynamics mentioned previously and illustrates state-of-the-art of methods engaged in the application of spatial analysis, sociodemographic analysis, and perception analysis.

## Spatial Analysis: Mobility Patterns, Functional Areas, and Event Detection

### Mobility Patterns: Dynamic Flows of Urban Activity

According to the development of ICTs, location-aware devices such as smartphones, mobile devices, and vehicles allow users to record and share their whereabouts of urban activity. Based on spatial and temporal information, most of the studies on human mobility have been carried out with the geospatial data and analytical models (see Table 7) including the gravity model, the generalized potential model, the rank-based movement model, and the radiation model which have all been proposed (Abbasi et al. 2017; Barbosa et al. 2018).

These methods have been applied to different scales of mobility studies including intraurban, interurban, national, and global studies. At the intraurban level, Wu et al. (2014) developed a model based on transition probability of travel demands and then validated it with the agent-based model. When it comes to the interurban level, Liu et al. (2014) fitted the gravity model to analyze the

**Table 6.** Main applications and representative studies by POIs

Main applications	Representative studies
Urban structure	Pan et al. (2018), Song et al. (2018), and Deng et al. (2019)
Urban growth detection	Long et al. (2015), and Daggitt et al. (2016)
Human activity pattern	Noulas et al. (2013), Hasan and Ukkusuri (2015), Sun and Li (2015), and Pouke et al. (2016)
Urban functional use detection	Zhan et al. (2014), Frias-Martinez and Frias-Martinez (2014), Jiang et al. (2015), Long and Liu (2016), Yao et al. (2017), Liu et al. (2017), Gao et al. (2017), Li et al. (2018), and Zhai et al. (2019)
Urban vibrancy	Jin et al. (2017), Yue et al. (2017), and Wu et al. (2018)
Urban deprivation	Quercia and Saez (2014), and Shelton et al. (2015)

**Table 7.** Representative studies of mobility patterns

Research topic	Methods	Data set	Case study	Authors and year
Intraurban mobility patterns	Temporal transition probability matrix; ABM for simulation	Social media	Shanghai, China	Wu et al. (2014)
	LDA topic model and perplexity	Twitter	New York City	Hasan and Ukkusuri (2014)
	Geo-SOM/H-SOM for clustering; LDA topic model for semantic similarity	Twitter	Greater London	Steiger et al. (2016)
Interurban human mobility and spatial interaction	Gravity model; Complex network analysis	JiePang POIs	China	Liu et al. (2014)
Global mobility patterns	The radius of gyration; Temporal variations; Network of tweet flows	Twitter	Global	Hawelka et al. (2014)
Human mobility prediction	Rank-based model	Foursquare	New York City	Abbasi et al. (2017)



underlying patterns of trips and spatial interaction between 370 cities in China. The study suggested that crowdsourced data perform better when revealing the collective level of spatial interaction, compared with intraurban human mobility. At the global level, social media has been a suitable enough form of data source to explore mobility between nations or districts, since it is difficult for mobile phone data to cover worldwide scales due to the high fragmentation of the mobile telecom market. Hawelka et al. (2014) uncovered global patterns of human mobility against the volume of international travelers, characteristics of flows between nations, temporal patterns of international travel, and mobility networks. Recently, a rank-based model has been developed to predict human mobility by ranking the probability of commuting between venues (Noulas et al. 2012; Liang et al. 2015; Chen et al. 2016). Furthermore, Abbasi et al. (2017) applied check-in weighting schema to rank the probability. However, the accuracy of the computing rank is not sufficient, which affects the estimation of mobility. To further subset mobility patterns, Yang et al. (2017) labeled users as natives and nonnatives and proposed the *indigenization coefficient* to estimate the extent of natives. Similarly, Li et al. (2018) examined the spatial interaction between locals and tourists.

Studies elaborated previously only depict the spatiotemporal distribution of individuals, without considering many other features embedded in crowdsourced data. Steiger et al. (2016) identified spatiotemporal clusters of urban activity with a self-organizing maps (SOM) algorithm, which handles high-dimensional data sets well. Importantly, the clustering process also considers the semantic similarity of tweets supported by the LDA model, which further reveals the patterns of mobility.

### Functional Areas: Activity-Based Analysis

To understand urban activity, it is not enough to identify the mobility patterns of human beings while neglecting the purpose and content of their activities. Due to the strong connection between human behavior and its linked built environment, activities happening in the urban space define the function of urban areas (Crooks et al. 2015). In addition, activities and habitats of citizens also form functional areas differing from administrative units in terms of extent

and structure. To identify those functional areas in cities, traditional methods such as remote sensing technologies present limitation in capturing the socioeconomic attributes of human dynamics, and some social science methods such as interviewing, observing, and cognitive mapping are usually costly and time-consuming for researchers (Zhou and Zhang 2016; Chen et al. 2017). By leveraging the geospatial information with fine granularity and continuously updated content of individuals' behavior, researchers have attempted to delineate urban functional areas with crowdsourced geospatial data (see Table 8). It is fair to say that the introduction of crowdsourced data allows studies to change from movement-based analysis to activity-based analysis (Wu et al. 2014).

Hollenstein and Purves (2010) analyzed 8 million Flickr images with georeferenced tags to understand how people describe city core areas with different names and how these areas are distributed. When it comes to the neighborhood level, Cranshaw et al. (2012) in the representative case, the *Livehoods project*, developed a clustering model considering spatial proximity and social proximity, with check-in data to map neighborhoods dynamically. To identify the functional areas or land use pattern, the clustering method shows its importance on aggregate objects into groups spatially. Wang et al. (2016) compared three representative spatial clustering algorithms, density-based spatial clustering of application with noise (DBSCAN), expectation-maximization (EM), and K-means, arguing that K-means, as an algorithm based on the distance of objects, is appropriate to process high-dimensional objects for identifying land use patterns. Differing from the commonly used clustering method, kernel density estimates (KDE), Aadland et al. (2016) developed an algorithm employing fuzzy-set theory to identify the boundary of neighborhoods.

In terms of identifying urban functional areas not just with spatial location, Yuan et al. (2012) introduced a probabilistic topic model which regards a region as a document and function as a topic, delineating urban functional areas through the clustering method based on LDA. From the perspective of urban planning, Crooks et al. (2015) provided insights into the urban forms and function powered by crowdsourced data and explained how to conduct implicit function classification based on POIs with the LDA topic model at three scales (buildings, streets, and neighborhood).

**Table 8.** Representative studies of urban functional areas

Research topic	Methods	Data set	Case study	Authors and year
Community detection	Spatial proximity; social proximity analysis LDA topic model	Twitter; Foursquare Foursquare	Pittsburgh New York City	Cranshaw et al. (2012) Hasan and Ukkusuri (2015)
Urban functional areas	LDA topic model	POIs; GPS trajectories	Beijing	Yuan et al. (2012)
	LDA topic model	POIs	Washington, DC	Crooks et al. (2015)
	LDA topic model	Foursquare	US (10 most populated)	Gao et al. (2017)
	Word2Vec model	Baidu POIs	Pearl River Delta, China	Yao et al. (2017)
	Simulated annealing and hill climbing algorithm	Yahoo POIs	Boston	Jiang et al. (2015)
	LRA-based model	Social media	Shanghai, China	Zhi et al. (2016)
	Topic models; Support vector machine	OpenStreetMap; Gaode POIs; Tencent	Hangzhou, China	Liu et al. (2017)
Mixture of urban function	SVM classification	Twitter; Foursquare	Boston and Chicago	Zhou and Zhang (2016)
	Dynamic time warping (DTW) distance-based k-medoids method	Social media data from Tencent	Guangzhou, China	Chen et al. (2017)
	Spatial entropy	POIs; Social media data from Tencent	Beijing	Li et al. (2016a)
	Hill number including richness, entropy, and Simpson's index	POIs	Shenzhen, China	Yue et al. (2017)
	Shannon entropy	POIs; social media	Shenzhen, China	Wu et al. (2018)

However, the LDA topic model only considers the frequencies of POIs neglecting the inner spatial correlations, so Yao et al. (2017) engaged a deep learning model (Google Word2vec) to identify functions by considering the high-dimensional features of POIs at the travel analysis zones.

In detecting functional areas, social media check-in data attract more attention than POIs since it is challenging to match human movements consistently just with POIs. However, as we mentioned before, not all social media comes with POIs reference. Zhou and Zhang (2016) trained a support vector machine (SVM) classifier based on tweets with foursquare venues and applied it to all the social media data to evaluate the content of activities. Liu et al. (2017) integrated the topic model with SVM for classification while including remote sensing to extract urban functional areas. Without predefining categories, Zhi et al. (2016) built up a model based on low-rank approximation (LRA) to detect functional regions and its temporal pattern with a large social media check-in data set in one year. To identify functional areas at the building level, Chen et al. (2017) applied a dynamic time warping (DTW) distance based on k-medoids to perform time series clustering. On top of all these, a group of researchers calculated the mixture of function to evaluate urban vibrancy through Shannon entropy (Wu et al. 2018), spatial entropy (Li et al. 2016a), and Hill number (Yue et al. 2017).

Despite the advantages of crowdsourced data in delineating functional areas, Li et al. (2016) also highlighted the biases when introducing geospatial data to analyze urban activity. Since this data heavily rely on mobile phone devices, night activities cannot be captured when the mobile phone is powered off.

### Event Detection: Crowd-Based Monitoring

Crowdsourced data, especially social media data, are characterized as high frequency which describes what is happening in which parts of the city (Xia et al. 2015). Therefore, multiple crowdsourced data streams have been collected to detect and depict local or emergency events (see Table 9).

When information from crowdsourced data is extracted and aggregated, it is not enough to just collect the location and timestamp. Instead, it is essential to capture information from content, including text, tags, or images. Chen and Roy (2009) exploited event-related tags from annotated photos and then grouped photos based on tag usage occurrence. Multiple events could be identified in association with temporal and locational attributes. To reduce the workload of preselecting event-relevant tweets, Walther and Kaisser (2013) pre-selected posts based on geographical and temporal proximity and introduced a machine learning algorithm to evaluate whether detected events happen in the real world. As one type of event detection, some studies also attempted to identify traffic anomalies with crowdsourced data. For example, Pan et al. (2013) used the traditional

GPS trajectory data set to detect the change of routine from drivers and fuse the social media data which related to traffic anomalies to conduct an in-depth temporal analysis.

Another branch of study under this analysis is emergency events management. The web user can be seen as a *social sensor* who can provide more information when emergency events happen. To estimate the location of a specific emergency event, Sakaki et al. (2010) examined different methods and found that particle filtering works better in estimating the epicenter of earthquakes. Crooks et al. (2013) built a similar sensor system with social media data and engaged the signal-to-noise ratio to detect the epicenter and impact area. Shifting from the physical location of emergency events to public opinion, Xu et al. (2016) conducted semantic analyses to extract main topics from the related social media.

### Sociodemographic and Perception Analysis: City Attractiveness, Demographic Characteristics, and Sentiment Detection

The analyses summarized in the previous section are orientated toward answering the spatiotemporal variation of urban activity. Since the crowdsourced data are collected at the individual level, it comes with multiple features of data subjects beyond spatial and temporal information. In general, sociodemographic and perspective features may directly be included in the profile of users and the generated content, or it is hidden in the spatiotemporal preference of their activities. These features further allow the exploration of patterns of urban activity and their underlying mechanisms. This section summarizes the application of features beyond geographic information of crowdsourced data.

#### City Attractiveness

Evaluating city attractiveness based on crowdsourced data provides insights into several fields such as urban planning, flows forecasting, transportation, and economics. City attractiveness refers not only to mobility but also how people experience the city. Focusing on local attractiveness, Girardin et al. (2009) quantified attractiveness by fusing mobile phone data and geotagged photos from Flickr and tracked the evolution of central areas. However, this research could only depict the distribution and density of digital footprints without considering the driving force of attractiveness. To overcome this, Huang et al. (2010) used POIs and GPS trajectory to identify spatiotemporal attractiveness. In order to quantify city attractiveness more accurately, Sobolevsky et al. (2015) fused multisource data including Flickr, Twitter, and bank card transactions in order to identify foreign visitors and their mobility patterns. Regarding studies on global attractiveness, one representative case is that of Paldino et al. (2015), who analyzed the data set with geotagged photos over 10

**Table 9.** Representative studies of event detection

Research topic	Methods	Data set	Case study	Authors and year
Event detection	Wavelet-based spatial analysis	Flickr photos	US	Chen and Roy (2009)
	K-means clustering method;	Twitter	Japan	Lee and Sumiya (2010)
	Voronoi diagram			
	Machine learning component	Twitter	New York City	Walther and Kaisser (2013)
Traffic anomalies	Anomaly analysis	Weibo	Beijing	Pan et al. (2013)
Emergency events management	Particle filtering	Twitter	Japan	Sakaki et al. (2010)
	Signal-to-noise ratio	Twitter	US	Crooks et al. (2013)
	Text classification	Weibo	China	Xu et al. (2016)
	Generalized additive model (GAM)	Twitter	Germany	De Albuquerque et al. (2015)

**Table 10.** Representative studies of social demographics

Research topic	Methods	Data set	Case study	Authors and year
Sociodemographic characteristics	Home location identification	Twitter; Flickr photos	California	Li et al. (2013)
	Home location identification	Twitter	Washington, DC	Huang and Wong (2016)
	Name analysis	Twitter	London	Longley et al. (2015)
	Multiple regression analysis	Twitter	Madrid, Spain	García-Palomares et al. (2018)
	Exploratory spatial data analysis on GeoDA	Twitter	London	Hofer et al. (2015)
Social segregation	Home location identification	Twitter	Louisville, US	Shelton et al. (2015)
	Name analysis	Twitter	Chicago	Luo et al. (2016)
Socioeconomic characteristics	Home location identification	Yelp	New York City	Davis et al. (2019)

**Table 11.** Representative studies of sentiment analysis

Research topic	Methods	Data set	Case study	Authors and year
Happiness	Maximum entropy classifier	Twitter	London, UK	Quercia et al. (2012)
	Language Assessment by Mechanical Turk word list	Twitter	US	Mitchell et al. (2013)
	Language Assessment by Mechanical Turk word list	Twitter	US	Frank et al. (2013)
	Multinomial Naïve Bayes classifier	Twitter	New York City	Li et al. (2017)
	Unigram-based sentiment analysis	Twitter	London, UK	Guo et al. (2016)
Smart governance	AFINN dictionary	Twitter	US	Hollander and Hartt (2018)

years by ranking the total number of photographs taken by tourists. It also provided a novel method in terms of defining the home country of a user based on the photo numbers in different locations.

### Social Demographics

Although the total size of social media is massive, its users are still sample data rather than being the representative of the entire population. To understand the sociodemographic background of social media creators, Li et al. (2013) detected the home location of users based on their time lines and linked social media with sociodemographic characteristics from census data by location. However, those linked to sociodemographic features are just collective features around a census unit. To detect features such as gender, age, and ethnicity at the individual level, a group of researchers involved in name analysis (see Table 10). Longley et al. (2015) emphasized the relationship between demographic features and the characterization of forename–surname pairs and applied it to demographic classification. In line with this, both Hofer et al. (2015) and Luo et al. (2016) conducted text mining to explore the demographic characteristics of social media users with their profile information and investigate the spatiotemporal characteristics of spatial patterns.

No matter how socioeconomic/demographic features are extracted, those enriched features from crowdsourced data help to understand the urban dynamics of concentration, dispersion, and segregation. In the reputed areas of segregation, Shelton et al. (2015) proposed a methodological framework to group users according to the neighborhood they visit frequently visited and explored the sociospatial mobilities between those groups in Louisville, Kentucky. Davis et al. (2019) used Yelp reviews to infer home or work location of reviewers through mining location-related context and links to census data to identify the segregation of urban consumption in New York City. Focusing on internal migration, Fiorio et al. (2017) mined long-period data from Twitter and explored the characteristics of demographic mobility, which helps to understand long-term migration. Because of the high-dimensionality of crowdsourced data, more studies focused on age, gender, sexuality, consumption power, economic status, and other identities are likely to be produced, which will be accurately extracted from unstructured information (Shelton et al. 2015).

### Sentiment Analysis

Content, as the central part of social media data, has become a growing research subject in recent years with applications of natural language processing techniques increasing in recent years (see Table 11). Because of NLP technologies, the contextual information can be extracted and analyzed for detecting the content of activities and public sentiment. The increasing interactions between citizens and online social media have provided an opportunity to conduct sentiment analysis for a better understanding of urban human geography. Quercia et al. (2012) detected the sentiment variance in different areas of London and found a positive correlation between sentiment and socioeconomic well-being. As one of the most comprehensive studies, Mitchell et al. (2013) investigated the correlations between sentimental expression (happiness) from Twitter and the emotional and demographic characteristics through 50 states in the US. This study provides a novel methodology by using the mechanical Turk word list that scores the average happiness of each word. Similarly, Frank et al. (2013) applied the same assessment tool to examine the relationship between happiness and the patterns of life in the US. To explore the relationship between sentiment and socio-economic parameters, Guo et al. (2016) conducted unigram-based sentiment analysis with geotagged tweets for different socio-demographic groups and found that the number of jobs, children, and transportation availability can well explain the sentiment variations. However, the content from social media is not just text but also emojis used to express users' emotions. To fill this gap, Li et al. (2017) applied the multinomial Naïve Bayes classifier to evaluate these special features. Realizing the contribution of sentiment analysis to smart governance, Hollander and Hartt (2018) introduced sentiment analysis to investigate the propensity of resident sentiment in declining cities around the US.

### Potential Challenges of Crowdsourced Data

From the aforementioned review, it can be found that crowdsourced data have become widely used in the field of urban activity analysis. Although advantages are highlighted in the previous studies described in this paper, crowdsourced data also bring challenges and difficulties that need to be clarified and tackled, such as the



challenges involved in data collection, data processing, and analysis. However, there are potential challenges when dealing with crowdsourced data which have been identified. The first challenge of crowdsourced data is concerning its representativeness, which is proposed frequently regarding biases (Huang and Wong 2016; Liu et al. 2016a). Although the proliferation of crowdsourced data is obvious, the users of this new form of data are relatively small by comparison with the overall population that needs to be studied (even smaller percentages are represented when studies focus on geotagged data). For instance, approximately 1% of tweets worldwide are geotagged, i.e., including the location information (Morstatter et al. 2013). The problem of representativeness consequently brings another problem relevant to statistical analysis where appropriate sampling is needed for valid inference. This is because the collection of crowdsourced data is automatically completed through APIs. Hence, some data may be oversampled or less sampled. Another concern and a source of biases is the reliability of crowdsourced data since the data are generated by individuals who may upload false or fake information on social media or collaborative websites. For example, the location tagged on social media can be *any place* in the world.

Another challenge is linked with data processing originating from multiple sources (Li et al. 2016b); according to the review, one trend that has been identified is that researchers are attempting to fuse and integrate different types of data together. However, the data formats and structure of metadata are different. Specifically, researchers need to convert files such as CSV, KML, KMZ, AML, TXT, and JSON into a uniform format for conducting analysis. Adding to this is the danger of merging data sets of unknown granularity levels of crowdsourced data during the data fusion. When it comes to analysis, Liu et al. (2016a) also point out that this new type of data is facing a methodological challenge, since traditional approaches are limited to fully leverage the value of crowdsourced data because of its volume, granularity, structure, and so forth.

Among the studies reviewed in this paper, there have been various attempts to eliminate the biases in the studies mentioned previously and they have leveraged crowdsourced data in specific fields with appropriate methods, whether engaging data fusion or applying mixed-method research. This paper would argue that crowdsourced mining data have provided an unexpected opportunity to produce novel and meaningful research regarding urban activity.

## Summary

This paper conducted a systematic review of studies in crowdsourced data mining for urban activity analysis. While there is no standard definition of crowdsourced data (Crooks et al. 2015), they can be explained as types of data that are collected from the *crowd* actively and passively (contributed by the user based on the terms of services) through the interaction between citizens and ICT-support services. Following the coordinated bottom-up process, crowdsourced data contain rich spatial, temporal, sociodemographic, and perception information related to urban activity, providing opportunities to get insights into urban dynamics from a perspective of the public. In the era of big data, crowdsourced data have advantages due to the massive volume of data, available access, fine granularity, real time, and high frequency. Given these characteristics, there is an increasing number of studies which vary in nature and scope that conduct urban activity analysis by using the main crowdsourced data sources, social media, POIs data, and collaborative websites, with different content such as text,

images, tags, profile, and so forth, and each data source has its advantage in specific domains.

This review highlighted the application of crowdsourced data on spatial analysis, including mobility patterns, functional areas, and event detection, with comprehensive studies. The high-volume spatial-temporal information provides chances for mobility analysis exploring dynamic flows of urban activity rather than static distribution. In other words, the content of crowdsourced data is used to identify the purpose of movement and functional areas, which leads to activity-based analysis. Other contents such as text, tags, and images provide crowd-based information for event detection and management. In addition, this review examined the application of sociodemographic and perception analysis and states the possibility of crowdsourced data mining. Three main fields, city attractiveness, demographic characteristics, and sentiment analysis, are identified. By reviewing the various applications listed previously, it was found that crowdsourced data support the shift from static analysis to human dynamic analysis in the field of urban studies. This also provides building blocks for real-time modeling and dynamic simulation in the future.

Potential challenges, such as biases of crowdsourced data, are mentioned at the end of this review. Problems in data collection, data processing, and analysis, i.e., representativeness, coverage bias, and heterogeneity of data frame, should be realized by researchers. These also need to be tackled through eliminating irrelevant content, fusing with multisource data, introducing algorithms of data cleaning, and integrating both qualitative and quantitative methods. While there are concerns and challenges about crowdsourced data, it is important to value how such new forms of data can be explored and leveraged for revealing spatial, temporal, sociodemographic, and perception characteristics of urban activity and realize that a new data-driven urban analysis, involving GIScience, human geography, urban studies, and data science has been developed during the era of the digital data revolution.

## Data Availability Statement

Some or all data, models, or code generated or used during the study are available from the corresponding author by request, such as the network file of cocitation of authors and the network file of cocitation of references.

## Acknowledgment

This research is funded by a scholarship from the China Scholarship Council (CSC No. 201808060346). We thank the anonymous reviewers for their many insightful comments and suggestions.

## References

- Aadland, M., C. Farah, and K. Magee. 2016. "μ-shapes: Delineating urban neighborhoods using volunteered geographic information." *J. Spatial Inf. Sci.* 12 (2016): 29–43.
- Abbasi, O. R., A. A. Alesheikh, and M. Sharif. 2017. "Ranking the city: The role of location-based social media check-ins in collective human mobility prediction." *ISPRS Int. J. Geo-Inf.* 6 (5): 136.
- Ashkezari-Toussi, S., M. Kamel, and H. Sadoghi-Yazdi. 2019. "Emotional maps based on social networks data to analyze cities emotional structure and measure their emotional similarity." *Cities* 86: 113–124.
- Barbosa, H., M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. 2018. "Human mobility: Models and applications." *Phys. Rep.* 734: 1–74.

- Batty, M. 2012. "Smart cities, big data." *Environ. Plann. B: Plann. Des.* 39 (2): 191–193.
- Batty, M. 2013. "Big data, smart cities and city planning." *Dialogues Hum. Geogr.* 3 (3): 274–279.
- Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet allocation." *J. Mach. Learn. Res.* 3 (Jan): 993–1022.
- Bocconi, S., A. Bozzon, A. Psyllidis, C. Titos Bolivar, and G.-J. Houben. 2015. "Social glass: A platform for urban analytics and decision-making through heterogeneous social data." In *Proc., 24th Int. Conf. on World Wide Web*, 175–178. New York: ACM.
- CASA (UCL Centre for Advanced Spatial Analysis). 2018. "Colouring London." Accessed July 10, 2019 <https://www.pages.colouring.london>.
- Chen, L., and A. Roy. 2009. "Event detection from Flickr data through wavelet-based spatial analysis." In *Proc., 18th ACM Conf. on Information and Knowledge Management*, 523–532. New York: ACM.
- Chen, W., Q. Gao, and H.-G. Xiong. 2016. "Uncovering urban mobility patterns and impact of spatial distribution of places on movements." *Int. J. Mod. Phys. C* 28 (01): 1750004.
- Chen, Y., X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, and F. Pei. 2017. "Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method." *Landscape Urban Plann.* 160: 48–60.
- Cheng, Z., J. Caverlee, K. Lee, and D. Z. Sui. 2011. "Exploring millions of footprints in location sharing services." In *Proc., 5th Int. AAAI Conf. on Weblogs and Social Media*, 81–88. Menlo Park, CA: AAAI.
- Crampton, J. W., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, and M. Zook. 2013. "Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb." *Cartography Geogr. Inf. Sci.* 40 (2): 130–139.
- Cranshaw, J., R. Schwartz, J. I. Hong, and N. Sadeh. 2012. "The livelihoods project: Utilizing social media to understand the dynamics of a city." In *Proc., 6th Int. AAAI Conf. on Weblogs and Social Media*, 58–65. Menlo Park, CA: AAAI.
- Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski. 2013. "#Earthquake: Twitter as a distributed sensor system." *Trans. GIS* 17 (1): 124–147.
- Crooks, A., D. Pfoser, A. Jenkins, A. Croitoru, A. Stefanidis, D. Smith, S. Karagiorgou, A. Efentakis, and G. Lamprianidis. 2015. "Crowdsourcing urban form and function." *Int. J. Geogr. Inf. Sci.* 29 (5): 720–741.
- Daggitt, M. L., A. Noulas, B. Shaw, and C. Mascolo. 2016. "Tracking urban activity growth globally with big location data." *R. Soc. Open Sci.* 3 (4): 150688.
- Davis, D. R., J. I. Dingel, J. Monras, and E. Morales. 2019. "How segregated is urban consumption?" *J. Politic. Econ.* 127, 1684–1738.
- De Albuquerque, J. P., B. Herfort, A. Brenning, and A. Zipf. 2015. "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management." *Int. J. Geogr. Inf. Sci.* 29 (4): 667–689.
- Deng, Y., J. Liu, Y. Liu, and A. Luo. 2019. "Detecting urban polycentric structure from POI data." *ISPRS Int. J. Geo-Inf.* 8 (6): 283.
- Dubey, A., N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo. 2016. "Deep learning the city: Quantifying urban perception at a global scale." In *Computer Vision—ECCV 2016*, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling, 196–212. Cham: Springer.
- Elwood, S., M. F. Goodchild, and D. Z. Sui. 2012. "Researching volunteered geographic information: Spatial data, geographic research, and new social practice." *Ann. Assoc. Am. Geogr.* 102 (3): 571–590.
- Estima, J., and M. Painho. 2015. "Investigating the potential of OpenStreetMap for land use/land cover production: A case study for continental Portugal." In *Openstreetmap in GIScience: Experiences, research, and applications*, Lecture Notes in Geoinformation and Cartography, edited by J. Jokar Arsanjani, A. Zipf, P. Mooney, and M. Helbich, 273–293. Cham: Springer.
- Feick, R., and C. Robertson. 2015. "A multi-scale approach to exploring urban places in geotagged photographs." *Comput. Environ. Urban Syst.* 53: 96–109.
- Fiorio, L., E. Zagheni, G. Abel, I. Weber, J. Cai, and G. Vinué. 2017. "Using twitter data to estimate the relationships between short-term mobility and long-term migration." In *Proc., WebSci—ACM Web Sci. Conf.*, 103–110. Troy, NY: ACM.
- Foursquare Labs. (2013). "The last three months on Foursquare." *Foursquare*. Accessed October 18, 2018. <https://foursquare.com/infographics/500million>.
- Frank, M. R., L. Mitchell, P. S. Dodds, and C. M. Danforth. 2013. "Happiness and the patterns of life: A study of geolocated tweets." *Sci. Rep.* 3 (1): 2625.
- French, S. P., C. Barchers, and W. Zhang. 2017. "How should urban planners be trained to handle big data?" In *Seeing cities through Big data: Research, methods and applications in urban informatics*, Springer Geography, edited by P. Thakuriah, N. Tilahun and M. Zellner, 209–217. Cham: Springer.
- Frias-Martinez, V., and E. Frias-Martinez. 2014. "Spectral clustering for sensing urban land use using Twitter activity." *Eng. Appl. Artif. Intell.* 35: 237–245.
- Gabrielli, S., P. Forbes, A. Jylhä, S. Wells, M. Sirén, S. Hemminki, P. Nurmi, R. Maimone, J. Masthoff, and G. Jacucci. 2014. "Design challenges in motivating change for sustainable urban mobility." *Comput. Hum. Behav.* 41: 416–423.
- Gao, S., K. Janowicz, and H. Couclelis. 2017. "Extracting urban functional regions from points of interest and human activities on location-based social networks." *Trans. GIS* 21 (3): 446–467.
- Garcia-Molina, H., M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios. 2016. "Challenges in data crowdsourcing." *IEEE Trans. Knowl. Data Eng.* 28 (4): 901–911.
- Garcia-Palomares, J. C., J. Gutiérrez, and C. Mínguez. 2015. "Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS." *Appl. Geogr.* 63: 408–417.
- Garcia-Palomares, J. C., M. H. Salas-Olmedo, B. Moya-Gómez, A. Condego-Melhorado, and J. Gutiérrez. 2018. "City dynamics through Twitter: Relationships between land use and spatiotemporal demographics." *Cities* 72: 310–319.
- Girardin, F., A. Vaccari, A. Gerber, and A. Biderman. 2009. "Quantifying urban attractiveness from the distribution and density of digital footprints." *Int. J. Spatial Data Infrastruct. Res.* 4: 26.
- González, M. C., C. A. Hidalgo, and A.-L. Barabási. 2008. "Understanding individual human mobility patterns." *Nature* 453 (7196): 779–782.
- Goodchild, M. F. 2007. "Citizens as sensors: The world of volunteered geography." *GeoJournal* 69 (4): 211–221.
- Granell, C., and F. O. Ostermann. 2016. "Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management." *Comput. Environ. Urban Syst.* 59: 231–243.
- Gray, S., R. Milton, and A. Hudson-Smith. 2015. "Advances in crowdsourcing: Surveys, social media and geospatial analysis: Towards a big data toolkit." In *Advances in crowdsourcing*, edited by F. J. Garrigos-Simon, I. Gil-Pechuán, and S. Estelles-Miguel, 163–179. Cham: Springer.
- Guo, W., N. Gupta, K. Pogrebna, and S. Jarvis. 2016. "Understanding happiness in cities using twitter: Jobs, children, and transport." In *Proc., IEEE Int. Smart Cities Conf.: Improv. Citizens Qual. Life (ISC2)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Haklay, M. 2010. "How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets." *Environ. Plann. B: Plann. Des.* 37 (4): 682–703.
- Harvey, F. 2013. "To volunteer or not to contribute locational information? Towards truth in labeling for crowdsourced geographic information." In *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice*, edited by D. Sui, S. Elwood, and M. Goodchild, 31–42. Dordrecht, Netherlands: Springer.
- Hasan, S., C. M. Schneider, S. V. Ukkusuri, and M. C. González. 2013. "Spatiotemporal patterns of urban human mobility." *J. Stat. Phys.* 151 (1–2): 304–318.
- Hasan, S., and S. V. Ukkusuri. 2014. "Urban activity pattern classification using topic models from online geo-location data." *Transp. Res. Part C—Emerg. Technol.* 44: 363–381.
- Hasan, S., and S. V. Ukkusuri. 2015. "Location contexts of user check-ins to model urban geo life-style patterns." *PLoS One* 10 (5): e0124819.
- Hawelka, B., I. Sitko, E. Beinart, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. 2014. "Geo-located Twitter as proxy for global mobility patterns." *Cartography Geogr. Inf. Sci.* 41 (3): 260–271.

- Hecht, B., and M. Stephens. 2014. "A tale of cities: Urban biases in volunteered geographic information." In *Proc., 8th Int. Conf. on Weblogs and Social Media (ICWSM)*, 197–205. Menlo Park, CA: AAAI.
- Hofer, B., T. J. Lampoltshammer, and M. Belgiu. 2015. "Demography of twitter users in the city of London: An exploratory spatial data analysis approach." In *Modern trends in cartography*, edited by J. Brus, A. Vondrakova, and V. Vozenilek, 199–211. Cham: Springer.
- Hollander, J. B., and M. D. Hartt. 2018. "Big data and shrinking cities: How Twitter can help determine urban sentiments." In *Big data for regional science*, edited by L. A. Schintler, and Z. Chen, 265–273. Abingdon, UK: Routledge.
- Hollenstein, L., and R. S. Purves. 2010. "Exploring place through user-generated content: Using Flickr tags to describe city cores." *J. Spatial Inf. Sci.* 1 (2010): 21–48.
- Howe, J. 2006. "The rise of crowdsourcing." *Wired Mag.* 14 (6): 1–4.
- Huang, L., Q. Li, and Y. Yue. 2010. "Activity identification from GPS trajectories using spatial temporal POIs attractiveness." In *Proc., 2nd ACM SIGSPATIAL Int. Workshop on Location Based Social Networks*, 27–30. New York: ACM.
- Huang, Q., and D. W. S. Wong. 2016. "Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us?" *Int. J. Geogr. Inf. Sci.* 30 (9): 1873–1898.
- Internet Live Statistics. (2016). "Twitter usage statistics." Accessed April 5, 2019. <http://www.internetlivestats.com/twitter-statistics/>.
- Jiang, S., A. Alves, F. Rodrigues, J. Ferreira, and F. C. Pereira. 2015. "Mining point-of-interest data from social networks for urban land use classification and disaggregation." *Comput. Environ. Urban Syst.* 53: 36–46.
- Jin, X., Y. Long, W. Sun, Y. Lu, X. Yang, and J. Tang. 2017. "Evaluating cities' vitality and identifying ghost cities in China with emerging geographical data." *Cities* 63: 98–109.
- Kim, J., and M. Hastak. 2018. "Social network analysis: Characteristics of online social networks after a disaster." *Int. J. Inf. Manage.* 38 (1): 86–96.
- Kitchin, R. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Thousand Oaks, CA: SAGE.
- Kitchin, R., T. P. Lauriault, and G. Mcardle. 2017a. *Data and the city*. Abingdon, UK: Routledge.
- Kitchin, R., T. P. Lauriault, and M. W. Wilson. 2017b. *Understanding spatial media*. Thousand Oaks, CA: SAGE.
- Lansley, G., and P. A. Longley. 2016. "The geography of Twitter topics in London." *Comput. Environ. Urban Syst.* 58: 85–96.
- Lee, R., and K. Sumiya. 2010. "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection." In *Proc., 2nd ACM SIGSPATIAL Int. Workshop on Location Based Social Networks*, 1–10. New York: ACM.
- Lenormand, M., T. Louail, O. G. Cantu-Ros, M. Picornell, R. Herranz, J. Murillo Arias, M. Barthelemy, M. San Miguel, and J. J. Ramasco. 2015. "Influence of sociodemographics on human mobility." *Sci. Rep.* 5: 10075.
- Li, D., X. Zhou, and M. Wang. 2018. "Analyzing and visualizing the spatial interactions between tourists and locals: A flickr study in ten US cities." *Cities* 74: 249–258.
- Li, L., M. F. Goodchild, and B. Xu. 2013. "Spatial, temporal, and socioeconomic patterns in the Use of twitter and flickr." *Cartography Geogr. Inf. Sci.* 40 (2): 61–77.
- Li, M., E. Ch'ng, A. Chong, and S. See. 2017. "The new eye of smart city: Novel citizen sentiment analysis in twitter." In *Proc., Int. Conf. Audio, Lang. Image Process (ICALIP)*, edited by F.-L. Luo, X. Yu, and W. Wan, 557–562. Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Li, M., Z. Shen, and X. Hao. 2016a. "Revealing the relationship between spatio-temporal distribution of population and urban function with social media data." *GeoJournal* 81, 919–935.
- Li, S., et al. 2016b. "Geospatial big data handling theory and methods: A review and research challenges." *ISPRS J. Photogramm. Remote Sens.* 115: 119–133.
- Liang, X., J. Zhao, and K. Xu. 2015. "A general law of human mobility." *Sci. China Inf. Sci.* 58 (10): 1–14.
- Liu, J., J. Li, W. Li, and J. Wu. 2016a. "Rethinking big data: A review on the data quality and usage issues." Supplement, *ISPRS J. Photogramm. Remote Sens.* 115(SC): 134–142.
- Liu, L., B. Zhou, J. Zhao, and B. D. Ryan. 2016b. "C-IMAGE: City cognitive mapping through geo-tagged photos." *GeoJournal* 81 (6): 817–861.
- Liu, X., J. He, Y. Yao, J. Zhang, H. Liang, H. Wang, and Y. Hong. 2017. "Classifying urban land use by integrating remote sensing and social media data." *Int. J. Geogr. Inf. Sci.* 31 (8): 1675–1696.
- Liu, X., and Y. Long. 2016. "Automated identification and characterization of parcels with OpenStreetMap and points of interest." *Environ. Plann. B: Plann. Des.* 43 (2): 341–360.
- Liu, Y., X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi. 2015. "Social sensing: A new approach to understanding our socioeconomic environments." *Ann. Assoc. Am. Geogr.* 105 (3): 512–530.
- Liu, Y., Z. Sui, C. Kang, and Y. Gao. 2014. "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data." *PLoS One* 9 (1): e86026.
- Long, Y., H. Han, Y. Tu, and X. Shu. 2015. "Evaluating the effectiveness of urban growth boundaries using human mobility and activity records." *Cities* 46: 76–84.
- Long, Y., and L. Liu. 2016. "Transformations of urban studies and planning in the big/open data era: A review." *Int. J. Image Data Fusion* 7 (4): 295–308.
- Longley, P. A., M. Adnan, and G. Lansley. 2015. "The geotemporal demographics of Twitter usage." *Environ. Plann. A* 47 (2): 465–484.
- Luo, F., G. Cao, K. Mulligan, and X. Li. 2016. "Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago." *Appl. Geogr.* 70: 11–25.
- Martin, A., A. B. A. Julián, and F. Cos-Gayón. 2019. "Analysis of Twitter messages using big data tools to evaluate and locate the activity in the city of Valencia (Spain)." *Cities* 86: 37–50.
- Miller, H. J., and M. F. Goodchild. 2015. "Data-driven geography." *GeoJournal* 80 (4): 449–461.
- Mitchell, L., M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. 2013. "The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place." *PLoS One* 8 (5): e64417.
- Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose." Preprint, submitted June 21, 2013. <http://arXiv.org/abs/1306.5204v1>.
- Noulas, A., C. Mascolo, and E. Frias-Martinez. 2013. "Exploiting four-square and cellular data to infer user activity in urban environments." In *Proc., 2013 IEEE 14th Int. Conf. on Mobile Data Management*, 167–176. Piscataway, NJ: IEEE.
- Noulas, A., S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. 2012. "A tale of many cities: Universal patterns in human urban mobility." *PLoS One* 7 (5): e37027.
- Paldino, S., I. Bojic, S. Sobolevsky, C. Ratti, and M. C. Gonzalez. 2015. "Urban magnetism through the lens of geo-tagged photography." *EPJ Data Sci.* 4 (1): 5.
- Pan, B., Y. Zheng, D. Wilkie, and C. Shahabi. 2013. "Crowd sensing of traffic anomalies based on human mobility and social media." In *Proc., 21st ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, 334–343. New York: ACM.
- Pan, H., B. Deal, Y. Chen, and G. Hewings. 2018. "A reassessment of urban structure and land-use patterns: Distance to CBD or network-based?—Evidence from Chicago." *Reg. Sci. Urban Econ.* 70: 215–228.
- Pouke, M., J. Goncalves, D. Ferreira, and V. Kostakos. 2016. "Practical simulation of virtual crowds using points of interest." *Comput. Environ. Urban Syst.* 57: 118–129.
- Quercia, D., J. Ellis, L. Capra, and J. Crowcroft. 2012. "Tracking 'gross community happiness' from tweets." In *Proc., ACM 2012 Conf. on Computer Supported Cooperative Work (CSCW '12)*, 965. Seattle, WA: ACM.
- Quercia, D., and D. Saez. 2014. "Mining urban deprivation from four-square: Implicit crowdsourcing of city land use." *IEEE Pervasive Comput.* 13 (2): 30–36.



- Rae, A. (2016). "Crowdsourced city boundaries." *Stats, Maps n Pix*. Accessed July 10, 2019. <http://www.statmapsnpix.com/2016/10/crowdsourced-city-boundaries.html>.
- Ratti, C., D. Frenchman, R. M. Pulselli, and S. Williams. 2006. "Mobile landscapes: Using location data from cell phones for urban analysis." *Environ. Plann. B: Plann. Des.* 33 (5): 727–748.
- Roberts, H., B. Resch, J. Sadler, L. Chapman, A. Petutschnig, and S. Zimmer. 2018. "Investigating the emotional responses of individuals to urban green space using twitter data: A critical comparison of three different methods of sentiment analysis." *Urban Plann.* 3 (1): 21–33.
- Sakaki, T., M. Okazaki, and Y. Matsuo. 2010. "Earthquake shakes twitter users: Real-time event detection by social sensors." In *Proc., 19th Int. Conf. on World Wide Web*, 851–860. New York: ACM.
- Salas-Olmedo, M. H., B. Moya-Gómez, J. C. García-Palomares, and J. Gutiérrez. 2018. "Tourists' digital footprint in cities: Comparing big data sources." *Tourism Manage.* 66: 13–25.
- Schlesinger, J. 2015. "Using crowd-sourced data to quantify the complex urban fabric—OpenStreetMap and the urban-rural index." In *Openstreetmap in GIScience: Experiences, research, and applications*. Lecture Notes in Geoinformation and Cartography, edited by J. Jokar Arsanjani, A. Zipf, P. Mooney, and M. Helbich, 295–315. Cham: Springer.
- See L et al. 2016. "Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information." *ISPRS Int. J. Geo-Inf.* 5 (5): 55.
- Shelton, T., A. Poorthuis, and M. Zook. 2015. "Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information." *Landscape Urban Plann.* 142 (SI): 198–211.
- Sobolevsky, S., I. Bojic, A. Belyi, I. Sitko, B. Hawelka, J. M. Arias, and C. Ratti. 2015. "Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity." In *Proc., 2015 IEEE Int. Congress on Big Data (BigData Congress)*, 600–607. Piscataway, NJ: IEEE.
- Song, Y., Y. Long, P. Wu, and X. Wang. 2018. "Are all cities with similar urban form or not? Redefining cities with ubiquitous points of interest and evaluating them with indicators at city and block levels in China." *Int. J. Geogr. Inf. Sci.* 32 (12): 2447–2476.
- Steiger, E., J. P. Albuquerque, and A. Zipf. 2015. "An advanced systematic literature review on spatiotemporal analyses of Twitter data." *Trans. GIS* 19 (6): 809–834.
- Steiger, E., B. Resch, and A. Zipf. 2016. "Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks." *Int. J. Geogr. Inf. Sci.* 30 (9): 1694–1716.
- Sui, D. Z., S. Elwood, and M. F. Goodchild. 2013. *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice*. New York: Springer.
- Sun, Y., and M. Li. 2015. "Investigation of travel and activity patterns using location-based social network data: A case study of active mobile social media users." *ISPRS Int. J. Geo-Inf.* 4 (3): 1512–1529.
- Thakuriah, P. V., N. Tilahun, and M. Zellner. 2016. *Seeing cities through Big data: Research, methods and applications in urban informatics*. Cham: Springer.
- Thatcher, J, A Shears, and J Eckert. 2018. *Thinking Big data in geography: New regimes, new research*. Lincoln, NE: Univ. of Nebraska Press.
- Van Eck, N. J., and L. Waltman. 2011. "Text mining and visualization using VOSviewer." Preprint, submitted September 9, 2011. <http://arxiv.org/abs/1109.2058v1>.
- Walther, M., and M. Kaiser. 2013. "Geo-spatial event detection in the twitter stream." In *European conf. on information retrieval*, Lecture Notes in Computer Science, edited by P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, 356–367. Berlin: Springer.
- Wang, Y., T. Wang, M.-H. Tsou, H. Li, W. Jiang, and F. Guo. 2016. "Mapping dynamic urban land use patterns with crowdsourced geo-tagged social media (Sina-Weibo) and commercial points of interest collections in Beijing, China." *Sustainability* 8 (11): 1202.
- White, H. D., and B. C. Griffith. 1981. "Author cocitation: A literature measure of intellectual structure." *J. Am. Soc. Inf. Sci.* 32 (3): 163–171.
- Wu, C., X. Ye, F. Ren, and Q. Du. 2018. "Check-in behaviour and spatio-temporal vibrancy: An exploratory analysis in shenzhen, China." *Cities* 77: 104–116.
- Wu, L., Y. Zhi, Z. Sui, and Y. Liu. 2014. "Intra-Urban human mobility and activity transition: Evidence from social media check-in data." *PLoS One* 9 (5): e97010.
- Xia, C., J. Hu, Y. Zhu, and M. Naaman. 2015. "What is new in our city? A framework for event extraction using social media posts." In *Proc., 19th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, 16–32. Cham: Springer.
- Xu, Z., Y. Liu, J. Xuan, H. Chen, and L. Mei. 2017. "Crowdsourcing based social media data analysis of urban emergency events." *Multimedia Tools Appl.* 76 (9): 11567–11584.
- Xu, Z., H. Zhang, V. Sugumaran, K.-K. R. Choo, L. Mei, and Y. Zhu. 2016. "Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media." *EURASIP J. Wireless Commun. Networking* 2016 (1): 44.
- Yang, Z., D. Lian, N. J. Yuan, X. Xie, Y. Rui, and T. Zhou. 2017. "Indigenization of urban mobility." *Physica A* 469: 232–243.
- Yao, Y., X. Li, X. Liu, P. Liu, Z. Liang, J. Zhang, and K. Mai. 2017. "Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model." *Int. J. Geogr. Inf. Sci.* 31 (4): 825–848.
- Yu, Z., H. Xu, Z. Yang, and B. Guo. 2016. "Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints." *IEEE Trans. Hum.-Mach. Syst.* 46 (1): 151–158.
- Yuan, J., Y. Zheng, and X. Xie. 2012. "Discovering regions of different functions in a city using human mobility and pois." In *Proc., 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 186–194. New York: ACM.
- Yue, Y., Y. Zhuang, A. G. O. Yeh, J.-Y. Xie, C.-L. Ma, and Q.-Q. Li. 2017. "Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy." *Int. J. Geogr. Inf. Sci.* 31 (4): 658–675.
- Zhai, W., X. Bai, Y. Shi, Y. Han, Z.-R. Peng, and C. Gu. 2019. "Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs." *Comput. Environ. Urban Syst.* 74: 1–12.
- Zhan, X., S. V. Ukkusuri, and F. Zhu. 2014. "Inferring urban land use using large-scale social media check-in data." *Netw. Spatial Econ.* 14 (3–4): 647–667.
- Zhi, Y., H. Li, D. Wang, M. Deng, S. Wang, J. Gao, Z. Duan, and Y. Liu. 2016. "Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data." *Geo-Spatial Inf. Sci.* 19 (2): 94–105.
- Zhou, X., and L. Zhang. 2016. "Crowdsourcing functions of the living city from Twitter and Foursquare data." *Cartography Geogr. Inf. Sci.* 43 (5): 393–404.