Travel Behaviour and Society 20 (2020) 192–212

Contents lists available at ScienceDirect

Travel Behaviour and Society

journal homepage: www.elsevier.com/locate/tbs

Crowd characterization for crowd management using social media data in city events

Vincent X. Gong[a,b,*], Winnie Daamen[a], Alessandro Bozzon[b], Serge P. Hoogendoorn[a]

[a] Dept. Transport & Planning, Delft University of Technology, Faculty of Civil Engineering and Geosciences, Stevinweg 1, 2628 CN Delft, The Netherlands
[b] Dept. Software Technology, Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands

ARTICLE INFO

Keywords:
Social media
Crowd management
Pedestrian behaviour
Crowds demographics

ABSTRACT

Large-scale events are becoming more frequent in contemporary cities, increasing the need for novel methods and tools that can provide relevant stakeholders with quantitative and qualitative insights about attendees' characteristics. In this work, we investigate how social media can be used to provide such insights. First, we screen a set of factors that characterize crowd behavior and introduce a set of proxies derived from social media data. We characterize the crowd in two city-scale events, Sail 2015 and King's Day 2016, analyzing several properties of their attendees, including demographics, city-role, crowd temporal distribution, social media post locations, Point of Interest (PoI.) preferences, and word use. We show that it is possible to characterize crowds in city-scale events using social media data, thus paving the way for new real-time applications on crowd monitoring and management for city-scale events.

1. Introduction

As cities compete for global importance and influence, city-scale public events are becoming an important ingredient to foster tourism and economic growth. Sports events, thematic exhibitions, and national celebrations are examples of city-scale events that take place in vast urban areas and attract large amounts of participants within short periods. The scale and intensity of these happenings demand technological solutions supporting stakeholders (e.g., event organizers, public and safety authorities, attendees) to monitor and manage the crowd.

These stakeholders aim to minimize the risk of incidents due to issues caused by external and internal threats. They usually apply predefined measures according to the qualitative interpretation of the crowd by police officers, stewards, or event organization employees.

As the efficiency and effectiveness of crowd management measures depend on pedestrian behavior (Still, 2000; Zomer et al., 2015), it is valuable for stakeholders to have information about the expected, and preferably actual, pedestrian behavior of the crowd. Pedestrian behavior is influenced by factors such as age, gender, and ethnicity (Martin, 2006). Insights into the distribution of these factors in an event's population can help to estimate and predicting crowd behavior, and as such, could be beneficial to crowd management.

However, information about these factors is difficult to capture. Traditionally, this information is manually sampled by stewards or staff members (Earl et al., 2004), a practice that is expensive and prone to biases. ICT solutions based on sensors (e.g., GPS, custom mobile apps) could provide spatio-temporal information (i.e., GPS coordinates and timestamps) that is useful to study crowd behavior (Jamil et al., 2015), but they are not broadly adopted, and might not provide demographic information. Camera sensors provide images or video clips that could be used to extract crowd features (Favaretto et al., 2016; Ryan et al., 2009), and detect crowd behavior (Wang et al., 2012; Zhan et al., 2006) through image recognition techniques. However, accessing the images or video recordings of the public area is computationally intensive, and often restricted due to privacy issues.

The advent of web-based technologies provides new social data sources that could be used to analyze and understand pedestrian behavior. Several platforms, such as Twitter, Instagram and Foursquare, are widely used. Social media content (e.g. text messages, images) is time-stamped and often geo-tagged, and it inherently containssni rich semantic information that could be used for characterizing the crowd from a pedestrian behavior perspective. For instance, the text content of posts sent by the crowd may indicate what the people are talking about, in order to see e.g. whether participants are enthusiastic about the event they are participating in or whether (security) issues are discussed. Likewise, the profiles of social media users can help to determine the crowd demographic characteristics. The rich semantic information makes social media a promising data source to provide

* Corresponding author at: Delft University of Technology, Mekelweg 5, 2628 CD Delft, The Netherlands.
E-mail address: X.Gong-1@tudelft.nl (V.X. Gong).

https://doi.org/10.1016/j.tbs.2020.03.011
Received 30 April 2019; Received in revised form 16 February 2020; Accepted 30 March 2020
Available online 13 April 2020

Table 1
Influencing factors with corresponding social media proxies.

Category	Factors	Social media proxy					
		Demographic	City-role	Crowd Temp. Dist.	Post position	PoI	Word use
Individual Characteristics	Demographic	x					
	Route familiarity		x				
	Perception of danger						x
	Type of destination					x	
Social Network	Household				x	x	
	Acquaintances				x		
	Neighborhood				x	x	
Trip characteristics	Trip purpose					x	
	Crowdedness						x
	Distance/ proximity				x		
	Capacity			x			
	Traffic volume			x			
Built environment	Type of area					x	
	Percentage of foreigners		x				
	Aesthetics						x
	Distance to nearest transit stop				x	x	
	Population density			x			
	Intersection density			x			
	Road density			x			

Crowd Temp. Dist.: Crowd temporal distribution.

information for crowd characterization in the city-scale events.

Previous works explored social media as a data source to analyze various aspects of human behavior and their characteristics for crowd management in the context of city events. Concerning human travel behaviors, Rashidi et al., 2017 explored the capacity of social media data for modeling travel behavior. Tyshchuk and Wallace (2018) explored a set of behaviors that are associated with a warning response process using social media. Roy et al. (2019) quantified and analyzed human mobility resilience to extreme events using geo-located social media. Krueger et al. (2019) proposed a visual analysis framework of city dynamics, including temporal patterns of visited places and citizens' mobility routines, using geo-located social media data. To explore the characteristics of human behavior, Abbasi et al., 2015 investigated a set of travel attributes that are extracted from social media data, such as trip purpose and activity location. Also, several studies are performed in the context of city events. Yang et al. (2019), Gao (2015), Hawelka et al. (2014) and Yang et al. (2019) use social media data collected in city events to investigate mobility issues. Cottrill et al. (2017) studied how attendees' behavior are affected in a large city event, in terms of providing and sharing transport-related information and responding to requests, based on social media. Pramanik et al. (2019); Hochmair et al. (2018), Balduini et al. (2014a) and Balduini et al. (2014b) proposed methods to provide real-time Point of Interest (PoI) recommendations in city events using social media. Alkhatib et al. (2019) proposed a framework for monitoring incidents during events in cities. Though the utility of social media data has been shown in urban application domains, no previous work aimed at characterizing the crowd of city-scale events, with a specific focus on crowd management. What is lacking is an in-depth understanding of which factors could be extracted from social media data, and which automatic user modeling techniques can provide an accurate and reliable estimation of such factors.

In this paper, we perform a study to show to what extent social media data could be used for characterizing crowds in city-scale events using factors for crowd management. First, we identify a set of factors that are relevant for pedestrian behavior analysis for crowd management and explore existing methods for extracting information about these factors from social media data. To showcase the application of these methods, we perform two case studies having different properties. In each case, we collect social media data from multiple platforms and extract the required information using SocialGlass (Bocconi et al., 2015), an integrated system for processing social media data. We then perform an exploratory analysis of these factors and correlate them with the corresponding event to check their accuracy and reliability. Discussions and conclusions are included at the end of the paper.

2. Crowd characterization

In this work, we seek a better understanding of how social media data can be used to support crowd management. To this end, we provide insights about factors that are known to influence pedestrian behavior. In this section, we first introduce a selection of factors that are relevant to pedestrian behavior analysis; then we describe how such factors could be calculated from social media data.

2.1. Characterization factors

Following the above discussion, criteria for selecting factors are:

(1). The factors should be identified as influencing the pedestrian behavior.
(2). The factors should be derived from social media.

As mentioned before, a set of factors has been discussed in (Martin, 2006) that affect pedestrian behavior. These factors can be classified into 6 categories, being Individual characteristics, Social network, Trip characteristics, Built environment, Destination environment and Physical environment. These factors with the corresponding social media proxies are listed in Table 1. These factors influence different types of pedestrian behavior, i.e. activity choice behavior, destination choice behavior, mode choice behavior, and route choice behavior (see (Wegener, 2004; Hoogendoorn and Bovy, 2005; Daamen, 2004) and Fig. 1). Obtaining information about these factors may help with understanding such types of pedestrian behavior and further support crowd management.

As indicated in the previous section, crowd managers usually apply predefined measures according to the information about these factors. This is implemented in a crowd management plan (Still, 2014; Tubbs

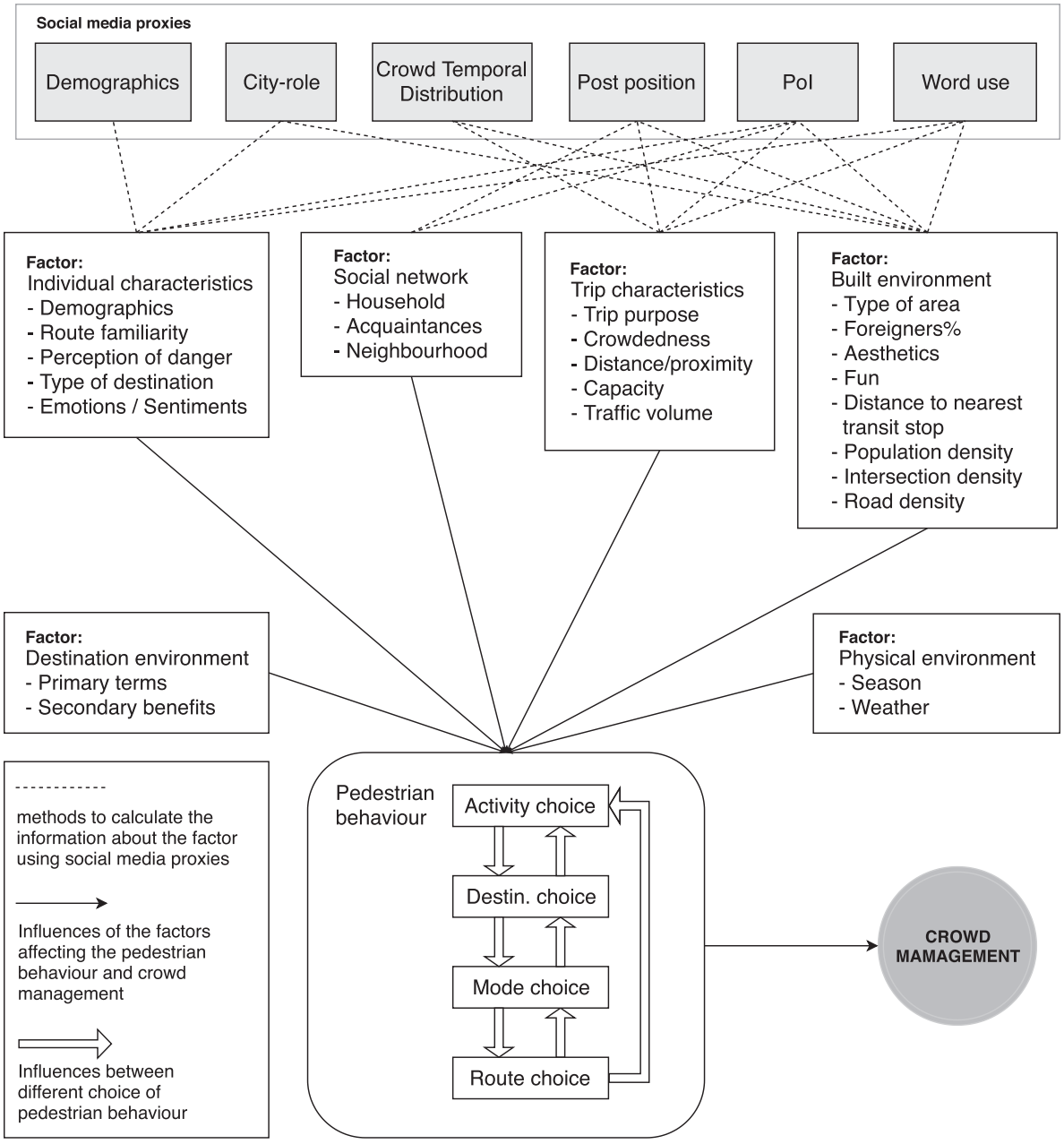


Fig. 1. Illustration of relationship between crowd management, pedestrian behavior, influencing factors, and social media proxies. The numbers in the brackets denote the references.

Table 2
An example of questions in crowd management plan and the social media proxies which can help answering these questions.

Questions	Social media proxies
What is the demographics composition of the participants?	Demographics
What is the percentage of people from other cities?	City role
What is the crowd density during the event?	Crowd temporal distribution
Where is the most crowded area?	Post position
What kinds of places are to be most visited by the crowd in different region?	PoI
What is the sentiment of the crowd?	Word use

and Meacham, 2007; Abbott and Geddie, 2000), in which a set of questions are to be answered. Answering these questions require qualitative and quantitative information about the crowd. Examples of these questions or required information in a crowd management plan are shown in Table 2.

In the following sections, we explain why those factors and the social media proxies are connected, and which methods are used to calculate these proxies from social media data.

2.2. Social media data analysis for crowd characterization

Among all popular social media platforms, we focus our studies on

three applications that provide data retrieval APIs, namely Twitter, Instagram and Foursquare. Twitter and Instagram provide posts and user profiles, while Foursquare provides Points of Interest (PoI) – the information about a location where people send posts. Twitter is a text-based social media platform, and one of the oldest social networking applications; Instagram is an image-based social media platform, which is particularly welcomed by female users (Yang et al., 2016; Gong, 2016). Data available for retrieval from such platforms include user profile information and submitted posts, their content, and time-stamp. A certain amount of posts contain coordinates where posts are sent, namely the geo-referenced posts, and the PoI information determined from geo-referenced posts. As city events take place at specific locations or areas, in this study we focus on social media data with geo-referenced posts. Based on the collected social media data, several proxies for crowd characterization factors can be calculated, namely demographic characteristics (i.e. age, gender), city-role, post spatio-temporal distribution, PoI, and word use. Each of them is addressed in the following subsections. The accuracy of these techniques is influenced by the amount and representativeness of information such as user profiles. Though social media is not used by everyone in the event, it could be considered as a partially representative sample of the crowd during events.

2.2.1. Demographic characteristics

Demographic characteristics, i.e. age (Berrigan and Troiano, 2002), gender (Berrigan and Troiano, 2002; Panter and Jones, 2010), have been identified as factors affecting pedestrian activity choice, destination choice, mode choice and route choice. This information could be derived from social media by approaches using text categorization (Peersman et al., 2011), first name (Lansley and Longley, 2016; Mislove et al., 2011), and profile picture (Bocconi et al., 2015; Longley et al., 2015). In our study, we use the user profile picture to determine user's age (Bocconi et al., 2015; Zhou et al., 2015; Psyllidis et al., 2015), and a multi-modal decision tree classifier (Yang et al., 2016; Titos Bolivar, 2014) combining the user's profile picture (Zhou et al., 2015) and the first name (Lansley and Longley, 2016) to detect user's gender information. A manual check with 628 labelled social media profiles performed by Yang et al., 2016 shows that both age and gender detection reach promise performance, i.e. 88% precision for age detection when faces are present, and 85% precision for gender detection.

People can be classified according to different indicators. One of the well-known ones is gender, i.e. male and female. Age is also known to influence behavior, often using four groups (Berrigan and Troiano, 2002). The range of each age group is defined considering social and physiological science (Al-Zahrani et al., 2003; Young et al., 1993) as follows:

- Young: user between 0 and 18,
- Young-adult: user between 18 and 30,
- Adult: user between 30 and 65,
- Old: user older than 65.

2.2.2. City-role

The city-role describes the relationship between the people and the city, i.e.:

- Resident: attendees living in the city of the event;
- Local traveler: attendees living in the same country, but in another city;
- Foreign traveler: attendees from a foreign country.

The percentage of foreigners (Kim et al., 2014; Rietveld and Daniel, 2004) and people's familiarity with a route (Kim et al., 2014) are identified as factors affecting mode choice behavior (Kim et al., 2014; Rietveld and Daniel, 2004) and route choice behavior (Kim et al., 2014), respectively. Information about these factors can be derived using social media by checking a user's home location through a recursion search method (Cheng et al., 2011; Titos Bolivar, 2014), which shows promise accuracy (covering about 0.004 square miles) according to the comparison (Yang et al., 2016).

2.2.3. Crowd temporal distribution

The temporal distribution of a crowd, i.e. the distribution of persons present at a certain area over time, is identified as a factor affecting destination choice (Han et al., 2010; Zahran et al., 2008), mode choice (Handy, 1996; Zahran et al., 2008; Guo, 2009; Rodríguez et al., 2009) and route choice (Zahran et al., 2008; Guo, 2009), as illustrated in Fig. 1. Calculating the temporal distribution of the crowd during an event requires information about the amount of people in an event area during a predefined period of time.

In social media, each post is sent with a timestamp. This information may be used to count the amount of posts sent by different people in a period of time. It is then used as a proxy for the temporal distribution of crowds (Yang et al., 2016; Gong, 2016; Titos Bolivar, 2014), which is temporally correlated with the estimated ground truth from sensor data according to a comparison (Gong et al., 2018).

2.2.4. Post position

Distance/proximity (Van der Waerden et al., 1998; Maley and Weinberger, 2011; Panter and Jones, 2010) is identified as a factor affecting all four pedestrian behaviors mentioned in Fig. 1. To calculate the distance, e.g. the distance between a pedestrian and a certain object in the area, having the position of the pedestrian is required.

In social media, the geo-referenced posts contain the coordinate of the location they have been sent. This position data can be a proxy to calculate distances (Yang et al., 2016; Gong, 2016; Titos Bolivar, 2014).

2.2.5. Points of Interest

Factors such as type of destination (Eash, 1999), diversity of land use (Rodríguez et al., 2009; Panter and Jones, 2010; McCormack and Shiell, 2011), and trip purpose (Handy, 1996) are identified as factor affecting destination choice (Eash, 1999), mode choice (Eash, 1999; Handy, 1996), and route choice (Rodríguez et al., 2009; Panter and Jones, 2010; McCormack and Shiell, 2011), respectively. These factors require information about a location with its functionality category as well as popularity, which can be provided by the Point of Interest (PoI), a particular location that someone may find useful or interesting, such as a hotel, a restaurant, or a bus station. A social media post sending from a PoI indicates a PoI has been visited by this user. With such information, we may extract the set of PoIs visited by people during an event, as well as PoI functionality categories and popularity. The destination of a pedestrian's trip as well as the trip purpose could be examples for which the data can be analyzed.

The PoI information can be derived from social media through various techniques, such as Natural Language Processing (Lingad et al., 2013), user relationship analysis (Davis et al., 2011), and the Venues Mapping method (Noulas et al., 2012). The Venues Mapping method proposed by Noulas et al., 2012 establishes a model to determine the venue visited by each user considering multiple aspects in their approach, i.e. popular places, similar places, users' preferences in selecting places, places visited by friends, and places in short distance. In our research, we employ the Venues Mapping method (Noulas et al., 2012) to get the PoI visited by social media users as it results in 5% to 18% improvement over other methods (Noulas et al., 2012). We record the top-level PoI category defined by Foursquare visited by social media users for analysis.

2.2.6. Word use

Influencing factors such as Crowdedness (Pratiwi et al., 2015; Duives et al., 2016), Aesthetics (Guo, 2009; Panter and Jones, 2010; McCormack and Shiell, 2011), Fun (Florez et al., 2014), and Perception of danger (Panter and Jones, 2010) affect mode choice and route

Table 3
The measurements of social media proxies to derive property of factors.

Aspect	Proxy	Measurement
Demographics	Gender	#male, #female, M/F
	Age	#young, #young-adult, #adult, #old, SD
City-role	City-role	#resident, #local_traveler, #foreign_traveler, SD, R/L
Crowd temp. dist.	Post amount	#GP of day, Max #GP and Time, Min #GP and Time
Position	Coordinates	latitude, longitude
PoIs	PoIs	#PoI_visit
	PoI category	#PoI functionality category
Word use	Text content	Word count

#male: number of people determined as male,
Crowd temp. dist.: Crowd temporal distribution.
M/F: the rate of Male with Female,
SD: Standard Deviation.
R/L: the rate of Resident with Local.
#GP: amount of Geo-posts.
Max #GP and Time: the max amount of Geo-posts, and the time of a period during which this amount is observed.

choice. These factors require information about a pedestrian's expressions and feelings. This information can be derived from social media data.

A social media post usually consists of a texture attribute which can be used to infer topics the people talk about, and their feelings. In this research, we visualize the frequently used words (word-cloud) sent by the crowd in order to provide such information, see also (Yang et al., 2016; Schwartz et al., 2013; Chen et al., 2014; Gong, 2016).

2.3. Summary

The sections above introduced a set of factors, about which information can be derived from social media data, the so-called proxies. We further described each proxy with properties and methods to calculate them. An overview of the measurements of the proxies is shown in Table 3. In the remainder of the paper, we will apply these techniques in two city events and compare and analyze the estimated information with the events programs.

3. Applying crowd characterization based on Social Media data in two city-scale events

In this section, we showcase how social media data (and related methods) can be used to characterize the crowd in two city-scale events by providing information about the factors described in the previous section. Furthermore, we relate the derived information with the event programs, to discuss its accuracy and reliability.

3.1. Case selection

We investigate two events that took place in Amsterdam, the Netherlands, respectively Sail 2015 (in the following referred to as Sail) and King's Day 2016 (Kingsday). We selected the two events for their similarities and their differences. On the one hand, these events have similar properties, being *city-scale*, and taking place in the same urban environment and *planned*, *temporally constrained*, and thoroughly organized (in contrast to seasonal events, such as Christmas shopping, or serendipitous events, like protests) and *popular* and *generalist*, as they attract large crowds with diverse demographics.

On the other hand, the two events also differ from each other in terms of duration, topic, crowd composition and event terrain. For instance, for duration, Sail lasts for 5 days, ending in a weekend, whereas Kingsday is a single-day event, and a public holiday. As to the topic, Sail being a naval event offering, for instance tall-ship exhibition, nautical history experience, fireworks show, while Kingsday is a recurrent national celebration, which offers a boat parade, free market and parties. As for the crowd composition, Sail is known to attract visitors from the whole world, while Kingsday is a national event. For the event terrain, Sail has activities centred around the IJhaven area (where ships docked), while Kingsday activities are scattered throughout the city.

To compare the analysis with the actual situation where quantitative ground truth existed, we also perform an analysis for two terrains (sub-area), based on the findings in our previous work (Gong et al., 2018), where the number of people calculated from social media in each terrain is temporally correlated with the estimated ground truth calculated using sensor data. The social media users in these two terrains, i.e. Javakade in Sail and Zuidplein in King's Day. The Javakade located on Java Island, directly faces the IJHaven, the bay area where the boats docked. This terrain is residential, with no recreational businesses. Areas separated by canals are connected by small pedestrian bridges, where several docked boats can be accessed during Sail event. The Zuidplein is the forecourt of the station Amsterdam Zuid, which is a popular pedestrian square connecting the station with the CBD area, and the Amsterdam OUD-Zuid. Around the square, there are various shops and restaurants, attracting a large amount of people during King's Day event. The number of users of which the demographic and city-role have been derived from these two terrains are listed in Table 4.

We selected these two events to compare the introduced crowd characterization for events with different fingerprints. The areas where these two events took place are shown in Fig. 2. Further details about the events are introduced in the following sub-sections.

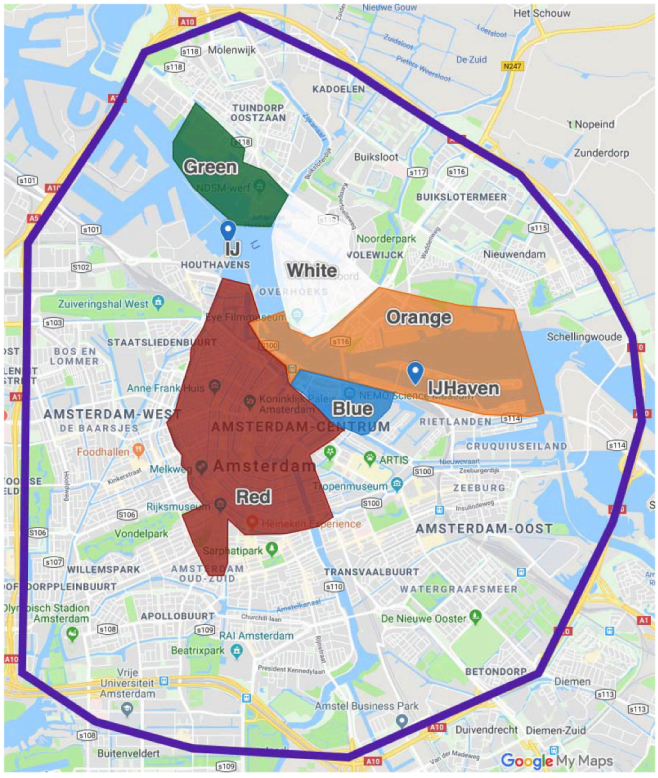
3.1.1. Case 1: Sail 2015

SAIL Amsterdam is a quinquennial maritime event in Amsterdam, the Netherlands. Tallships from all over the world come to the city to be visited and visitors join activities. It is the largest public event in the Netherlands: the 2015 edition of the event lasted 5 days, from August

Table 4
Number of users of which the demographic and city-role have been derived from social media in two terrains during the Sail and King's Day events, respectively.

	Terrain		Age				Gender			City-Role				
			Young	You. Adult	Adult	old	Sum	Male	Female	Sum	Resident	Loc. Tour	For. Tour	Sum
Sail	Twitter	Javakade	8.4%	39.6%	52.0%	0.0%	94	68.2%	31.8%	163	44.6%	33.6%	21.9%	187
	Instagram		49.5%	31.0%	0.2%	367	46.5%	53.5%	757	48.8%	21.9%	29.2%	1018	
King's Day	Twitter	Zuidplein	12.9%	43.9%	43.3%	0.0%	69	61.6%	38.4%	98	45.6%	15.2%	39.2%	191
	Instagram		49.8%	26.9%	0.0%	637	39.6%	60.4%	1032	44.0%	21.1%	35.0%	3965	

You. Adult: Young Adult.
Age for Young: 0–18, **Young Adult**: 18–30, **Adult**: 31–64, **Old**: 65+.
The scope of the terrains are illustrated in **Fig. 3** in "Using Social Media for Attendees Density Estimation in City-Scale Events".
The users in each terrain is identified using **speed- and flow-based** density estimation methods (K3/K4) in **Table 2** in "Using Social Media for Attendees Density Estimation in City-Scale Events".



(a) Area of two events that took place in Amsterdam. Activities during Kingsday took place in the whole city of Amsterdam (area bounded by dark blue line). The other 5 coloured areas are for Sail, i.e. Orange, White, Blue, Green and Red activity areas, the so-called Oceans. Marked locations are further explained in the case introduction and analysis.

(b) Sail 2015 event

(c) King's Day 2016 event

Fig. 2. The two events in Amsterdam selected in this study.

Table 5
Overview of the data collected during Sail 2015 and King's Day 2016 in Amsterdam. The number of Users, Geo-Posts, and PoIs are expressed in thousands.

	Sail 2015 Aug 19–23, 2015		King's Day 2016 Apr 26–28 2016	
	Twitter	Instagram	Twitter	Instagram
#User	2.8	27.3	1.6	28.5
#Geo-Posts	11.6	60.4	4.6	44.1
#PoIs	2.5	8.0	1.4	2.4

19 to 23, and attracted more than 2 million people. The exhibition included tallships and historical ships, as well as a large number of other boats. The official event area covers most of the city center, and was organized into five so-called oceans, each devoted to a theme.

The program of this event included sub-events spanning all five days. On August 19, all tallships sailed from the coast towards Amsterdam and docked in the IJHaven. During the following three days, the tallships were open for visits from 10AM till 11PM. A set of ship related activities took place around IJhaven attracting a huge amount of people who are interested in this topic. The ships departed again on August 23 in the closing SAIL-out event after a Sail Thank You parade. Every day, a firework show took place at IJhaven lasting for 15 min between 22:00 to 23:00.

3.1.2. Case 2: King's Day 2016

King's Day is a national holiday held each year in April, celebrating the birthday of King Willem-Alexander. In major cities in the Netherlands it is celebrated with joyful open air festivities. People join this yearly regular event with their families and friends. In 2016, the King's Day celebration attracted more than 1.5 million people in Amsterdam, including Dutch tourists and a huge amount of foreign tourists.

Though it is a one day public holiday, it is certainly not a day of rest. The celebrations started on the eve of King's Day - named as the King's Night. Parties, music, and carnival atmosphere continuing throughout the city until the end of the day. Following King's Night, the most interesting activity on King's Day in Amsterdam is the boat parade. From 1 pm, canals are packed with boat parties, during which the boats are sailing along the canals throughout the city with people enjoying drinking and celebrating wearing orange. Besides, several large museums are open for people who would like to experience culture and history.

3.2. Data collection

For each case, we collected geo-referenced social media data on the Twitter and Instagram platforms. The geo-referenced social media posts were mapped with PoIs from Foursquare. Then, we derived information about the crowd, including age, gender, city-role, crowd temporal distribution, post position, PoIs, and word use. We analyzed the derived information for each case, looking for meaningful relationships with the events' programs. We also compared the outcomes of the analysis of the two cases, highlighting similarities and differences.

The data is collected and derived using SocialGlass (Bocconi et al., 2015), an integrated system for crawling and processing social media data. First, we set up a crawling task with a duration (starting and ending date) and an event area (a bounding-box for Twitter, and multiple circles for Instagram) to crawl geo-referenced social media posts sent during an event, through queries on Twitter and Instagram. Second, we screen out unique users from the captured social media posts, as one user may send multiple posts. Third, we crawl user profile data on both platforms. Next, we crawl historical geo-referenced posts for each user on Twitter and Instagram, respectively. Further, we calculate the demographic, city-role and word use information for each user, and generating PoIs information through venue mapping algorithm. Finally, we export demographic, city-role, crowd temporal distribution, position, PoIs and word use information from the SocialGlass platform.

With regard to Sail 2015, the data set includes posts generated from August 19 until August 23. For King's Day 2016, we collected data from April 26 to April 28, to respectively cover celebration starting the night before (i.e. King's Night), King's Day itself, and the following day (to capture celebrations lasting throughout the night).

In the crowd temporal distribution analysis, we further include social media data sent seven days before each event and seven days after, in order to compare the pattern of the crowd distribution between event days and regular days, as well as to compare event days with regular week and weekend days. However, for the other analyses we only use event dates, i.e. August 19 to 23 for Sail and April 26 to 28 for King's Day, shown in Table 5.

From both events we collected more posts on Instagram than on Twitter; it is caused by, on the one hand, the scarcity of the geo-referenced tweets which only accounts for 1–2% of all tweets (Paule et al., 2019). We decide to obtain more geo-referenced tweets in our future work using techniques such as geolocalisation (Middleton et al., 2018; Paule et al., 2019). On the other hand, it suggests that Instagram, being

Table 6
Number of users of which the demographic and city-role have been derived from social media during the Sail and King's Day events.

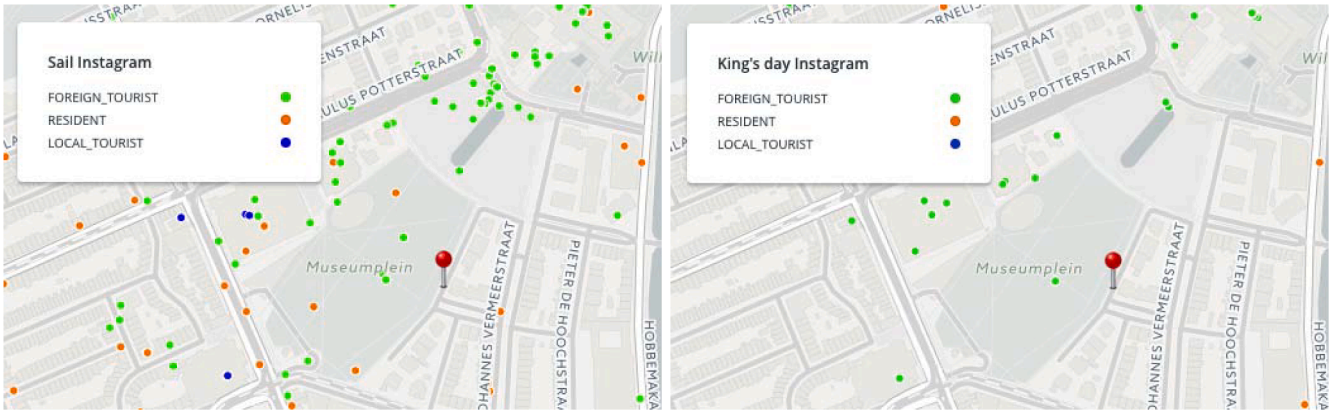
		Age					Gender			City-Role			
		Young	You. Adult	Adult	Old	Sum	Male	Female	Sum	Resident	Loc. Tour	For. Tour	Sum
Sail	Twitter	12.3%	42.7%	45.0%	0.0%	1225	56.7%	43.3%	2267	37.8%	21.3%	40.8%	2779
	Instagram	24.4%	49.8%	25.8%	0.1%	7130	42.0%	58.0%	19341	35.3%	14.7%	50.0%	26947
King's Day	Twitter	13.4%	44.1%	42.5%	0.0%	515	56.8%	43.2%	754	36.3%	14.1%	49.6%	1640
	Instagram	23.0%	44.6%	32.3%	1.0%	8747	42.1%	57.9%	8510	40.1%	14.6%	45.4%	7028

You. Adult: Young Adult. **Loc. Tour**: Local Tour. **For. Tour**: Foreign Tourist.
Age for Young: 0–18, **Young Adult**: 18–30, **Adult**: 31–64, **Old**: 65+.

Table 7
Number of users of which the demographic and city-role have been derived from social media during three sub-events of Sail.

		Age					Gender		
		Young	You. Adult	Adult	Old	Sum	Male	Female	Sum
Sail fireworks	Twitter	9.1%	42.7%	48.2%	0.0%	110	61.0%	39.0%	182
	Instagram	21.3%	50.5%	27.1%	1.1%	727	44.6%	55.4%	1403
Sail topic activities	Twitter	8.9%	38.3%	52.8%	0.0%	358	65.0%	35.0%	609
	Instagram	21.0%	47.7%	31.2%	0.1%	1745	46.1%	53.9%	3645
Sail parade	Twitter	6.3%	38.5%	55.2%	0.0%	96	59.8%	40.2%	132
	Instagram	20.3%	48.8%	30.9%	0.0%	602	43.1%	56.9%	954

You. Adult: Young Adult.
Age for Young: 0–18, **Young Adult**: 18–30, **Adult**: 31–64, **Old**: 65+.



(a) Sail, Instagram (b) Kingsday, Instagram

Fig. 3. The city-role distribution around the Museumplein area from Instagram across events.

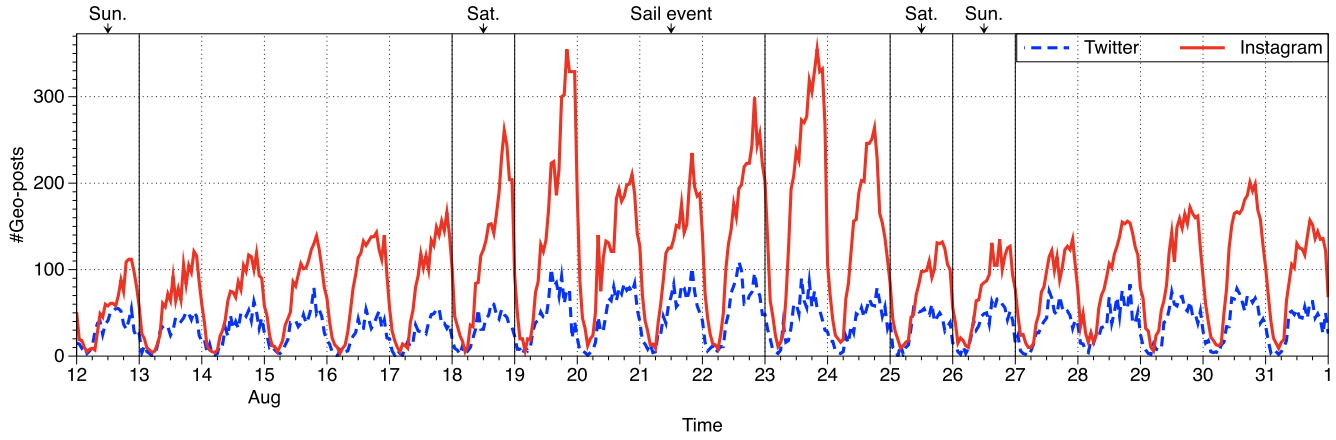
image based social media, is a preferred social networking choice during an event or festival. This result is consistent with findings from related work (Yang et al., 2016; Gong, 2016; Titos Bolivar, 2014). The sparsity of social media data currently may affect the representativeness of information for crowd management when deriving information from geo-referenced Tweets. However, such influence may be reduced by increasing the collecting of amount of social media data and applying geolocalisation methods.

The sparsity of social media data currently may affect the representativeness of information for crowd management. However, such influence may be reduced by increasing the amount of social media data collection and applying applying geolocalisation methods. Results also show that the amount of data collected during Sail is more than during King's Day except for the number of users on Instagram (Sail, Instagram users: 27.3 k; King's Day, Instagram users: 28.5 k), which indicates that Instagram has been used by more people during King's Day than during Sail.

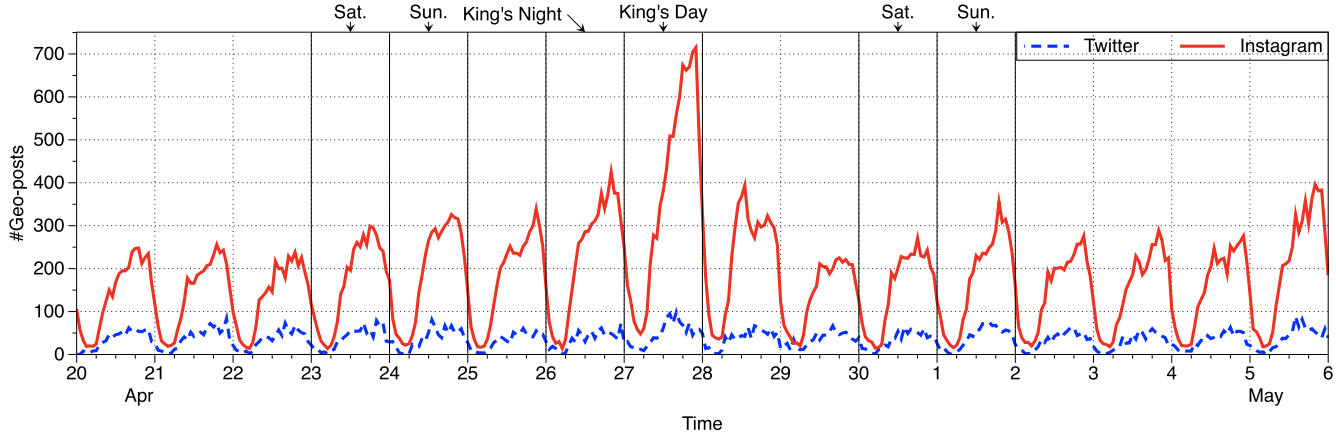
The number of users we collected on Twitter and Instagram may contain duplicated users, i.e. those who have profiles and use these two social networks in the same event. This is indeed a bias in the dataset. To counter this bias, we perform a manual check on counting the number of users who mention their tweets on Instagram posts, or mention their Instagram posts on tweets. The results show that less than 1% of users performed this operation. For users whose Twitter and Instagram accounts do not show any explicitly connections, we may not be able to count their amount. Even though, the amount of such user may not be high. We therefore assume that in this research all social media users collected on Twitter and Instagram are individual users.

3.3. Findings & analysis

In this section, we analyze the collected social media data on both social media networks, i.e. Twitter and Instagram, in terms of demographics, city-role, crowd temporal distribution, post location, PoIs and



(a) Sail



(b) King' s Day

Fig. 4. The temporal distribution of posts sent by people observed from social media.

Table 8
Sub-events of Sail and King's Day for temporal distribution analysis.

	Sub-event	Date	Duration Time	# Length	Area in Amsterdam
Sail	SAIL-in Parade	19–08-2015	13:30–16:00, Day	3.5 h	IJ, IJhaven
	Fireworks	19–08-2015 to 22–08-2015	22:00–23:00, Night	15 min in 1 h	IJhaven
King's Day	King's Night	26–04-2016 to 27–04-2016	18:00–02:00, Night	8 h	City Centre
	King's Day Boat Parade	27–04-2016	13:00–17:00, Day	4 h	City Centre

word-use. For each aspect, we compare findings across different events and social media platforms, and relate them to the respective event programs.

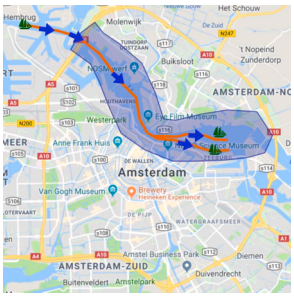
3.3.1. Demographic analysis

The demographic analysis consists of the analyses of age and gender.

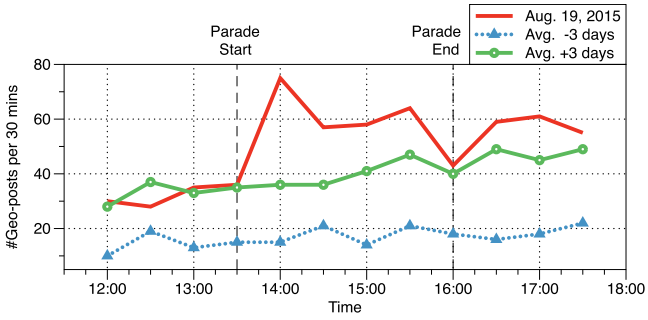
Age. Table 6 shows that more young-adult users (Sail: 42.7%, 49.8%; King's Day: 44.1%, 44.6%) are captured in both social media across events, followed by adult users and young users. Old users are extremely sparse, a result that we attribute to distinct technology penetration – old people seldom use social media. In the meantime, Instagram is far more used by young-adults, which might be because people in this age tend to share pictures taken during enjoyable events. The standard deviation of age of people during King's Day is lower than during Sail, indicating that more people across age ranges make use of social media during King's Day. The percentages of young and young-adult people observed during King's Day (Twitter: 13.4%, 44.1%, Instagram: 23.0%, 44.6%) are almost the same as during the Sail event (Twitter: 12.3%, 42.7%, Instagram: 24.4%, 49.8%).

We further zoom into 3 sub-events which attracted attendees with different demographic characteristics during Sail. According to the Sail official website, everyday between 09 h and 21 h – except for the last day – naval-related activities were organized around the IJhaven area where boats docked. Examples of such activities, which attracted people interested in this field, were tall-ship exhibitions, nautical history experiences, sports games on the water, and first aid in boat damage training. In contrast, the fireworks shows took place every night between 22 h and 23 h, and attracted more families [1]. The Sail parade had similar population distribution, but took place during the last day of the event between 12 h and 18 h. Table 7 shows the number of people in different age and gender groups of this three sub-events. The

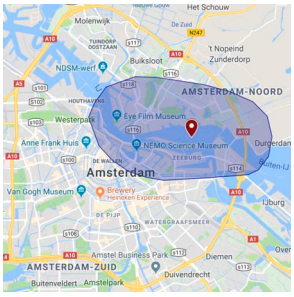
[1] https://www.abroad-experience.com/blog/sail-amsterdam-maritime-fun-for-the-whole-family/



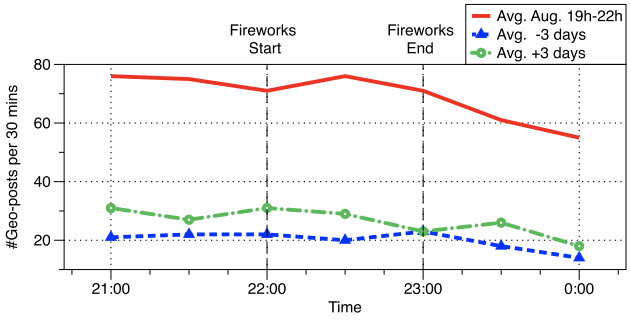
(a) Sail, In_Parade area



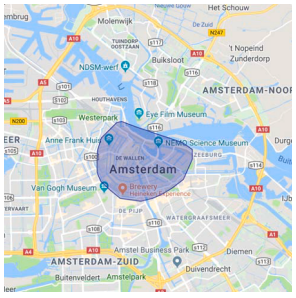
(b) Sail, In_Parade



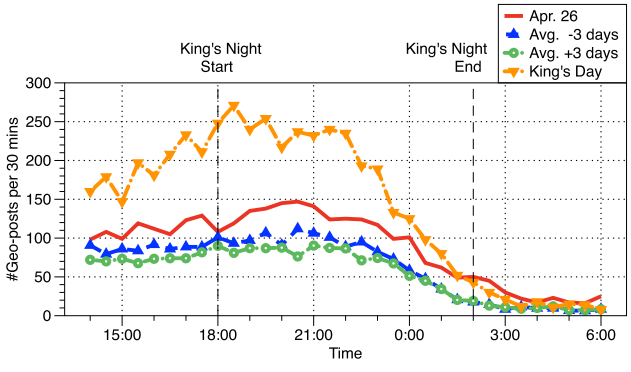
(c) Sail, Fireworks monitoring area



(d) Sail, Fireworks

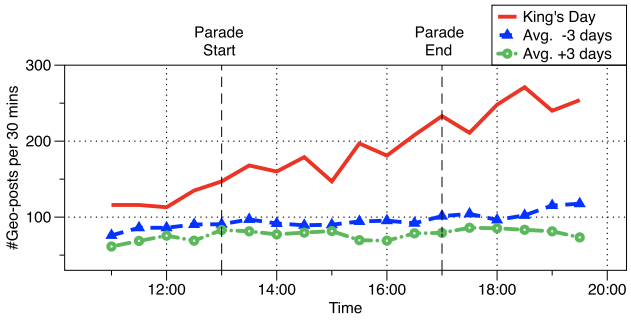


(e) King's Day, King's Night area



(f) King's Day, King's Night

(g) King's Day, Boat Parade area

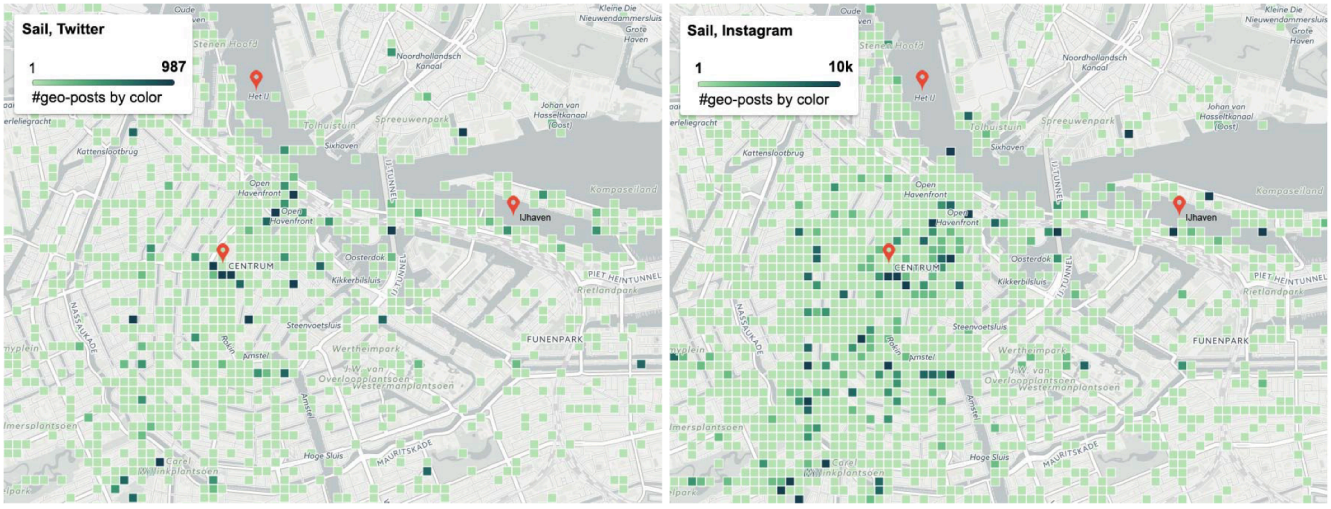


(h) King's Day, Boat Parade

Fig. 5. Temporal distribution of social media activity on sub-events of Sail and King's Day.

standard deviation of age during the Sail parade (Twitter: 23.9, Instagram: 86.9) is less than during the Sail fireworks show (Twitter: 23.3, Instagram: 112.3), followed by the Sail topic related activities (Twitter: 80.0, Instagram: 235.7). It is in accordance with the expectation that the Sail fireworks and parade attract more families which are more evenly distributed populations.

With regard to the social media users in two terrains according to Table 4, the standard deviation of age in Javakade in Sail event



(a) Sail, Twitter

(b) Sail, Instagram

(c) King's Day, Twitter

(d) King's Day, Instagram

Fig. 6. The heatmap of the position of people by color observed from social media posts.

(Twitter: 56.2, Instagram: 267.2) is larger than in Zuidplein during King's Day event (Twitter: 37.7, Instagram: 45.4). This is in line with the fact of different types of PoIs in terrains; Recreation amenities, such as bars, clubs, shops and restaurants which located in Zuidplein may attract more young people thus lead to more social media posts sent by young and young-adult than in residential area.

Gender According to Table 6 in both events, more male users are detected on Twitter while more female users are on Instagram, indicating that sharing pictures in such festivals is more popular among female users. The ratio between male and female users is similar for the two events (Twitter: 1.3, Instagram: 0.7). With regard to the three sub-events shown in Table 7, the standard deviation of ratio between male and female (Twitter: 61.4, Instagram: 57.3) is less obvious compared with standard deviation in age (Twitter: 32.6, Instagram: 79.5).

With regard to the gender distribution in the two terrains shown in Table 4, the standard deviation of gender on Instagram (Javakade: 131.5, Zuidplein: 53.1) is larger than on Twitter (Javakade: 101.8, Zuidplein: 39.6). In particular on Instagram, the gender of people in Javakade in Sail event is more equally distributed than on Zuidplein in King's Day. This may be caused that more recreation and activities on Zuidplein during King's Day gives rise to the usage of Instagram, the image based social media network, which are more popular for female users to share their feelings (Yang et al., 2016).

3.3.2. City-role

Table 6 shows that more foreign travelers than residents and local tourists are identified on social media in both events. As a popular touristic city, Amsterdam usually attracts huge numbers of tourists, and large scale events only increase these numbers. It is also likely that foreign travelers are more active on social media as they feel fresh and excited to be in a new place. This is also observed in related work (Yang et al., 2016). The ratio between local travelers and residents during King's Day is lower than during Sail, thus showing that Sail, an event taking place every 5-years, is more likely to attract visitors from other cities and countries. According to Table 5 and Table 6, the proportion of Kingsday Instagram users whose city-role is detected only accounts around 24.67% among all Instagram users who sent geo-reference posts on Kingsday in the event area. The lower rate of city-role detection may be caused by the updated privacy protection settings on Instagram and the limitation of Instagram API.

With regard to the city-role on the two terrains, according to Table 4, more residents are observed than any other types of users on both social media networks. It is in line with the actual situation that either the Javakade or Zuidplein is not the tourist attractions during events. Consequently, less local and foreign tourists than residents are captured on social media.

We further zoom into the Museumplein in Amsterdam, a popular

(a) Sail, Central Station

(b) King's Day, Central Station

(c) Sail, PoI population

(d) King's Day, PoI population

Fig. 7. The PoIs observed from social media across events.

tourist attraction, to show how social media characterize the city-role in this area, as compared to events. The Museumplein is a famous tourist attraction area that hosts several popular museums (e.g. the Van Gogh museum, and the Rijksmuseum). Also, the square hosts the famous 'I Amsterdam Sign' attracting numerous tourists to take pictures with. As the Twitter data is too sparse for such a small area (only 13 tweets during Sail, and 8 during King's Day), we focus on Instagram data. According to Fig. 3, more foreign tourists are observed in this area than local travellers and residents. This observation is in line with the fact that we see more pictures about this area on Instagram.

3.3.3. Crowd temporal distribution

Fig. 4 shows the temporal distribution of social media activities observed around the two events in Amsterdam. In order to compare the pattern of the crowd distribution between event days and regular days, we show the temporal distribution during seven days before and seven days after the respective events. For each event the temporal distribution clearly illustrates the daily pattern. According to Fig. 4, the amount of social media activities on event days increases faster than on regular days, and the total amount from one day before the event to one day after the event is larger than for ordinary days, including weekend days.

We further zoom into four attractive sub-events during Sail and King's Day event which attracts a large amount of people and may give rise to social media usage, listed in Table 8. We show the temporal distribution of social media activities per 30 min around each sub-event and compare it with three days before and three days after the event in the same time and area, shown in Fig. 5 based on Instagram posts which are far more frequently observed than Twitter.

We found that during the Sail-In parade and the King's Day boat parade the temporal distribution of social media activity shows significantly distinct patterns compared with three days before and three days after the event, which is in line with the reality that such sub-events attract a large number of people. During the other two sub-events, i.e. Sail Fireworks show and King's Night celebration, the temporal distribution of social media activity shows less distinct patterns compared with three days before and three days after the event. This observation indicates that sub-events such as the boat parade, which lasts for several hours and takes place in a city-scale area with magnificent views over the day is suitable to be characterized through a temporal distribution based on social media. Details about temporal distribution analyses of social media activities for the two events are presented in the Appendix.

3.3.4. Position

The positions of crowds in a city-scale event can be derived from social media data, as described in Section 2. Fig. 6 shows the position heatmap of social media users observed during the two events in Amsterdam. It shows that during both events crowds are active on social media mostly in the city center. However, attendees of Sail were more active in the area around the IJ and IJhaven, a clear indication for crowd managers that special activities taking place around that area during Sail event which attract a large amount of people.

We could also observe event-related social media activities outside the event area in Fig. 6, which may be caused by people who sent posts after attending the event.

3.3.5. PoI

As illustrated in Section 2, the PoI data help with deriving

(a) Sail, Sail-in Parade

(b) Sail, Fireworks

(c) King's Day, Boat Parade

(d) King's Day, King's Night

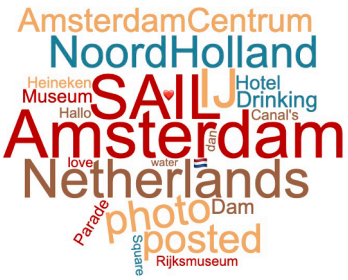
Fig. 8. The Area of Interest detected by clustering PoIs using social media posts for 4 sub-events listed in Table 8. The size and colour of each plot indicate the amount of visits of the covering area. The bigger size in Red colour denotes larger popularity, while smaller size in Blue colour denotes less popularity.

information from pedestrian behavior factors such as built environment, trip characteristics and social networks. We first zoom into the area of Central Station, which is the major transportation hub in Amsterdam. According to Fig. 7 (a) and (b), there are more Travel & Transport PoIs (red dot) visited by social media users than other PoIs, which corresponds to the expected indication of the land use of this area according to the land-use of Amsterdam [2]. According to Fig. 7 (c) and (d), popular PoIs during Sail are spread around 5 oceans, particularly in the Orange ocean (IJhaven area) where the boats docked, while during King's Day the popular PoIs are less distributed in the IJhaven area.

We can also use social media posts to discover PoIs which are most visited in different events. By clustering these PoIs we can further detect the most popular area visited by people during events, i.e. the Area of Interest, which is valuable for crowd managers to understand the most important area in different events.

To showcase the Area of Interest discovered using social media for crowd management, we generate a clustered map for each sub-events in 4 sub-events (Table 8) using DBSCAN algorithm (Ester et al., 1996)

[2] https://maps.amsterdam.nl/grondgebruik/?LANG = en



(a) Sail, Twitter

(b) Sail, Instagram

(c) King's Day, Twitter

(d) King's Day, Instagram

(e) Sail, Twitter, filtered

(f) Sail, Instagram, filtered

(g) King's Day, Twitter, filtered

(h) King's Day, Instagram, filtered

Fig. 9. Word-clouds during Sail and King's Day.

(a) Javakade, Sail, Instagram, filtered (b) Zuidplein, King's Day, Instagram, filtered

Fig. 10. Word-clouds in Javakade during Sail, and Zuidplein during King's Day.

which groups together points that are closely packed together. According to Fig. 8 (a)-(d), Amsterdam Central Station is a popular area across sub-events, which is natural as it is the biggest transport hub in the city center. During both Sail sub-events, shown in Fig. 8 (a) and (b), area close to IJ (bay area), such as Javakade and Eastern Docklands, are more popular than in King's Day event, which is in line with the fact that most activities in Sail event are carried out around IJ. In particular, during Fireworks show in the night during Sail event, shown in Fig. 8 (b), the Eastern Docklands are getting more popular. It may indicate that Eastern Docklands is a good place to have a full view of fireworks and thus attracts a large number of people. While, during King's Day, according to Fig. 8(c) and (d), the most popular areas are distributed in the city center rather than the IJ area. In Kingsday boat parade, shown in Fig. 8 (c), the most popular areas include city center where parade is taking place, the Westerpark where music performances are being held, and the Vondelpark where activities such as flea markets are carried out. During King's Night, according to Fig. 8(d), the most popular area, besides the Amsterdam city center, are the Westerpark and Leidseplein, where music performances and clubs (or bars) attract a large number of visitors, respectively. Crowd managers may perform measures in these areas to avoid crowdedness during sub-events. Details about clustering analyses of social media activities for 4 sub-events are presented in the Appendix.

3.3.6. Word use

Analogue to the word use of the crowd, we generate word clouds showing the most frequently used words observed from social media. The word clouds are generated using text content of the posts filtering out Hashtags and URLs from Twitter and Instagram, respectively. As shown in Fig. 9(a)–(d), the general words about the city (e.g. Amsterdam, Netherlands) and event name related (e.g. Sail: SAIL. King's Day: King's, King's Day, King's night) in either English or Dutch are most frequently used across events. In order to characterize what people talk about discarding these words, we exclude them and regenerate word-clouds in Fig. 9(e)–(h). Results shows that activity related words (e.g. Sail: ship, drinking, parade. King's Day: drinking, orange, kingsland) and emotional words including emojis are captured (e.g. happy, love, vertraagd (delayed), vechtpartij (fighting), best, amazing, exciting, fun, lovely). It indicates that the popular topics and sentiment of people in the crowd which are valuable for crowd management can be derived from social media.

We also zoom-into the two terrains, i.e. the Javakade in Sail event and Zuidplein in King's Day event. Fig. 10 (a) and (b) shows the most frequently used words captured in these two terrains on Instagram, excluding the general words and event names. These words characterise events in these two terrains in terms of crowd sentiment and activity. For instance, frequently used words such as 'thanks', 'wonderful', and 'love' describe a very good atmosphere during the activities, mentioned in words such as 'boatlife', 'Vuurwerk' ('fireworks' in Dutch), 'sunset' in Javakade in Sail, and in music performances such as 'martingarrix' and 'TIKTAK' taking place around the Zuidplein during Kingsday.

To showcase the availability of information related to the attendees' perception which can be used for crowd managers to estimate the crowdedness of the crowd during events. We analyze the occurrence distributions of words such as 'Crowded' and 'Druk' (*crowded* in Dutch) in both events in Table 9, normalizing by the total number of posts. During both events people on Twitter posted these words more often than on Instagram: it is possible that people are more willing to share their negative emotions, such as crowded perceptions, on Twitter. While comparing two events, more 'crowded' and 'Druk' words are captured during Sail than during King's Day. This may also indicate that people experience more crowdedness during the Sail event than during King's Day.

3.3.7. Discussion

In the above sections, we report the outcomes of the crowd characterization operations performed on social media data, which could be collected in a timely manner during the two events. These crowd properties are calculated to provide relevant insights for crowd management purposes. Crowd managers could apply crowd management measures by taking into consideration the semantic and qualitative interpretation of social media posts. Therefore, the validity and reliability of crowd characterization are essential for crowd management.

By relying on state-of-the-art techniques (Yang et al., 2016; Noulas et al., 2012), we show through two case studies how social media data can be enriched to surface socio-demographic properties of their users. We acknowledge that the limited availability of meaningful profile pictures and location information reduces the intrinsic utility of socio-demographic analysis of social media data. Although it is good to take these limitations into account and work further on more precise methods to derive this information, the resulting information may

Table 9
Word count of 'crowded' in word use across events. **Druk**: 'Crowded' in Dutch; **#p**: Number of posts; **%p**: Percentage of posts.

Event		Twitter				Instagram			
		Crowded		Druk		Crowded		Druk	
		#p	%p	#p	%p	#p	%p	#p	%p
Sail	Sail	1	0.00%	32	0.28%	30	0.05%	103	0.17%
Kingsday	Kingsday	1	0.00%	5	0.11%	2	0.00%	5	0.01%

benefit crowd management purposes, as currently hardly any information on crowd characteristics is available.

In the meantime, we noticed that the sparsity of social media data has less influence on the validity and reliability of characterizing the crowd for crowd management during city-scale events. For instance, the temporal and spatial distribution of posts, PoI popularity, as well as word use successfully characterize the crowd and show distinctions across events.

In contrast, the bias of social media usage affects crowd characterization for crowd management. The bias in age, e.g., older adults seldom use social media, affects the identified age distribution of the crowd. The bias in gender with respect to the platform, i.e., females are more active on Instagram, affects characterizing the gender composition of the crowd. Also, the bias in platform usage, i.e., the Instagram data, is more sensitive to the temporal distribution than Twitter, which presents more diverse distinctions between events or between different days in an event, affects characterizing the temporal distribution of the crowd. Besides, the bias in city-role, e.g., tourists are more active than residents, reduces the capability of crowd characterization using social media. However, the comparison of the distribution of age, gender, and city-role through social media in different events successfully characterized the distinction of events. E.g., the age distribution during King's Day is more uniform than during Sail. Also, one can try to compensate for these biases when more is known about them, such as event programs, sensor or stewards observations, and historical data on crowd characteristics. This information can be utilized in crowd management to gain event distinctions and apply measures accordingly.

4. Summary and conclusions

Nowadays, city-scale events are getting more popular. Stakeholders of these events demand qualitative and quantitative insights into the crowd to be managed. Conventional solutions depend on manual observations, which are expensive, prone to introduce observation biases, and not suitable for longitudinal observations.

In this paper, we advocate the use of social media data as a valuable and effective alternative data source for crowd characterization purposes. We screened a set of factors that are known to influence pedestrian behaviors, which, therefore, are relevant for crowd characterization. We also select examples of methods that could be used to derive information about these factors from social media data.

We apply these methods in the context of two city-scale events – Sail 2015 and King's Day 2016, in Amsterdam, the Netherlands – and reflect on the accuracy and reliability of these methods. Based on the results of the existing methods, we can conclude whether dedicated methods need to (and can) be developed. For instance, we observed that the age distribution during King's Day is more evenly distributed than during Sail, a result that complies with the expected composition of the events' crowds. We also found that fewer local tourists join the King's Day event in Amsterdam than during Sail, which may be explained by the fact that people in other cities are more willing to travel to Amsterdam for Sail, the event occurring once each 5-years than the yearly King's Day. We noticed that the amount of social media posts sent by people during events is far more than during regular days, which is in line with the fact that it is more crowded during the city events. The temporal distribution of sub-events, which take place in large areas, lasts for several hours during the day, e.g., the Sail-in parade and King's Day boat parade, illustrates clearly the increase of participants. Moreover, more social media usage and PoI visits are observed in the IJhaven area during Sail than during the King's Day event, which is in line with the fact that Sail activities take place around the IJhaven area where ships docked. The word use of the people across events from social media successfully captured the words about event topics and people's emotions. All of these observations indicate crowd management strategies could take into consideration the different characteristics of crowds, in terms of demographics, spatial–temporal distribution, Point of Interest (PoI) preferences, and word use, in the context of city events.

The social media sparsity has less impact on crowd characterization. However, the bias of social media usage in terms of age, gender, and city-role, as well as social media platform selection, affects providing information about crowd during events, such as absolute age, gender distribution, and temporal distribution of a crowd. Still, the comparison of results between the two events is in line with expected event characteristics.

In future work, we plan to continue studying in multiple topics based on the findings and analysis in the current study. To deal with social media data sparsity, we would like to use geolocalisation techniques to increase the amount of geo-referenced posts to counter the sparse of geo-referenced data on social media. To provide more relevant information for crowd management, we plan to explore methods to derive more social media proxies about people in the crowd, such as their sentiment, the main area they visited in different events. A further sensitivity analysis would be done in terms of the number of points and the radius (i.e., the "epsilon" parameter for the DBSCAN algorithm) for expanding clusters. We also would like to investigate the feasibility of the current study in other events which may have different characteristics.

CRediT authorship contribution statement

Vincent X. Gong: Conceptualization, Methodology, Software, Validation, Writing - original draft. **Winnie Daamen:** Conceptualization, Methodology, Supervision. **Alessandro Bozzon:** Conceptualization, Methodology, Supervision. **Serge P. Hoogendoorn:** Conceptualization, Methodology, Supervision, Resources, Funding acquisition.

Funding

The research leading to these results has received funding from the European Research Council under the European Union Horizon 2020 Framework Programme for Research and Innovation. It is established by the Scientific Council of the ERC Grant Agreement no. 669792 (Allegro).

Acknowledgement

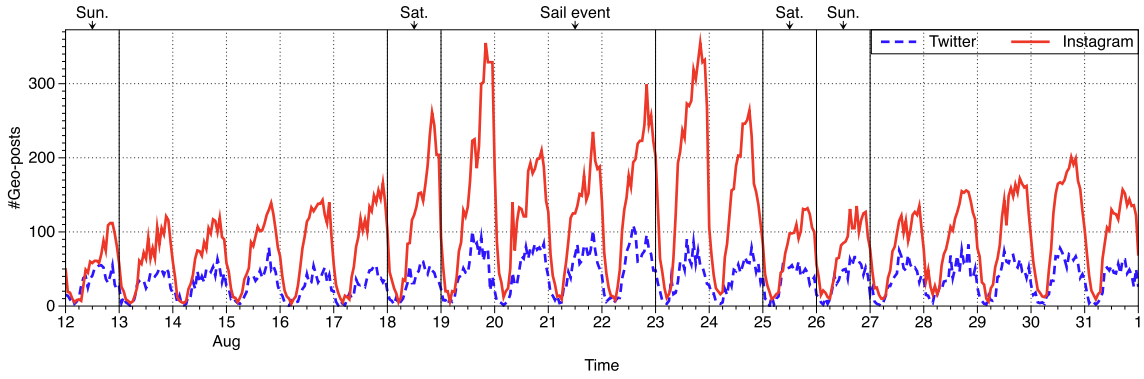
The authors would like to thank Robrecht Baving from The Security Company BV for kindly sharing his knowledge about the crowd management.

Appendix A. Appendix

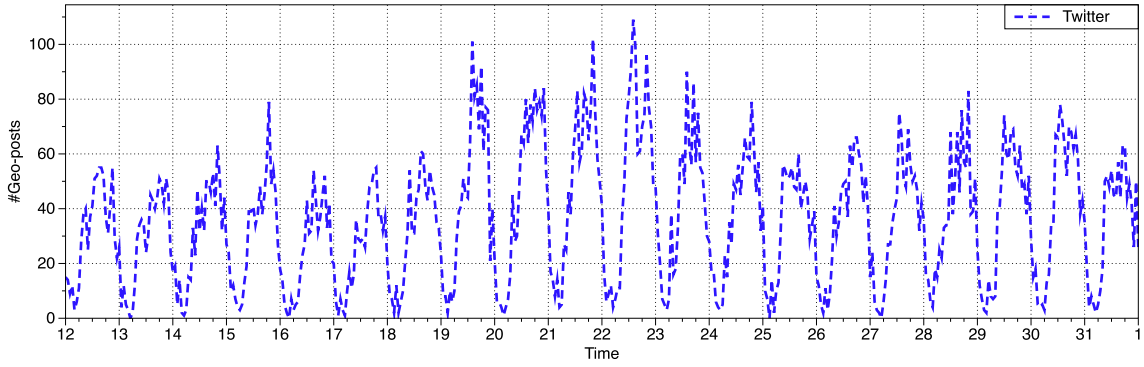
A.1. Social media geo-posts sent around Sail and Kingsday event

.

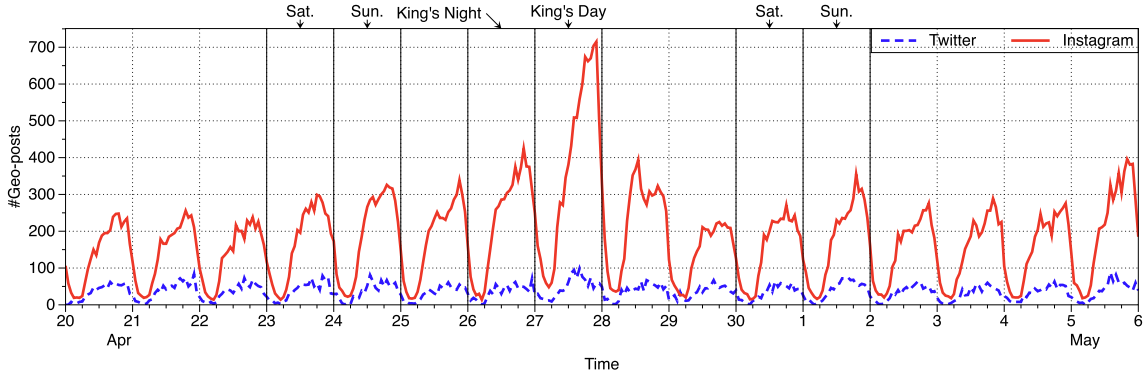
- The temporal distribution of posts sent by people observed from social media in Fig. 11.
- The Total, Max, and Minimal number of geo-posts sent around Sail event 2015 in Table 10 and around King's Day event 2016 in Table 11.



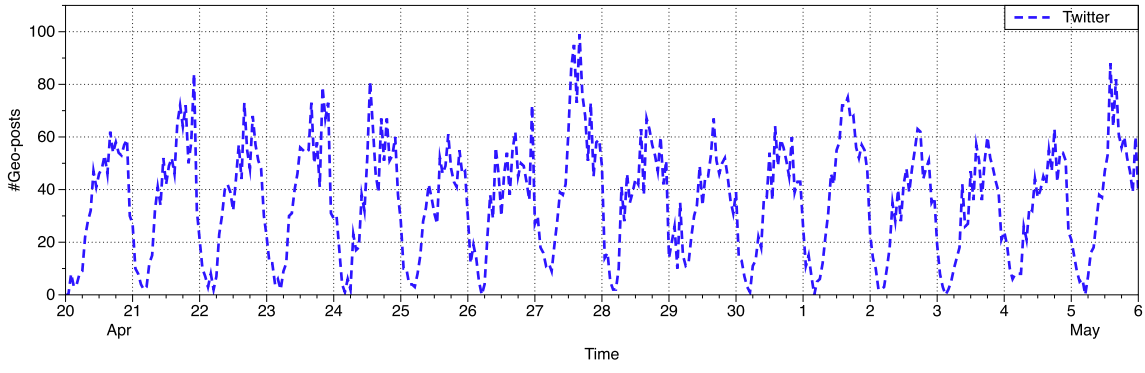
(a) Sail



(b) Sail, Twitter



(c) Kingsday



(d) Kingsday, Twitter

Fig. 11. The temporal distribution of posts sent by people observed from social media.

Table 10
The Total, Max, and Minimal Number of Geo-posts Sent Around Sail Event 2015

		Date																			
								Weekend	Sail Event						Weekend						
		Sun Aug 12	Mon Aug 13	Tue Aug 14	Wed Aug 15	Thu Aug 16	Fri Aug 17	Sat Aug 18	Sun Aug 19	Mon Aug 20	Tue Aug 21	Wed Aug 22	Thu Aug 23	Fri Aug 24	Sat Aug 25	Sun Aug 26	Mon Aug 27	Tue Aug 28	Wed Aug 29	Thu Aug 30	Fri Aug 31
Twitter	#P. of day	759	686	727	757	591	629	754	1083	1092	1154	1212	1002	892	834	868	832	926	905	980	836
	Max #P., Time (hour)	15:00	18:00	20:00	19:00	15:00	19:00	15:00	14:00	18:00	20:00	14:00	14:00	19:00	16:00	17:00	13:00	20:00	12:00	13:00	17:00
	Max #P.	55	51	63	79	54	55	61	101	84	102	109	90	79	60	66	75	83	74	78	63
	Min #P., Time (hour)	04:00	05:00	05:00	06:00	04:00	05:00	03:00	03:00	04:00	05:00	05:00	05:00	05:00	03:00	03:00	05:00	04:00	04:00	06:00	05:00
	Min #P.	3	0	1	3	0	1	2	2	1	4	4	3	3	0	2	0	4	2	3	2
Instagram	#P. of day	1291	1435	1510	1696	1943	2012	2796	3775	2967	2766	3429	4638	3278	1796	1751	1815	2156	2402	2754	2268
	Max #P., Time (hour)	19:00	21:00	21:00	20:00	19:00	22:00	20:00	20:00	21:00	20:00	20:00	20:00	18:00	20:00	18:00	22:00	19:00	19:00	18:00	17:00
	Max #P.	112	121	119	139	143	166	261	355	210	235	299	355	264	132	135	135	156	172	201	157
	Min #P., Time (hour)	03:00	05:00	05:00	05:00	05:00	04:00	05:00	04:00	04:00	06:00	06:00	05:00	05:00	04:00	06:00	05:00	06:00	05:00	06:00	05:00
	Min #P., Time (hour)	5	4	4	5	3	5	5	6	12	9	8	11	16	9	10	10	16	6	12	9

#GP.: amount of Geo-posts.
Max #GP. Time(hour): the time in hour during which the max amount of Geo-posts is observed.

Table 11
The Total, Max, and Minimal Number of Geo-posts Sent Around King's Day Event 2016.

Date																					
					Weekend			K. Night	K. Day				Weekend								
		Wed April 20	Thu April 21	Fri April 22	Sat April 23	Sun April 24	Mon April 25	Tue April 26	Wed April 27	Thu April 28	Fri April 29	Sat April 30	Sun May 1	Mon May 2	Tue May 3	Wed May 4	Thu May 5				
Twitter	#P. of day	822	926	839	918	886	768	855	1134	914	833	814	957	785	715	794	909				
	Max #P., Time (hour)	16:00	22:00	16:00	20:00	13:00	17:00	23:00	16:00	16:00	16:00	14:00	16:00	17:00	18:00	18:00	14:00				
	Max #P.	62	84	73	79	81	61	72	99	67	67	64	75	63	60	63	88				
	Min #P., Time (hour)	01:00	04:00	05:00	05:00	04:00	05:00	05:00	06:00	04:00	06:00	05:00	04:00	04:00	03:00	04:00	05:00				
	Min #P.	0	2	2	2	1	3	0	9	2	10	1	0	2	0	6	0				
Instagram	#P. of day	3369	3402	3325	4006	4574	4277	5450	8902	5575	3407	3622	4160	3666	3647	3723	4959				
	Max #P., Time (hour)	19:00	19:00	19:00	18:00	19:00	21:00	20:00	22:00	13:00	18:00	18:00	19:00	21:00	20:00	22:00	20:00				
	Max #P.	248	256	238	299	326	339	425	715	394	225	271	355	276	289	276	396				
	Min #P., Time (hour)	03:00	04:00	05:00	05:00	05:00	04:00	05:00	05:00	05:00	06:00	05:00	05:00	05:00	05:00	05:00	04:00				
	Min #P.	19	19	14	14	22	17	14	48	36	20	15	16	20	18	20	18				

A.2. Areas of Interest based on clustering social media posts

The Area of Interest detected by clustering PoIs using social media posts for 4 sub-events listed in Table 8 are shown in Fig. 12–15. These cluster maps are generated using DBSCAN algorithm (Ester et al., 1996), which groups points within distance of 50 meters (epsilon = 0.05). Such groups will be identified as a cluster if the number of points in a group is more than 15 (min_Points = 15), otherwise each points will be determined as outliers. Each colour represents one cluster. Black points are outliers which fail to form group with any other points. The count number denotes the amount of points in each cluster.

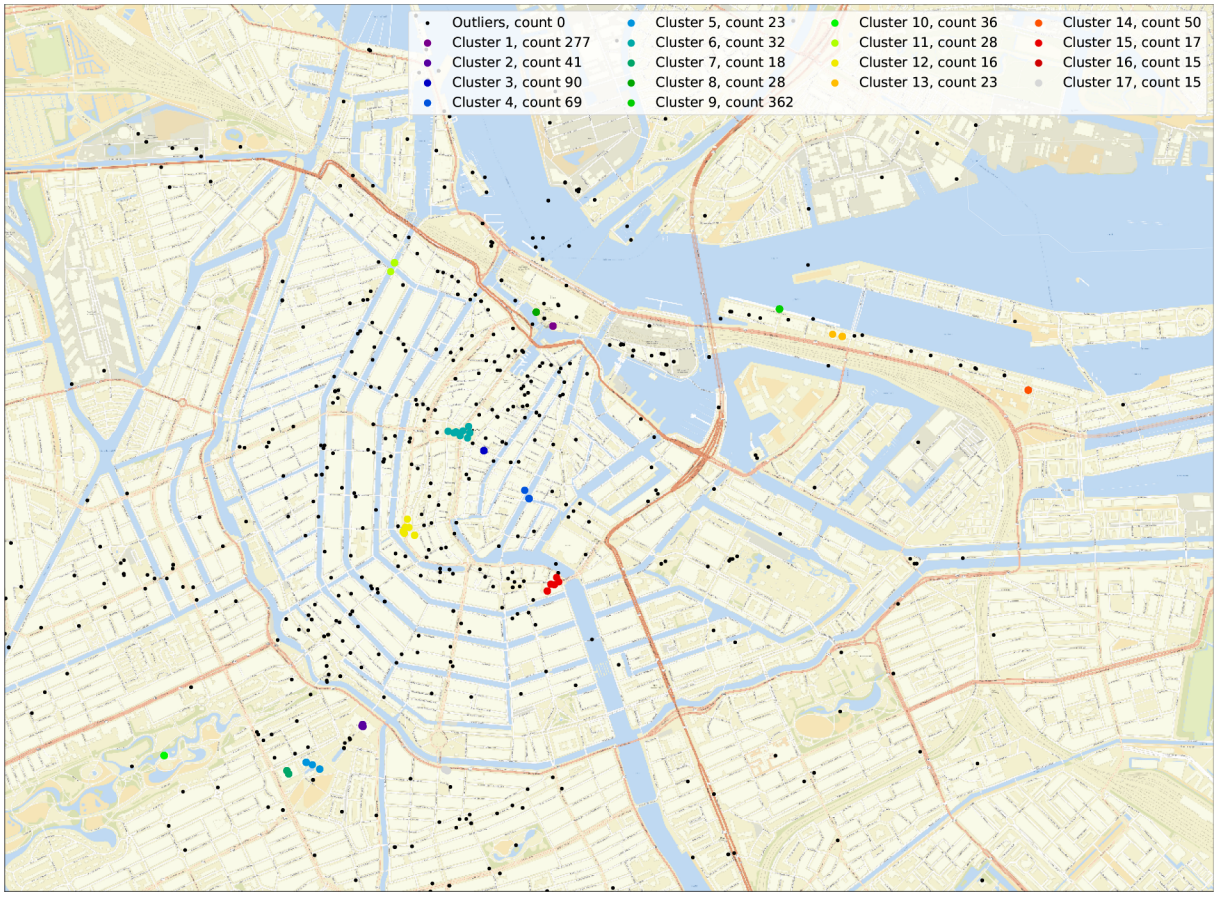


Fig. 12. The Area of Interest detected by clustering PoIs using social media posts in Sail, Sail-in Parade.

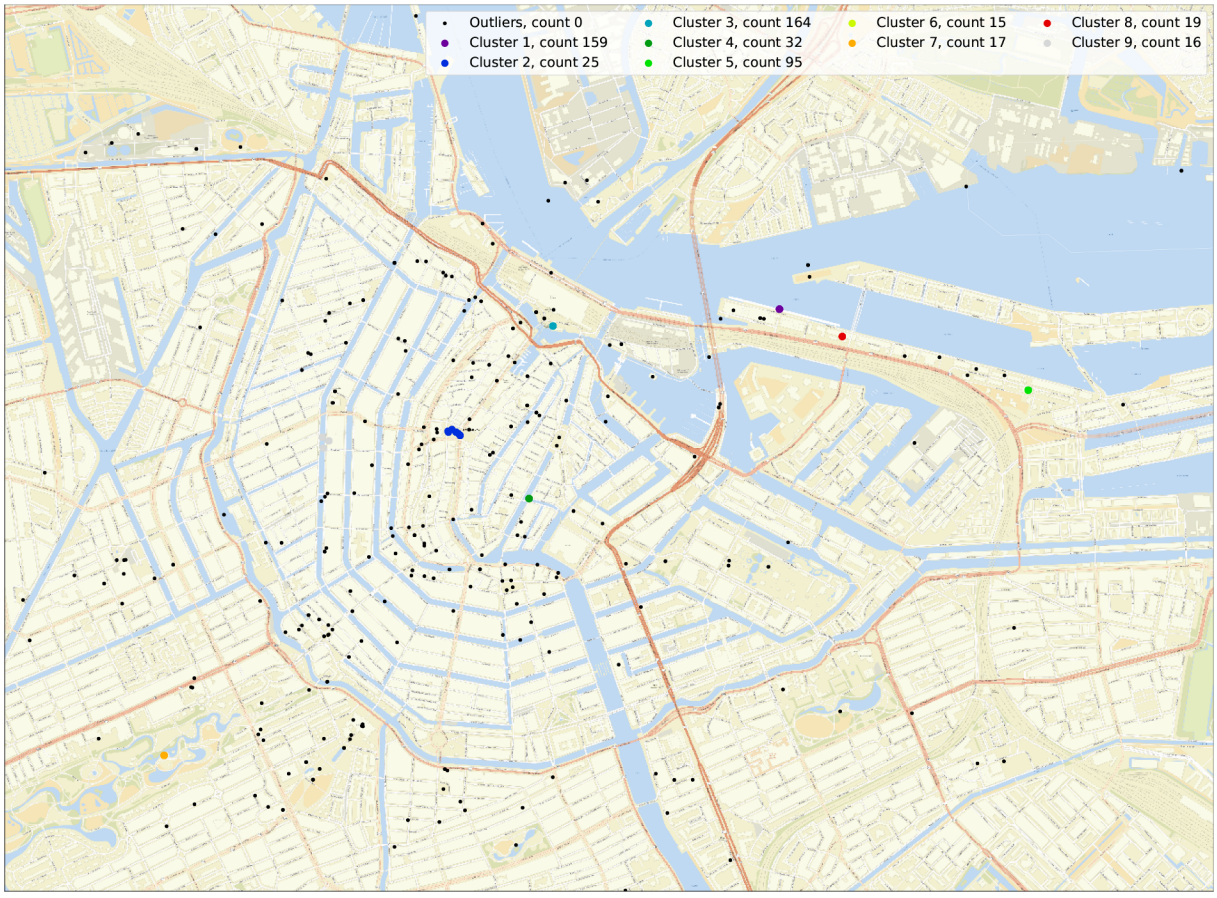


Fig. 13. The Area of Interest detected by clustering PoIs using social media posts in Sail, Fireworks.

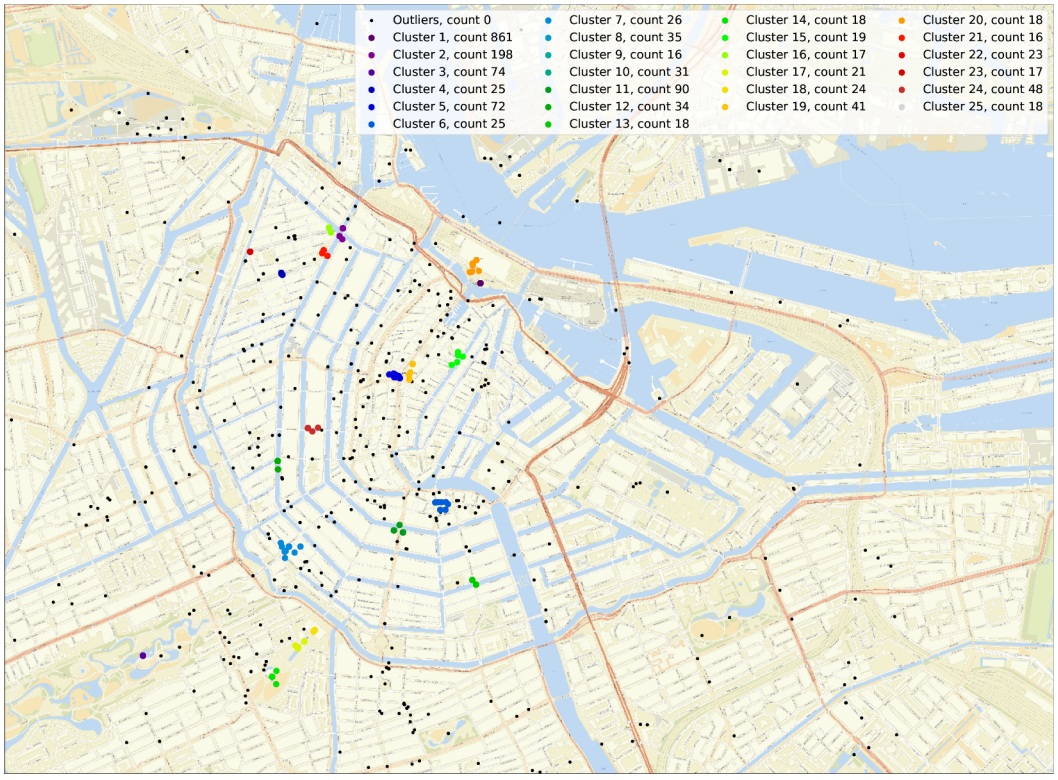


Fig. 14. The Area of Interest detected by clustering PoIs using social media posts in King's Day, Boat Parade.

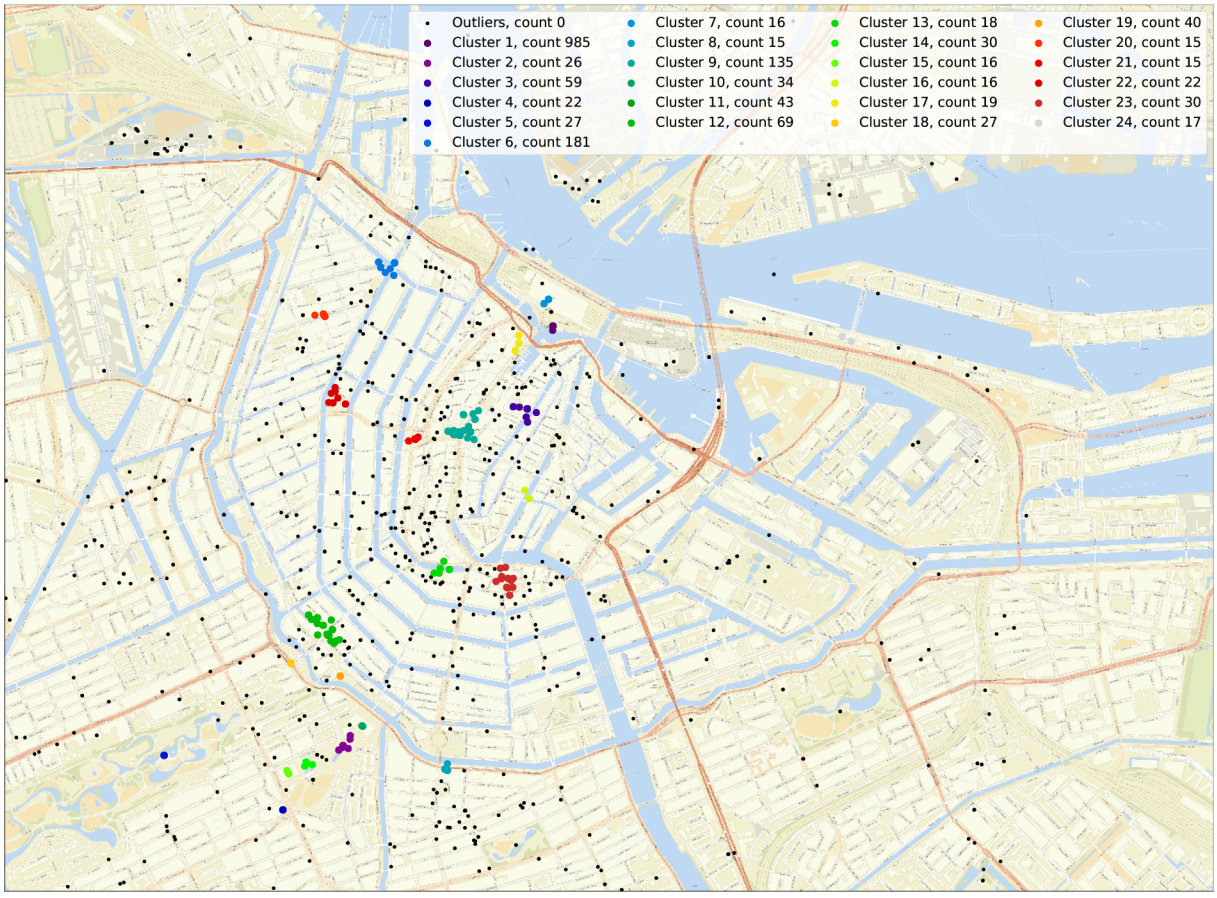


Fig. 15. The Area of Interest detected by clustering PoIs using social media posts in King's Day, King's Night.

References

Abbasi, A., Rashidi, T.H., Maghrebi, M., Waller, S.T., 2015. Utilising location based social media in travel survey methods: bringing twitter data into the play. In: Proceedings of the 8th ACM SIGSPATIAL international workshop on location-based social networks, ACM, p. 1.

Abbott, J., Geddie, M.W., 2000. Event and venue management: minimizing liability through effective crowd management techniques. Event Manage. 6, 259–270.

Al-Zahrani, M.S., Bissada, N.F., Borawski, E.A., 2003. Obesity and periodontal disease in young, middle-aged, and older adults. J. Periodontol. 74, 610–615.

Alkhatib, M., El Barachi, M., Shaalan, K., 2019. An arabic social media based framework for incidents and events monitoring in smart cities. J. Clean. Prod. 220, 771–785.

Balduini, M., Bocconi, S., Bozzon, A., Della Valle, E., Huang, Y., Oosterman, J., Palpanas, T., Tsytsarau, M., 2014. A case study of active, continuous and predictive social media analytics for smart city. In: Proceedings of the Fifth Workshop on Semantics for Smarter Cities a Workshop at the 13th International Semantic Web Conference (ISWC 2014), pp. 31–46.

Balduini, M., Bozzon, A., Della Valle, E., Huang, Y., Houben, G.J., 2014b. Recommending venues using continuous predictive social media analytics. IEEE Internet Comput. 18, 28–35.

Berrigan, D., Troiano, R.P., 2002. The association between urban form and physical activity in us adults. Am. J. Preventive Med. 23, 74–79.

Bocconi, S., Bozzon, A., Psyllidis, A., Titos Bolivar, C., Houben, G.J., 2015. Social glass: a platform for urban analytics and decision-making through heterogeneous social data. In: Proceedings of the 24th International Conference on World Wide Web ACM, pp. 175–178.

Chen, J., Hsieh, G., Mahmud, J.U., Nichols, J., 2014. Understanding individuals' personal values from social media word use. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing ACM, pp. 405–414.

Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z., 2011. Exploring millions of footprints in location sharing services. ICWSM 2011, 81–88.

Cottrill, C., Gault, P., Yeboah, G., Nelson, J.D., Anable, J., Budd, T., 2017. Tweeting transit: an examination of social media strategies for transport information management during a large event. Transp. Res. Part C: Emerging Technol. 77, 421–432.

Daamen, W., 2004. Modelling passenger flows in public transport facilities.

Davis Jr, C.A., Pappa, G.L., de Oliveira, D.R.R., de L Arcanjo, F., 2011. Inferring the location of twitter messages based on user relationships. Transactions in GIS 15, 735–751.

Duives, D.C., Daamen, W., Hoogendoorn, S.P., 2016. The influence of the interaction characteristics on the movement dynamics of pedestrians. In: Proceedings of the 8th International Conference on Pedestrian and Evacuation Dynamics (PED2016). University of Science and Technology of China Press.

Earl, C., Parker, E., Tatrai, A., Capra, M., et al., 2004. Influences on crowd behaviour at outdoor music festivals. Environ. Health 4, 55.

Eash, R., 1999. Destination and mode choice models for nonmotorized travel. Transp. Res. Rec.: J. Transp. Res. Board 1–8.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, pp. 226–231.

Favaretto, R.M., Dihl, L.L., Musse, S.R., 2016. Detecting crowd features in video sequences. In: Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on. IEEE, pp. 201–208.

Florez, J., Muniz, J., Portugal, L., 2014. Pedestrian quality of service: lessons from maracan a stadium. Procedia-Soc. Behav. Sci. 160, 130–139.

Gao, S., 2015. Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. Spatial Cognit. Computat. 15, 86–114.

Gong, V.X., Yang, J., Daamen, W., Bozzon, A., Hoogendoorn, S., Houben, G.J., 2018. Using social media for attendees density estimation in city-scale events. IEEE Access 6, 36325–36340.

Gong, X., 2016. Exploring human activity patterns across cities through social media data.

Guo, Z., 2009. Does the pedestrian environment affect the utility of walking? a case of path choice in downtown boston. Transp. Res. Part D: Transp. Environ. 14, 343–352.

Han, Q., Dellaert, B.G., Van Raaij, W.F., Timmermans, H.J., 2010. Visitors' strategic anticipation of crowding in scarce recreational resources. J. Retailing Consumer Serv. 17, 449–456.

Handy, S., 1996. Urban form and pedestrian choices: study of austin neighborhoods. Transp. Res. Rec.: J. Transp. Res. Board, 135–144.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geolocated twitter as proxy for global mobility patterns. Cartogr. Geographic Inf. Sci. 41, 260–271.

Hochmair, H.H., Juhász, L., Cvetojevic, S., 2018. Data quality of points of interest in selected mapping and social media platforms. In: LBS 2018: 14th International Conference on Location Based Services. Springer, pp. 293–313.

Hoogendoorn, S.P., Bovy, P.H., 2005. Pedestrian travel behavior modeling. Networks Spatial Econ. 5, 193–216.

Jamil, S., Basalamah, A., Lbath, A., Youssef, M., 2015. Hybrid participatory sensing for analyzing group dynamics in the largest annual religious gathering. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous

Computing ACM, pp. 547–558.
Kim, S., Park, S., Lee, J.S., 2014. Meso-or micro-scale? Environmental factors influencing pedestrian satisfaction. Transp. Res. Part D: Transp. Environ. 30, 10–20.
Krueger, R., Han, Q., Ivanov, N., Mahtal, S., Thom, D., Pfister, H., Ertl, T., 2019. Bird's-eye-large-scale visual analytics of city dynamics using social location data. Comput. Graph. Forum 595–607 Wiley Online Library.
Lansley, G., Longley, P., 2016. Deriving age and gender from forenames for consumer analytics. J. Retailing Consumer Serv. 30, 271–278.
Lingad, J., Karimi, S., Yin, J., 2013. Location extraction from disaster-related microblogs. In: Proceedings of the 22nd international conference on world wide web ACM, pp. 1017–1020.
Longley, P.A., Adnan, M., Lansley, G., 2015. The geotemporal demographics of twitter usage. Environ. Plann. A 47, 465–484.
Maley, D.W., Weinberger, R.R., 2011. Food shopping in the urban environment: Parking supply, destination choice, and mode choice. In: 90th Annual Meeting of the Transportation Research Board, Washington, DC.
Martin, A., 2006. Factors influencing pedestrian safety: a literature review. PPR241. TRL Wokingham, Berks.
McCormack, G.R., Shiell, A., 2011. In search of causality: a systematic review of the relationship between the built environment and physical activity among adults. Int. J. Behav. Nutrit. Phys. Activity 8, 125.
Middleton, S.E., Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, Y., 2018. Location extraction from social media: geoparsing, location disambiguation, and geotagging. ACM Trans. Inf. Syst. (TOIS) 36, 40.
Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N., 2011. Understanding the demographics of twitter users. ICWSM 11, 5th.
Noulas, A., Scellato, S., Lathia, N., Mascolo, C., 2012. A random walk around the city: New venue recommendation in location-based social networks. In: Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international conference on social computing (socialcom), IEEE. pp. 144–153.
Panter, J.R., Jones, A., 2010. Attitudes and the environment as determinants of active travel in adults: what do and don't we know? J. Phys. Activity Health 7, 551–561.
Paule, J.D.G., Sun, Y., Thakuriah, P.V., 2019. Beyond geotagged tweets: exploring the geolocalisation of tweets for transportation applications. In: Transportation Analytics in the Era of Big Data. Springer, pp. 1–21.
Peersman, C., Daelemans, W., Van Vaerenbergh, L., 2011. Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents ACM, pp. 37–44.
Pramanik, S., Haldar, R., Kumar, A., Pathak, S., Mitra, B., 2019. Deep learning driven venue recommender for event-based social networks. IEEE Trans. Knowl. Data Eng.
Pratiwi, A.R., Zhao, S., Mi, X., 2015. Quantifying the relationship between visitor satisfaction and perceived accessibility to pedestrian spaces on festival days. Front. Archit. Res. 4, 285–295.
Psyllidis, A., Bozzon, A., Bocconi, S., Bolivar, C.T., 2015. A platform for urban analytics and semantic data integration in city planning. In: International Conference on Computer-aided Architectural Design Futures. Springer, pp. 21–36.
Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. Transp. Res. Part C: Emerging Technol. 75, 197–211.
Rietveld, P., Daniel, V., 2004. Determinants of bicycle use: do municipal policies matter? Transp. Res. Part A: Policy Practice 38, 531–550.
Rodríguez, D.A., Brisson, E.M., Estupi nán, N., 2009. The relationship between segment-level built environment attributes and pedestrian activity around bogota's brt stations. Transp. Res. Part D: Transport Environ. 14, 470–478.
Roy, K.C., Cebrian, M., Hasan, S., 2019. Quantifying human mobility resilience to extreme events using geo-located social media data. EPJ Data Sci. 8, 18.
Ryan, D., Denman, S., Fookes, C., Sridharan, S., 2009. Crowd counting using multiple local features. In: Digital Image Computing: Techniques and Applications, 2009. DICTA'09, IEEE, pp. 81–88.
Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al., 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. PloS one 8, e73791.
Still, G.K., 2000. Crowd dynamics. University of Warwick (Ph.D. thesis).
Still, G.K., 2014. Introduction to crowd science. CRC Press.
Titos Bolivar, C., 2014. City usage analysis using social media.
Tubbs, J., Meacham, B., 2007. Egress Design Solutions: A Guide to Evacuation and Crowd Management Planning. John Wiley & Sons.
Tyshchuk, Y., Wallace, W.A., 2018. Modeling human behavior on social media in response to significant events. IEEE Trans. Comput. Soc. Syst. 5, 444–457.
Van der Waerden, P., Borgers, A., Timmermans, H., 1998. The impact of the parking situation in shopping centres on store choice behaviour. GeoJournal 45, 309–315.
Wang, B., Ye, M., Li, X., Zhao, F., Ding, J., 2012. Abnormal crowd behavior detection using high-frequency and spatio-temporal features. Mach. Vis. Appl. 23, 501–511.
Wegener, M., 2004. Overview of land use transport models. In: Overview of land use transport models, in: Handbook of transport geography and spatial systems. Emerald Group Publishing Limited, pp. 127–146.
Yang, C., Xiao, M., Ding, X., Tian, W., Zhai, Y., Chen, J., Liu, L., Ye, X., 2019. Exploring human mobility patterns using geo-tagged social media data at the group level. J. Spatial Sci. 64, 221–238.
Yang, J., Hauff, C., Houben, G.J., Bolivar, C.T., 2016. Diversity in urban social media analytics. In: International Conference on Web Engineering. Springer, pp. 335–353.
Yang, Y., Heppenstall, A., Turner, A., Comber, A., 2019. Who, where, why and when? using smart card and social media data to understand urban mobility. ISPRS Int. J. Geo-Inf. 8, 271.
Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., Badr, S., 1993. The occurrence of sleep-disordered breathing among middle-aged adults. N. Engl. J. Med. 328, 1230–1235.
Zahran, S., Brody, S.D., Maghelal, P., Prelog, A., Lacy, M., 2008. Cycling and walking: explaining the spatial distribution of healthy modes of transportation in the united states. Transp. Res. part D: Transport Environ. 13, 462–470.
Zhan, B., Remagnino, P., Velastin, S., Bremond, F., Thonnat, M., 2006. Matching gradient descriptors with topological constraints to characterise the crowd dynamics.
Zhou, E., Cao, Z., Yin, Q., 2015. Naive-deep face recognition: Touching the limit of lfw benchmark or not? arXiv preprint arXiv:1501.04690.
Zomer, L.B., Daamen, W., Meijer, S., Hoogendoorn, S.P., 2015. Managing crowds: the possibilities and limitations of crowd information during urban mass events. In: Planning Support Systems and Smart Cities. Springer, pp. 77–97.