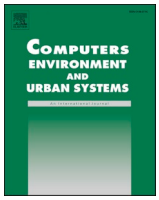
Computers, Environment and Urban Systems 85 (2021) 101551

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



Activity knowledge discovery: Detecting collective and individual activities with digital footprints and open source geographic data

Xinyi Liu [a,*], Qunying Huang [a,*], Song Gao [a], Jizhe Xia [b]

[a] Department of Geography, University of Wisconsin-Madison, 550 N Park St, Madison, WI 53706, USA
[b] School of Architecture & Urban Planning, Shenzhen University, Shenzhen, China

ARTICLE INFO

Keywords:
Activity space
Movement patterns
Human mobility
Spatial clustering
Multi-level clustering
Urban study
Trajectory knowledge discovery

ABSTRACT

Digital footprints collected from social media platforms are often clustered using methods such as the density-based spatial clustering of applications with noise (DBSCAN) and its variants to identify daily travel activities (e.g., dwelling, working, entertainment, and eating). However, these clustering methods mostly only consider the spatial distribution of travel activity points while ignoring their geographic context, resulting in the aggregation of digital footprints representing different activity types into one cluster. In addition, existing works only focus on examining people's travel activities at either the collective (i.e., macro) or individual (i.e., micro) level. To this end, this study utilizes geographic context information and develops a novel activity knowledge discovery framework to better detect frequent travel activities at both levels. First, we develop a multi-level spatial clustering method to aggregate digital footprints of a group of users into collective clusters (i.e., activity zones) by inferring and integrating the underlying activities performed at each zone with OpenStreetMap (OSM) datasets that can inform geographic context of the activity zones. Next, we introduce a location-aware clustering method to detect activity zones and associate activity types at the individual level by aggregating individual footprints based on the collective results. As case studies, digital footprints from 49 selected Twitter users are analyzed to evaluate the proposed framework. The results reveal that: (1) The multi-level spatial clustering method can often detect significant collective activity zones; and (2) The location-aware clustering method can aggregate individual digital footprints into activity zones more effectively compared with existing density-based spatial clustering methods (e.g., DBSCAN).

1. Introduction

Examining people's movement and activity patterns can benefit infrastructure construction and policymaking related to transportation, urban planning, disease control, and disaster management (Batty et al., 2012; Cheng et al., 2011; González, Hidalgo and Barabási, 2008; Song et al., 2010; Zhao et al., 2016). It is also essential for studying a variety of social problems, such as social segregation (Huang & Wong, 2015; Xu et al., 2018). To explore these topics, individual travel data have been collected traditionally through surveys, which is tedious and expensive. Recently, the advancement of smart devices has allowed collecting travel data by passively capturing a vast amount of tracking records from cell phones, GPS devices, and social media platforms. Among them, social media messages often include GPS locations as geo-tags to represent where they are posted (Blanford et al., 2015) and are increasingly used due to several unique advantages (Huang and Wong 2015; Yuan, Liu and Wei, 2017; Tu et al., 2017): (1) public availability, (2) capturing long-term trajectories, (3) a large number of participants, and (4) nearly real-time data availability.

Digital footprints collected from social media platforms are typically recorded as sequences of locations with timestamps to represent individual trajectories (Zheng, Zha and Chua, 2012). Using them to study human movement and patterns presents two major challenges: (1) Digital footprints record not only regular activities at sparse and irregular time intervals (Cao et al., 2014), but also random movements over space and time (González et al., 2008; Song et al., 2010). The captured individual travel patterns are thus not immediately visible and identifiable (Cheng et al., 2011; Gao and Liu, 2014); and (2) Digital footprints do not carry semantic information that can describe people's activities, such as the purpose of people's visit to a location (e.g., going to work), and context of the location (e.g., office), which is paramount for interpreting these data (Alvares et al., 2007).

* Corresponding authors.
E-mail address: xliu636@wisc.edu (X. Liu).

https://doi.org/10.1016/j.compenvurbsys.2020.101551
Received 28 January 2020; Received in revised form 25 September 2020; Accepted 28 September 2020
Available online 22 October 2020

To address these challenges, an important research topic, which we define as activity knowledge discovery (AKD), emerges for human mobility analysis with digital footprints (Huang and Wong 2015; Comito, Falcone and Talia, 2016; Chaniotakis et al., 2017). Specifically, AKD includes two tasks: (1) identifying travel trajectory clusters, also known as activity space or zones where people frequently visit, and (2) inferring the activity type (e.g., dwelling, working, and entertainment) performed at each zone. The former task can handle the sparseness and irregularity of digital footprints by aggregating them into clusters to represent regular daily activities. The latter offers semantic information of each footprint and travels between footprints, which improves the understanding of individual travel patterns and enables analyzing mobility patterns of groups with different socioeconomic status (Huang & Wong, 2016). AKD is different from semantic trajectory knowledge discovery in previous works (Alvares et al., 2007; Chaniotakis et al., 2017; Comito et al., 2016; Tu et al., 2017), which focus on activity type inference of a set of locations, without optimizing the footprint aggregation process to enable the detection of most frequent activities.

Currently, various clustering methods (Chen et al., 2011; Gao et al., 2018; Linton et al., 2014; Zhao et al., 2017) are developed to identify activity zones and unveil the complexity of millions of individual footprints. However, these methods require relatively dense and successive footprints to identify individual activities and thus are mostly applicable for traditional datasets or GPS trajectories extracted from cell phone logs. As such, these methods often fail to detect frequently visited locations from digital footprints (Huang, Cao, & Wang, 2014), which are sparser and contain more irregularities (Cao et al., 2014). In addition, these methods mostly consider the spatial distribution of points while ignoring their geographic context, resulting in the aggregation of digital footprints representing different activities (e.g., eating and education; Fig. 1b) into one cluster (Fig. 1c).

To infer activity types, researchers leveraged multiple types of data, including movement patterns and behaviors (e.g., transitions among locations; Wu et al., 2014), temporal information (e.g., stay time at a location; Steiger, Resch and Zipf, 2015), geographic and environmental context (e.g., land use data, POIs; Huang and Wong, 2016; Jiang et al., 2015), and textual information (e.g., content of posted messages or user profiles; Preoţiuc-Pietro and Cohn, 2013; Chaniotakis et al., 2017). For example, Huang and Wong (2016) inferred home area of social media users as the residential land use zone with the largest number of transitions, as well as workplace as the commercial zone with the largest number of transitions. Several other activities (e.g., entertainment, education) in a particular area of interest (AOI) were inferred through incorporating land use data and Google Place Services (Huang, Cao, & Wang, 2014). However, individual representative daily travel trajectory with detailed and accurate activity information was not yet produced. Furthermore, the land use data are typically published by a local authoritative organization and need to be collected and processed case-by-case for different regions.

Finally, people's travel activities are mostly detected at either the collective (i.e., macro) or individual (i.e., micro) level to investigate human mobility patterns (Comito et al., 2016). Collective activities are typically detected as travel hot-spots by aggregating trajectory datasets of a group of users (Wu et al., 2014; Yin et al., 2011; Zhang et al., 2014; Zhou and Zhang, 2016), while individual activities are delineated as movement behaviors of an individual over a given period of time (e.g., one day, one year) (Wu et al., 2014; Huang and Wong, 2015; Hasan and Ukkusuri, 2014). Although Comito et al. (2016) examined travel activity patterns at both levels, their application of density-based clustering for spatial aggregation and online check-ins as activity type labels can lead to inaccurate or absent detection of frequent activities.

To this end, this paper integrates open source geographic data to inform geographic context and optimize the aggregation of digital footprints, and develops a two-step AKD framework to detect both collective and individual frequent travel activities. Within this framework, collective activity zones are first derived through a multi-level spatial clustering method based on footprints of all users, which innovatively integrates semantic information of activity types inferred with OpenStreetMap (OSM) datasets. Next, individual travel trajectories are aggregated using a location-aware clustering method based on the semantic collective activity zones to generate a set of semantic individual frequent activity zones. Using geo-tagged tweets of 49 selected users in Madison, Wisconsin (U.S.), we tested the application of both clustering methods in identifying collective and individual activity zones. Finally, we evaluated the aggregation and activity type identification results of location-aware clustering by comparing with manually labeled results (i.e., ground truth), as well as outcomes generated by other methods (i.e., DBSCAN and M-DBSCAN). To sum up, the contributions of this work are highlighted in three areas:

1. We propose a two-step framework to effectively detect both collective and individual activity zones and activity types. To our best knowledge, this work is the first to examine people's frequent activities at both levels.
2. The activity zone detection methods, with multi-level spatial clustering for the collective level, and location-aware clustering for the individual level, consider the geographic context and enable the accurate detection of travel activities from sparse digital footprints.
3. With open source OSM datasets, we provide a detailed inference of activity types (ten in total; Table 3), contributing to the comprehensive understanding and analysis of human movement behaviors.

2. Related work

2.1. Spatial clustering methods for activity zone detection

People's representative travel activities are typically denoted as clusters that are generated from their travel trajectories using aggregation methods (Lu et al., 2011). To date, researchers have developed a variety of clustering algorithms for aggregating spatial data points with noise, including partitioning algorithms, hierarchical algorithms, density-based algorithms, grid-based algorithms (Maciąg, 2017), and so on. Among these methods, partitioning algorithms predefine the number of clusters and can only detect clusters of round shapes (Velmurugan and Santhanam, 2011). Hierarchical algorithms either gradually merge smaller clusters or divide larger clusters to reach a predefined number of clusters (Karypis, Han and Kumar, 1999). Grid-based algorithms first divide data space into cells and then merge certain cells based on

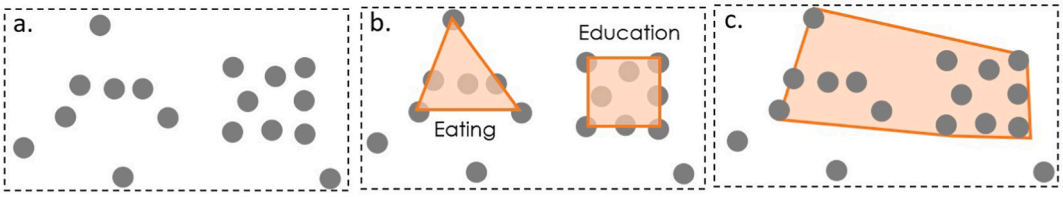


Fig. 1. Spatial clustering without considering geographic context.

predefined conditions (Cai, Lee and Lee, 2016). Graph-based algorithms achieve automatic clustering based on Voronoi modelling and Delaunay Diagrams (Estivill-Castro and Lee, 2000).

Compared with other spatial clustering methods except for the graph-based ones, density-based algorithms can detect arbitrary cluster shapes and need less predetermined input parameters that rely on domain knowledge (Ester et al., 1996). Particularly, DBSCAN can efficiently aggregate a large set of sparse digital footprints (Kang et al., 2010; Zheng et al., 2012; Huang, Cao, & Wang, 2014; Huang and Li, 2018), and thus becomes widely used in aggregating footprint trajectories for activity zone detection (Marques, 2014; Huang and Wong, 2015; Chaniotakis et al., 2017). Although graph-based algorithms (Estivill-Castro and Lee, 2000) and grid-based algorithms (Cao et al., 2014) can produce spatial clusters of arbitrary shapes and varied densities, they are more complicated while less efficient to deal with noise when processing sparse dataset. This study therefore discusses the optimization of spatial clustering for activity detection based on DBSCAN.

DBSCAN clustering results vary with different *eps* (neighborhood radius) and *minPts* (the minimum number of points required to form a dense region) as input parameters (Bäcklund, Hedblom and Neijman, 2011). One single *eps* often fails to detect clusters of different densities (Gong et al., 2015; Liu, Zhou and Wu, 2007). Several modified algorithms based on DBSCAN thus have been designed to overcome this limitation, such as VDBSCAN (Liu et al., 2007), HDBSCAN (Campello et al., 2015), M-DBSCAN (Liu, Huang, & Gao, 2019), and OPTICS (Ankerst et al., 1999). These methods can detect clusters of varying densities based on the spatial distribution of digital footprints. Hierarchical algorithms can also detect spatial clusters with different densities, while the criteria to merge or split clusters needs to be explicit (Karypis et al., 1999).

However, clustering digital footprints for activity zone detection not only depends on their spatial distribution (Fig. 1). For example, a large spatial cluster should be collapsed into smaller ones when they are geographically close to each other but represent diverse activities; Spatially adjacent digital footprints surrounding a dining hall and an academic building should be clustered as an eating zone and an education zone respectively (Fig. 1b), instead of being considered as a uniform activity zone (Fig. 1c). Therefore, our proposed framework implements the aggregation of activity footprints with a similar idea as the hierarchical clustering and uses activity type as the criteria to merge adjacent clusters (i.e., clusters of the same activity type are merged).

2.2. Inference of travel semantics

Individual travel trajectories denote a series of places people visit along a timeline. These places (e.g., home, workspace, and park) reflect people's corresponding activities (e.g., dwelling, work, and entertainment), which are discussed as semantic knowledge and could be implicit under raw digital footprints (Cai et al., 2016; Yan et al., 2013). Semantic inference methods can be differentiated by the source and representation of semantic knowledge. A typical approach is to represent travel semantics with manually defined activities using the geographic coordinates and content of user-generated place labels, such as Foursquare labels (Hasan and Ukkusuri, 2014). In these methods, boundaries of activities (i.e., activity zone) are not defined while could be useful (Chaniotakis et al., 2017). Another approach is to detect activity zone (e.g., location or region; Huang, Cao, & Wang, 2014) at first and then to associate an activity zone to an activity type using predictive models which are calibrated with digital footprints' metadata (e.g., content, timestamp; Yang et al., 2014, Steiger et al., 2015, Aurelio Beber et al., 2016).

Additionally, the underlying geographic context of digital footprints is directly applied to improve activity detection accuracy. For example, Huang, Cao, & Wang (2014) identified the activity type of activity zones based on a regional land use maps and Google Places application programming interface (API), which returns information about a place given the place location. Du et al. (2016) imposed geographic background constraints on density-based clustering to yield more relevant clustering results, such as separating a cluster which spans over a river into two clusters. Alternatively, some studies aggregate geographic context data into semantic types and then label trajectory segments with these types. For example, Cai et al. (2016) identified semantic clusters representing people's activities by aggregating outside POIs which indicate geographic contexts. Semantics of individual trajectories is thus detected by searching their spatially overlapped semantic clusters. Jiang et al. (2015) estimated land use types at the census block level by classifying POI types with a machine learning model and identified local individual activities based on the land use distribution.

2.3. Activity identification using digital footprints and geographic data

Using user-generated place labels for activity identification is insufficient simply because they are always not available and often miss frequent activities as people usually check in at limited types of locations (Hasan and Ukkusuri, 2014). Semantic metadata attached to digital footprints are usually leveraged via topic modelling (Steiger et al., 2015), which generate semantic clusters based on topic hotspots. However, this approach was unable to establish a complete activity classification mechanism for daily activity identification.

While incorporating underlying geographic context for activity detection (Huang, Cao, & Wang, 2014; Du et al., 2016), officially published datasets are often difficult and time-consuming to acquire. Officially published land use data are mostly defined for a small area (e.g., city-level), and researchers need to search for specific data based on their study areas. Moreover, detailed land use data for statewide or citywide regions are produced in different standards, and thus create various land use type subdivisions. Meanwhile, commercial geographic services (e.g., Google Places API), require API keys to be generated before people can use these interfaces, and allow limited number of free-charged requests within a period (Huang, Cao, & Wang, 2014). Therefore, open source data are increasingly leveraged (Estima and Painho, 2013; Fonte et al., 2015). For example, Cai et al. (2016) used GeoNames, an accessible gazetteer database, which contains a large amount of qualified data to identify geographic contexts (Ahlers, 2013). Jiang et al. (2015) widely collected open source POIs, including those from Yahoo!. However, the accuracy and coverage of open source geographic data need to be considered (Zielstra and Zipf, 2010).

POIs indicating geographic context are mostly clustered with grid-based algorithms (Cai et al., 2016; Jiang et al., 2015). Nevertheless, it is computationally intensive for classification at small scales (e.g., distinguish different activities happening in neighboring buildings) as the grid needs to be finely divided for multiple times to detect arbitrary shapes (Hio et al., 2013). Moreover, these methods are sensitive to parameters such as the size of cell, which is hard to specify (Njoo et al., 2015). Furthermore, for sparse digital footprints, it is inefficient to identify semantics of every grid cell, since only a small portion of them are covered by trajectory points.

Additionally, only a few clusters can be detected for most individuals by directly aggregating their travel trajectories due to data sparsity. Meanwhile, each cluster often takes a small area, which can rarely provide sufficient context information to identify associated activities. For example, a person may regularly visit a restaurant and tweet at one fixed location outside the restaurant. As a result, the geographic feature of the restaurant, typically represented as a POI, is outside the spatial extent of these tweets. As the geometry of this tweeting area (i.e., zone) may not overlap any eating place or zone, its activity type can hardly be inferred without referring to other resources such as message content.

In sum, previous methods for identifying activity zone types using social media data are not sufficient at a collective level with relatively dense data due to the lack of semantic information. They show extra problems when dealing with activity identification at an individual level

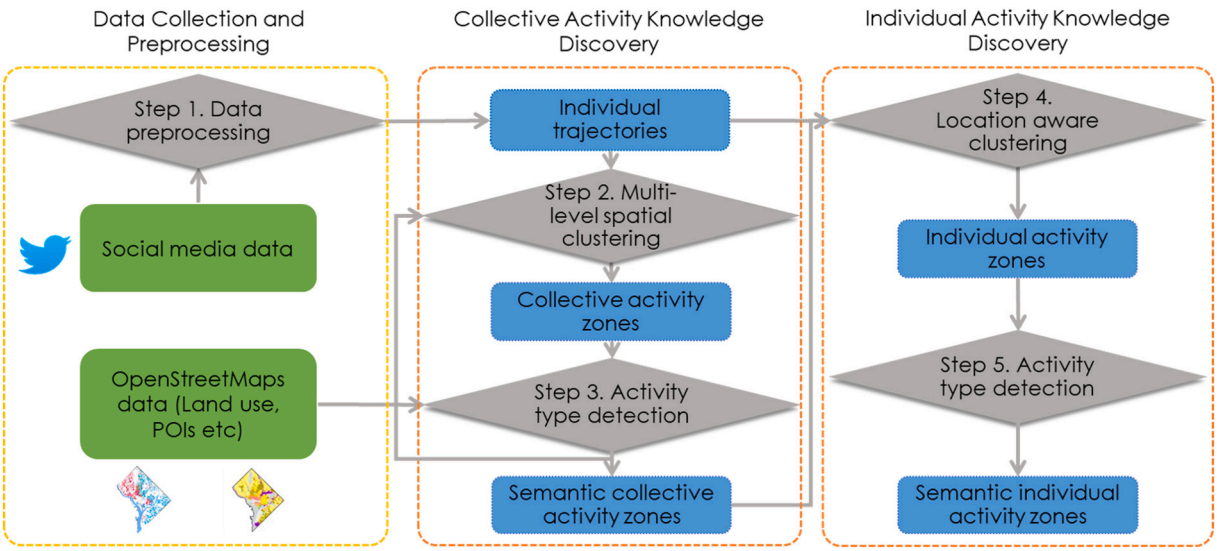


Fig. 2. Workflow of activity knowledge discovery at both collective and individual levels.

due to data sparsity. We propose a framework of activity knowledge discovery to detect frequent activities and determine activity zone types at both levels more effectively with sparse digital footprints.

3. Methodology

Using geotagged tweets as an example, this research proposes a framework to first discover collective activity zones and types by fusing OSM datasets. Individual activity zones and types are subsequently detected. This framework includes three components and five primary steps (Fig. 2): (1) Data pre-processing removes inactive or abnormal users (e.g., users representing a company), and represents users' tweet records as a set of *spatiotemporal* (ST) points with each point representing individual presence (i.e., footprint) at a location and a time point; (2) Multi-level spatial clustering aggregates ST points of all selected users into collective activity zones and derives the centroid of each cluster as a representative location where corresponding activities are performed; (3) Activity type identification detects the activity type (e.g., dwelling, work, entertainment, and eating) of each zone by integrating multiple datasets from OSM and merges adjacent activity zones of the same activity type into semantic collective activity zones; (4) Location-aware clustering aggregates every individual travel trajectory into a series of representative individual activity zones based on the spatial relationship between each individual footprint and the detected collective activity zones; and (5) Representative activity type of each individual zone is detected based on its attached collective zone type.

3.1. Data preprocessing

Using Twitter streaming API, users are identified within a geographic boundary, and their historic data are then collected with Twitter Search API. Users with more than 50 geotagged tweets are selected to generate representative daily travel paths (Chaniotakis et al., 2017). Twitter accounts representing organizations or companies are shared by many individuals and are thus removed, which are identified as users at a relocation speed over 240 m/s (Yin and Wang, 2016). Next, intertown travels of the valid users are discarded as this study focuses on local travel patterns. Specifically, a box boundary is predefined to remove invalid footprints located outside the target area, and to exclude corresponding long-distance travels. The remaining geotagged tweets are then processed and organized as a set of ST points to indicate individual trajectories.

3.2. Collective activity knowledge discovery

3.2.1. Activity zone identification with multi-level spatial clustering

In this paper, a multi-level spatial clustering method based on DBSCAN with varying *eps* is implemented to aggregate digital footprints of a group of users into collective activity zones of different densities. Specifically, three distinct *eps* values are first used as DBSCAN input to detect potential clusters at different density levels. Considering the common *eps* values used in the literature, 50 m, 100 m, and 200 m are selected; while 50 m can detect relatively small clusters capturing collective activities in a compact space (Hwang, Hanke and Evans, 2018), *eps* as 200 m is reported to be able to identify most AOIs in cities using geotagged Flickr photos and tweets (Hu et al., 2015; Liu, Huang, & Gao, 2019).

After applying DBSCAN on the ST points of all users at three different levels (Algorithm 1 line 1), a set of candidate clusters (C1, and C2, and C3) are produced. Next, a series of activity zones (Z1, and Z2, and Z3) are generated (Algorithm 1, line 2) based on these candidate clusters. Specifically, a convex hull represented as a polygon feature is derived from all points within each spatial cluster Huang & Li et al., 2016 to reveal its geometrical features (e.g., location, shape, and range) and to represent the activity zone. The geometric median for each spatial cluster is calculated to denote the cluster centroid. Since at least three different points are required to generate a convex hull, the points representing cluster centroids are used to denote activity zones containing less than three different locations.

Algorithm 1: Multi-level spatial clustering with geographical context

Input: ST point set

Output: activity zone set (Z^{I}) with activity type (T) identified

1: generate cluster set C^{I} (*eps*: 50m), C^{II} (*eps*: 100m) and C^{III} (*eps*: 200m) using DBSCAN (*minPts*: 4)

2: generate activity zone set Z^{I}, Z^{II}, Z^{III} by deriving the convex hull and the centroid of each cluster in C^{I}, C^{II}, C^{III}

3: identify activity zone type (T) of each Z in Z^{I}, Z^{II} and Z^{III} using Algorithm 2

4: for N in {2, 3}

5: for each Z in Z^{N-1} and each Z' in Z^{N}

6: if $T' = T$ and Z' overlaps Z

7: replace Z with Z'

8: if T' != 'Others' and Z' does not overlap Z

9: add Z' to Z^{I}

10: return Z^{I}

Algorithm 2: Identify zone type for an activity zone

Input: activity zone (Z) and OSM datasets (Table 1)

Output: activity zone type (T)

1: for each OSM land use feature F within the target area

2: if F overlaps Z

3: get the T of F ($T = t$) referring to Table 2

4: set default zone type of Z as t

5: for each other OSM feature F within the target area

6: if F overlaps Z

7: get the T of F referring to Table 3

8: if T = "unknown"

9: $T = t$

10: increase the count of T by 1 in potential zone type map $TM = (T, \mathrm{count})$

11: sort entries in TM based on the count of T in descending order

12: pick the first T in TM

13: return T

Next, the associated zone type of each activity zone is identified with Algorithm 1 (Algorithm 1, line 3; Section 3.2.2) by selecting optimal clusters at different density levels to represent collective activity zones. Specifically, we iterate clusters generated at each density level using a nested loop (Algorithm1, lines 4–5). Within each inner loop, the type of each activity zone is checked and compared with the surrounding activity zones detected with the *eps* at the next level. Smaller activity zones are replaced by a larger one at the next level that shares the same zone type and overlaps these smaller zones (e.g., Fig. 3a; Algorithm1, lines 6–7). However, if they have different activity types (Fig. 3b and c), the algorithm will not merge them (i.e., the larger activity zone will not replace smaller ones).

Larger activity zones with zone type identified without overlapping any activity zone detected at the previous level are also kept (Algorithm1, lines 8–9). This is because mergence could improperly aggregate diverse activity zones into a generalized one (Fig. 3b and c), where important activity zones (i.e., Z_a in Fig. 3b; Z_a, and Z_b in Fig. 3c) are merged as part of the large surrounding activity zone (i.e., zone Z^N) detected with the next-level *eps*. As a result, their associated activity zone types are generalized (e.g., Type 2 in Fig. 3b, Type 1 and 2 in Fig. 3c) and cannot be accurately identified. This mergence process eventually generates collective activity zones from three sets of candidate activity zones.

3.2.2. Activity type inference

As an open data source, OSM increasingly contributes to location collections, and provides comprehensive location sets with detailed information such as location types and names. For example, more than 20 OSM records while less than 5 from GeoNames could be detected within an activity zone based on our experiment. In addition to point data (e.g., POI dataset), OSM also contains data (Table 1) of other geometries such as polygon (e.g., street building dataset and water area dataset) and polyline (e.g., waterways dataset) to be able to improve detection resolution. Therefore, this work utilizes OSM datasets (Table 1) to identify activity zone types (Fig. 4).

Specifically, ten activity types are predefined to capture the majority of people's daily activities in an urban area (Huang, Cao, & Wang, 2014; Chaniotakis et al., 2017), including *Eating, Shopping, Education, Work, Health, Entertainment, Service, Dwelling, Transportation, and Transportation network* (Table 3). OSM land use datasets are labeled with vague feature classes, thus are mapped to eight general types (i.e., land use types; Table 2) to produce a land use layer corresponding to the ten activity types (Fig. 4 Step 1). A spatial join operation is performed on all

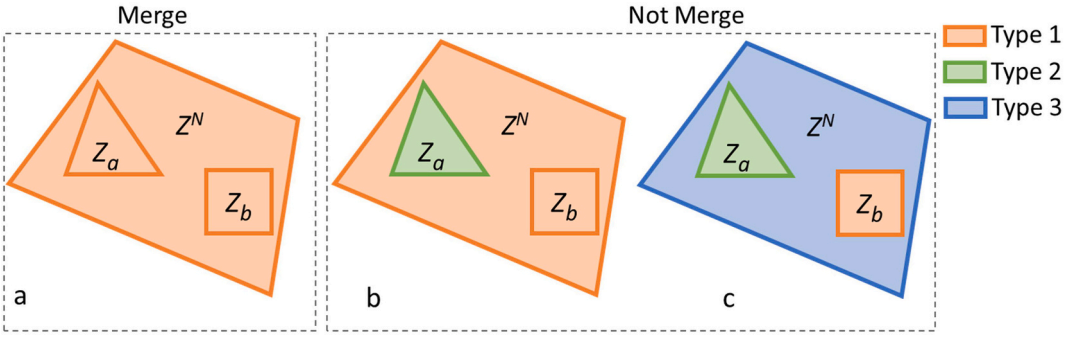


Fig. 3. Activity zone detection at current level and the next level: activity zone Z_a,$Z_b \in Z^{I}$or Z^{II}, detected with smaller *eps* at current level (e.g., *eps* = 50 m); activity zone $Z^N \in Z^{II}$or Z^{III}, detected with larger *eps* at the next level (e.g., *eps* = 100 m).

Table 1
OSM dataset description (please refer to Table 1 in appendix for complete features classes).

Dataset	Feature geometry	Distinct feature class
landuse	polygon	residential, park, …, retail, heath, cemetery, industrial
pois	point	dentist, college, cinema, …, restaurant, bank, hotel,
pois_a	polygon	laundry, florist, track, camera_surveillance
pofw	point	jewish, hindu, …, muslim
pofw_a	polygon	
natural	point	cliff, beach, peak, spring, …, glacier, volcano, tree
natural_a	polygon	
traffic	point	waterfall, dam, fuel, stop, …, street_lamp,
traffic_a	polygon	traffic_signals
transport	point	railway_halt, tram_stop, …, ferry_terminal, taxi,
transport_a	polygon	bus_stop
buildings	polygon	apartment, cinema, clinic, …, mall, station, offices, empty
water	polygon	water, wetland, dock, reservoir, river
waterways	polygon	river, stream, drain, canal

activity zones and the land use layer to identify the land use type of each zone (Step 3), which in turn can indicate whether an activity type is considered as *Dwelling* (e.g., the activity performed within the area of the residential land use type) or other seven general types (e.g., *Work* and *Entertainment*).

Given an activity zone, the associated land use type helps determine its activity type. However, three activity zone types, including *Eating*, *Transportation*, and *Transportation network*, are missing in land use types. Furthermore, commercial land use type cannot be directly mapped to an activity type as it could be used for diverse activities (e.g., *Eating* and *Entertainment*). Moreover, many land use types could include a large area with mixed activities (e.g., *Dwelling* area may consist of *Eating* activities). Therefore, additional OSM datasets (Table 1) are mapped to establish POI layers with activity zone types (Table 3; Fig. 4 Step 2) and spatially joint with activity zones to further indicate relevant activities (Step 3). These supplementary datasets include poi, building, and water features, which amplify the type reference for the activity zones.

POIs of an unknown feature class (e.g., blank or meaningless class name) are usually anonymous buildings (e.g., buildings at a *Shopping* area), thus the land use type identified for the activity zone is considered as its representative zone type (Fig. 5). Activity zones without POIs included are also assigned its land use type if applicable (i.e., the land use type is identical to one of the ten activity zone types). The type of an activity zone is determined by the maximum votes of the activity zone types associated with all its included OSM POIs (Fig. 4 Step 3). However, the most frequently appearing POI type is not necessarily selected. Specifically, ten activity zone types are classified into four priorities (Table 3) to maximize the identification accuracy of activity zone types based on our experiments. Activity types of lower-level priorities are more specific and could be distributed at smaller scales (i.e., areas of smaller size), which could also be conducted at activity zones of higher-level-priority types. For example, *Eating* activities could occur in the campus primarily for *Education* activities; Pharmacies and retails can be located at a *Dwelling* area for people's convenience. Based on the priority, activity zone types of higher priorities (e.g., priority I > priority II) should be taken before considering others of lower priorities, reducing the possibility of misclassifying finely divided zones as broader or large-scale types.

3.3. Individual activity knowledge discovery

Next, an improved clustering method, defined as location-aware clustering, is used to generate activity zones for individuals (Fig. 2). Specifically, given an individual ST point (e.g., point a, b, c, d, e in Fig. 6), a buffer zone with a radius of 200 m (the same as the largest *eps* to detect activity zones) is generated to search for spatially overlapped

Table 2
Mapping of land use types and OSM land use feature classes.

Land Use Type	Distinct Feature Class of OSM Land Use Dataset
Dwelling	residential
Entertainment	recreation_ground, vineyard, park, orchard
Health	health
Commercial	commercial
Service	cemetery
Shopping	retail
Work	farm, meadow, military, industrial, quarry
Others	scrub, nature_reserve, grass, forest, allotments

Table 3
Mapping of activity zone types and feature classes of OSM POIs in point/polygon format (please refer to Table 2 in appendix for activity zone type definition and complete OSM features classes).

Activity Zone Type	Priority	OSM Distinct Feature Class
Eating	I	bakery, cafe, fast_food, …, restaurant, food_court, caboose, cafeteria
Work		embassy, police, bank, …, community_centre, courthouse, factory, barn, industrial
Shopping	II	beauty_shop, bookshop, …, clothes, beverages, bicycle_shop, butcher, computer_shop, florist
Education		college, kindergarten, …, library, school, academic, university, academic_building
Health		dentist, doctors, …, hospital, pharmacy, chemist, optician, nursing_home
Entertainment	III	alpine_hut, artwork, …, archaeological, attraction, arts_centre, bar, battlefield, bench, biergarten
Service		bicycle_rental, atm, …, post_box, car_repair, fort, car_wash, graveyard, guesthouse, hostel, laundry
Dwelling		dormitory, houses, …, house, family_house, shed, condominium, townhouse, apartment, home
Transportation	IV	airplane, terminal, …, station, train_station, track, bridge, transportation, car_park
Transportation network		parking_shelter, …, camera_surveillance

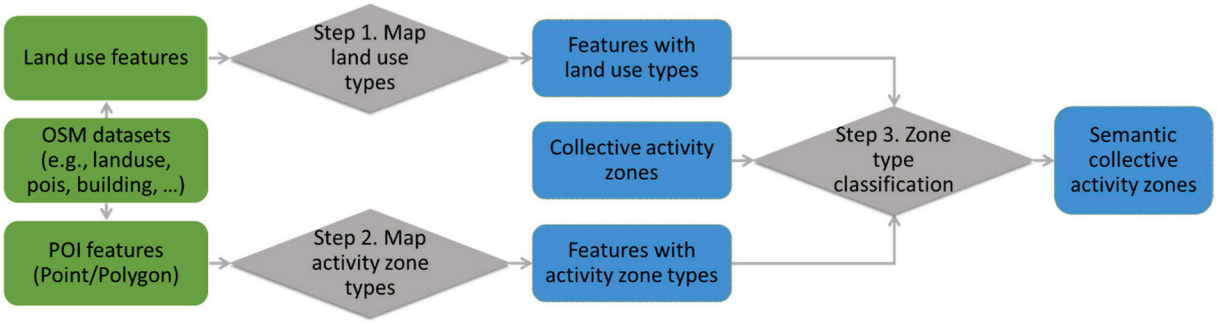


Fig. 4. Workflow of activity type identification.

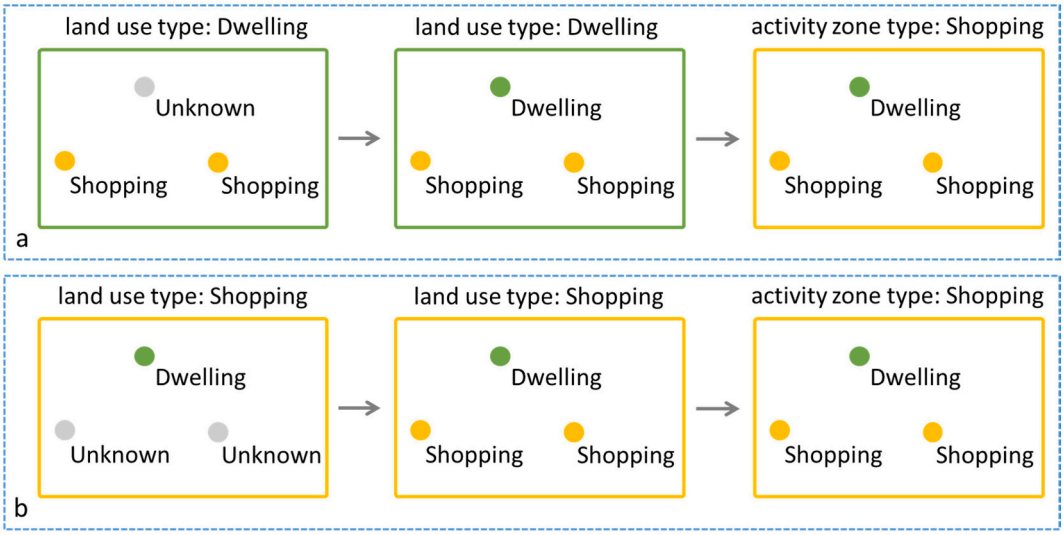


Fig. 5. Activity zone type identification in different scenarios (a. Mapping a POI of unknown type to *Dwelling* type; b. Mapping two POIs of unknown type to *Shopping* type).

collective activity zones (Section 3.2; zone A, B in Fig. 6). An individual point is attached to the collective zone which has the largest overlap with its buffer zone. For example, the buffer zone of point c is overlapped with both zone A and zone B, and then point c is attached to zone A as it has a larger overlap with the buffer zone of point c. ST points without any activity zone attached (e.g., point e in Fig. 6) are considered as noise and discarded. ST points attached to the same activity zone (e.g., points a, b, and c) are grouped as a distinct cluster, defined as an individual activity zone.

The activity type of individual activity zone is inferred based on its attached collective zone type (Fig. 6). Note that an individual activity zone should be classified as *Dwelling* type if it, as the largest cluster of this individual, is overlapped with both *Dwelling* and *Entertainment* land use areas. This is because people usually spend most of their time in *Dwelling* zones and post lots of messages. *Entertainment* land use type is often mapped by parks, which is closely related to home locations.

Each derived cluster indicates an activity zone that an individual visits, which is defined as a semantic individual activity zone. However, not every individual zone represents a regular travel activity of the individual. In fact, many zones capture random activities. Specifically, if an individual activity zone includes a large number of points, it indicates that this individual frequently visits the zone in a typical day Huang & Li et al., 2016. To differentiate regular zones and random ones, the minimum number of points for each zone, similar to the parameter of *minPts* in DBSCAN, is defined to remove random ones. As discussed in previous research Huang & Li et al., 2016, we use 4 as the minimum number. After removing insignificant zones, semantic individual frequent activity zones are derived.

4. Results

This study uses publicly accessible digital footprints, collected as geotagged tweets via Twitter's streaming API, to identify frequent activity zones with semantics at both collective and individual levels. A total of 115,065 tweets posted by 13,922 users within the geographic boundary of Madison, Wisconsin from September 2013 to June 2015 were archived. Next, 49 unique eligible users were selected for our experiments with each user having posted more than 50 tweet records in total, since these users have adequate trajectory points to unfold their activity patterns (Section 3.1).

4.1. Detection of semantic collective activity zones

After aggregating all ST points representing these users' geotagged tweets using the multi-level spatial clustering method discussed in Section 3.2, 345 activity zones are detected, and 330 of them are identified with an activity type (Table 3). Compared with the base map provided by Google Maps API (Fig. 7b), most activity zones are detected

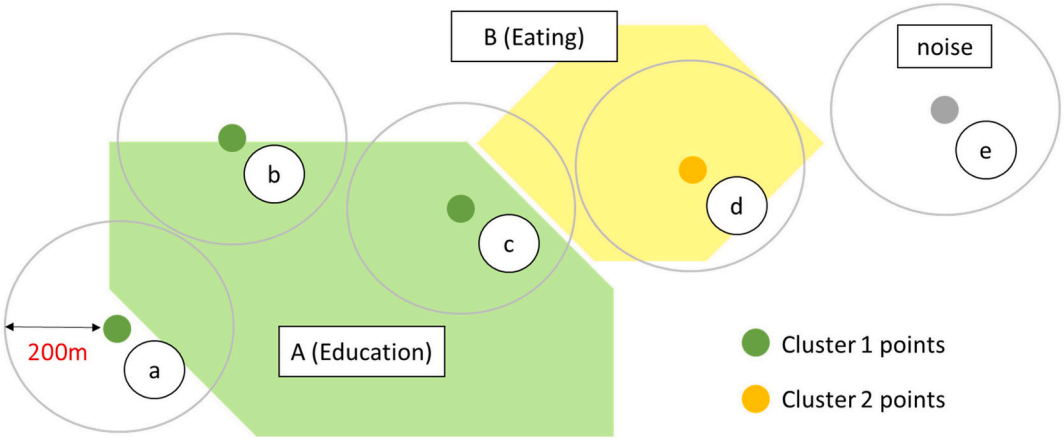


Fig. 6. Demonstration of location-aware clustering.

with the zone types that convey the same context information as the base map (Fig. 7a).

In Fig. 7a showing the detection results surrounding a campus area, polygons in different colors represent collective activity zones with the symbols indicating different activity types. Although some small zones are adjacent to each other, they are not replaced by a single large zone detected with a larger *eps* since the large zone is not of the same zone type. At the lower-left side of Fig. 7, many zones where academic buildings are located are correctly identified as *Education* types. Outside the *Education* zones, especially on the east side of the campus close to the downtown area of the city, many *Shopping* and *Eating* activities are detected. Further outside are dormitories, apartments, and houses mostly for students and faculties, where many *Dwelling* zones are detected. These detections are consistent with the manual interpretation based on the base map as well as in-person observations and understanding of the whole area. The detected semantic collective zones are used to identify semantic individual activity zones in the next subsection, and the evaluation result (Section 5) indicates that our classification mechanism based on DBSCAN and only OSM datasets are applicable and efficient.

4.2. Detection of individual activity zones

Individual travel trajectories are clustered based on their overlapped collective activity zones. Fig. 8a shows seven clusters (clusters A, B, C, D, E, F, and G) detected for a typical user using the proposed location-aware clustering method, with each cluster displayed in a distinct color. Among these clusters, clusters A, B, C, D, E, and F are also detected with DBSCAN (Fig. 8b, c, and d). However, some points relatively far away from the main cluster, are classified as noise. Moreover, cluster G is spatially scattered and not detected by the DBSCAN method until *eps* is increased to be 300 m (Fig. 8d), which is extremely large and runs the risk of merging clusters representing different activities or integrating random activity points (i.e., noise) into clusters.

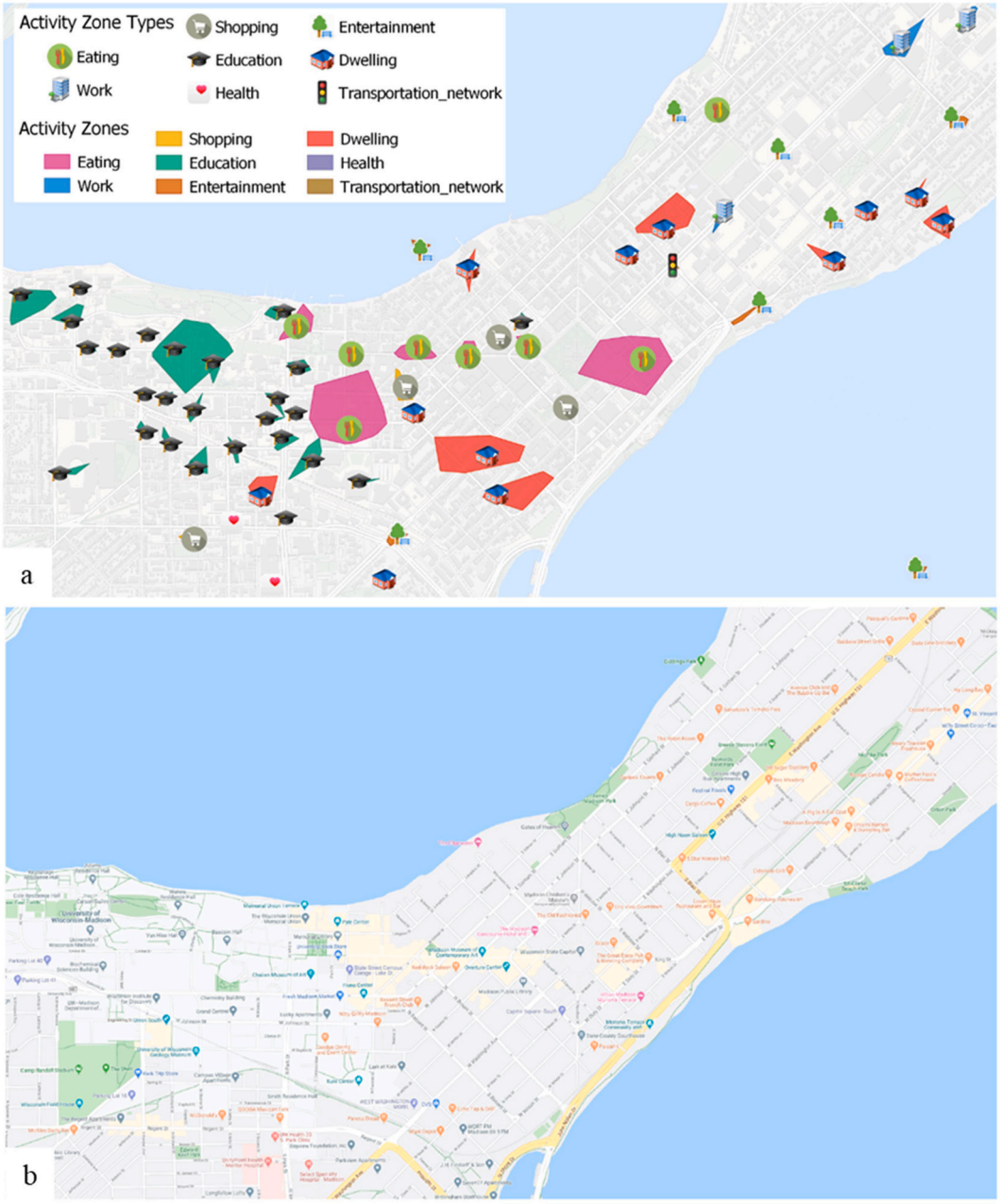


Fig. 7. Visualization of semantic collective activity zones around a campus (a) and geographic context of the area by Google Maps (b).

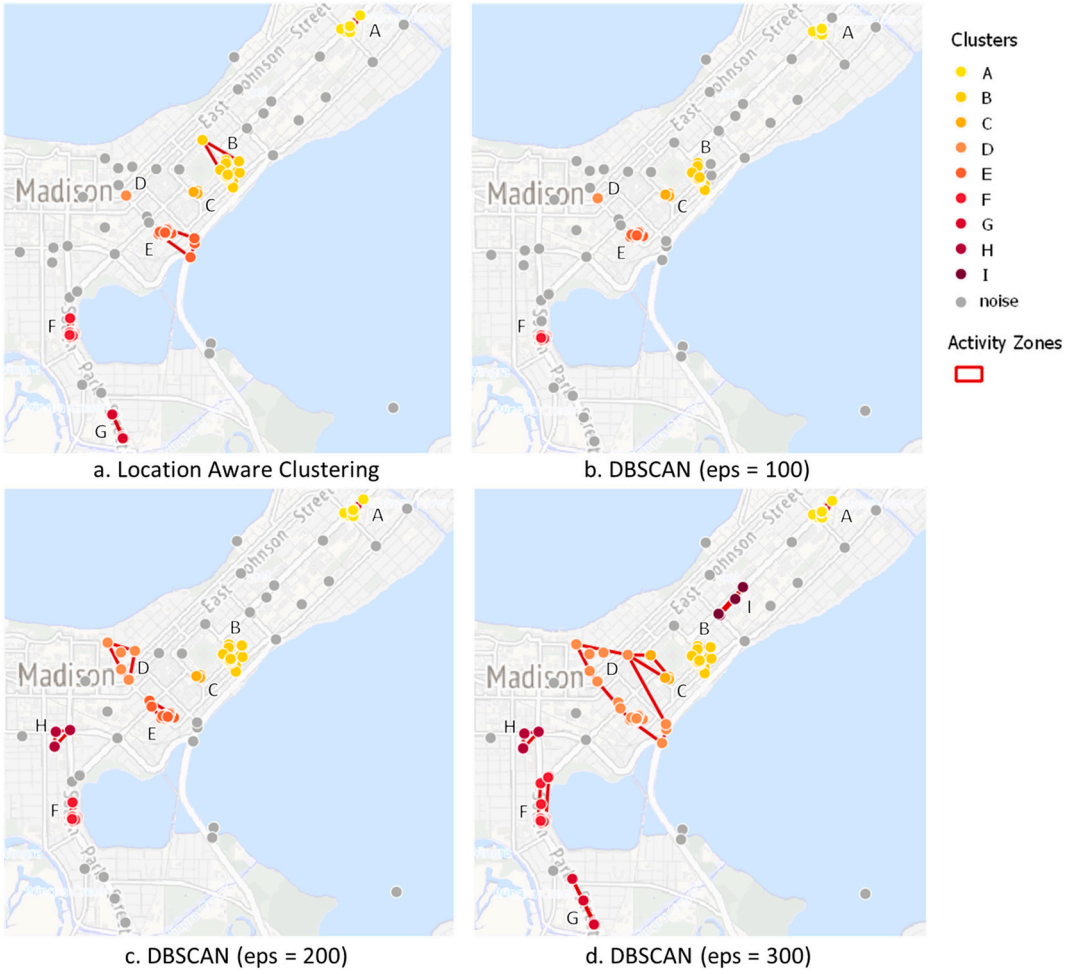


Fig. 8. Detection of individual activity zones using location-aware clustering and DBSCAN with different *eps* values for selected Twitter user 1.

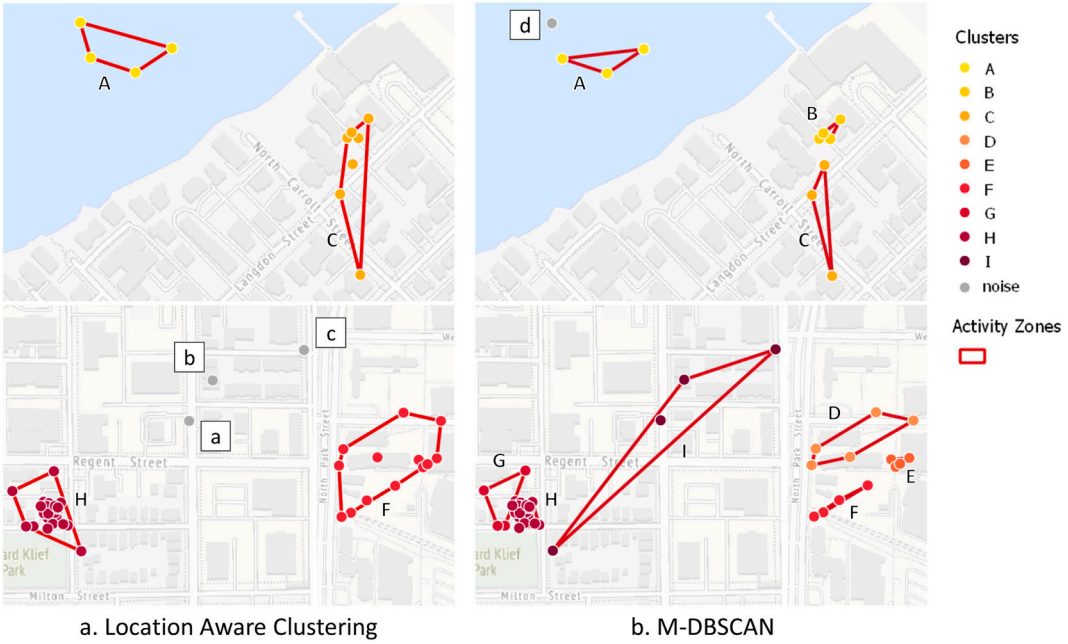


Fig. 9. Detection of activity zones using location-aware clustering and M-DBSCAN for selected Twitter user 2.

Fig. 9 shows the clustering results of another selected Twitter user using location-aware clustering (Fig. 9a) and M-DBSCAN (Fig. 9b). M-DBSCAN performs well in detecting clusters of different spatial densities (e.g., G, H, and I in Fig. 9b). It automatically generates *eps* values based on the spatial distribution of all the footprints, and then aggregates the footprints in various scales by applying different eps values. However, it does not consider the semantic homogeneity of these clusters, and thus breaks the integrated activity zones detected by location-aware clustering (e.g., cluster B, F, and H in Fig. 9a). Additionally, M-DBSCAN may misclassify semantic noise because noise (e.g., point a, b, c in Fig. 9a) can form a spatial cluster of small densities (e.g., cluster I in Fig. 9b) and non-noise (e.g., point d in Fig. 9n) can be located relatively far away (i.e., beyond the selected *eps*) from the main cluster (e.g., cluster A in Fig. 9b).

Based on the above analysis, location-aware clustering can detect clusters representing individual travel activities from digital footprints more effectively (Figs. 8a and 9a). To this end, individual daily travel patterns can be extracted with the activity identification results. For example, Fig. 10 demonstrates the detected daily frequent travel activities for the same individual as Fig. 9. We can observe that *Dwelling*, *Education*, *Entertainment*, *Shopping*, and *Eating* activities constitute this person's daily life. In fact, *Dwelling* and *Work/Education* activities are detected for most individuals. Besides, this selected person conducts *Entertainment* activities at multiple locations, and might travel a long distance westward for *Eating*. These activities can also be verified using the reference map (Fig. 10b).

5. Evaluation

Among 49 unique Twitter users in Madison, Wisconsin, 6 users with the largest number of distinct dates when one or more geotagged messages were posted, are manually investigated to validate the clustering and activity type inference results. These users are selected to ensure that each user generated diversified trajectory data on different dates over a long period for our analysis. Next, activity clusters are manually identified for each user based on the spatial distribution of individual digital footprints, the activity type information drawn from geographic context and message content (Fig. 11). Three primary labeling principles are followed: (1) frequent tweets (i.e., ≥ 4 points) located at a specific place or its neighboring places representing one type of activity (e.g., *Dwelling, Shopping*) are clustered (i.e., cluster A, B, C in Fig. 11); (2) tweets outside the cluster while representing the same or related activity (e.g., parking) are clustered; (3) infrequent tweets (i.e., < 4 points) at places described as (1) and (2) (i.e., noise points in Fig. 11) and spatially adjacent tweets spreading over multiple places of distinct activity types (points of cluster A and point a in Fig. 11) are not clustered. In addition, to differentiate clusters based on the activity types performed at an activity zone, activity clusters of small spatial scales are labeled separately if their surrounding clusters cover relatively large zones with more general or different activity types (e.g., zone A is labeled as *Dwelling* although located between two large *Entertainment* zones B and C in Fig. 11). This principle ensures that small-scale activities are not neglected and misclassified as the type of its surrounding large-scale activities. Finally, through examining the overlapped land use map and POIs of each cluster, its activity type is manually identified, and every ST point within it is assigned this type.

Next, we applied the proposed location-aware clustering method on each selected user's footprints and evaluated clustering results by examining how well different activities are detected and classified. Specifically, ST point grouping schemes are evaluated with Rand Index (RI) and Adjusted Rand Index (ARI). Activity type identification accuracy is evaluated with several quality measures in comparison with manually labeled results in Section 5.2.

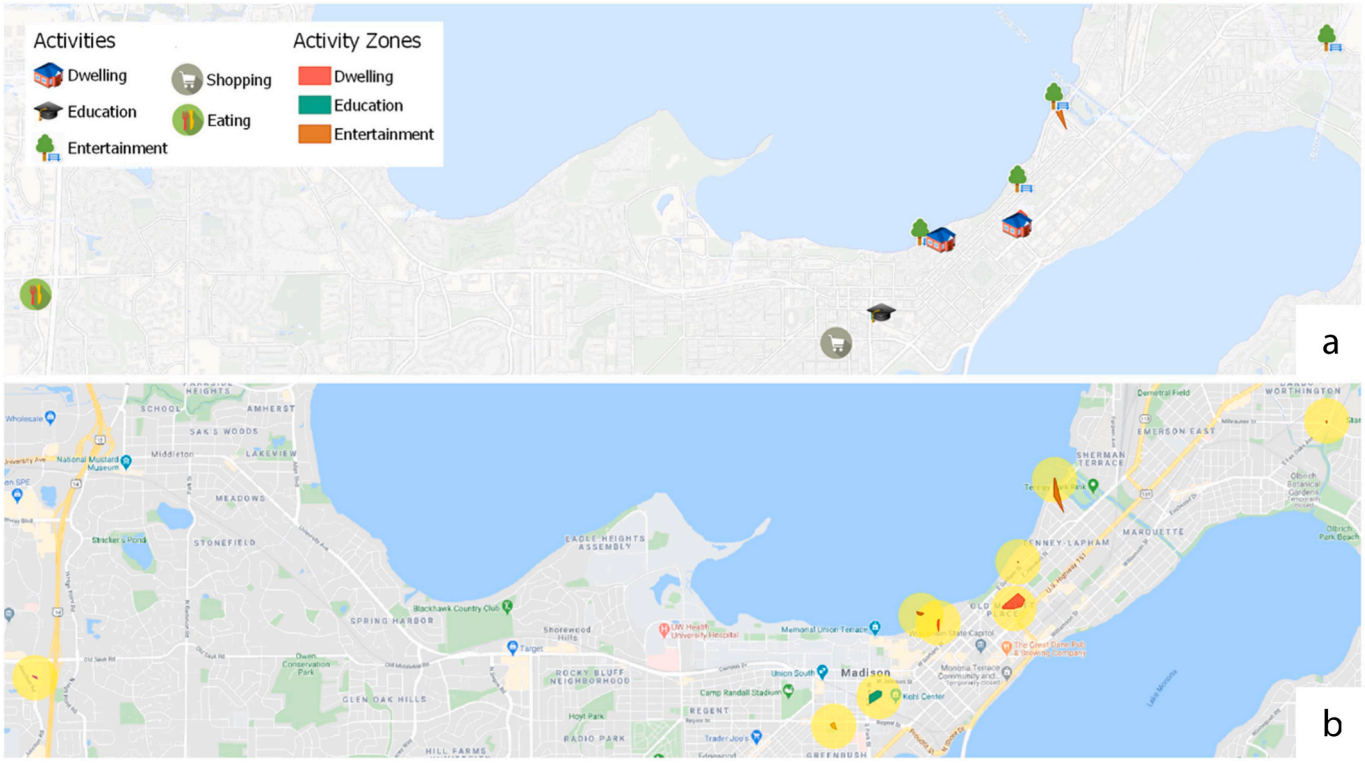


Fig. 10. Individual daily frequent activities detected with location-aware clustering for selected Twitter user 2 (a) and geographic context of these activities highlighted with circles in yellow by Google Maps (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

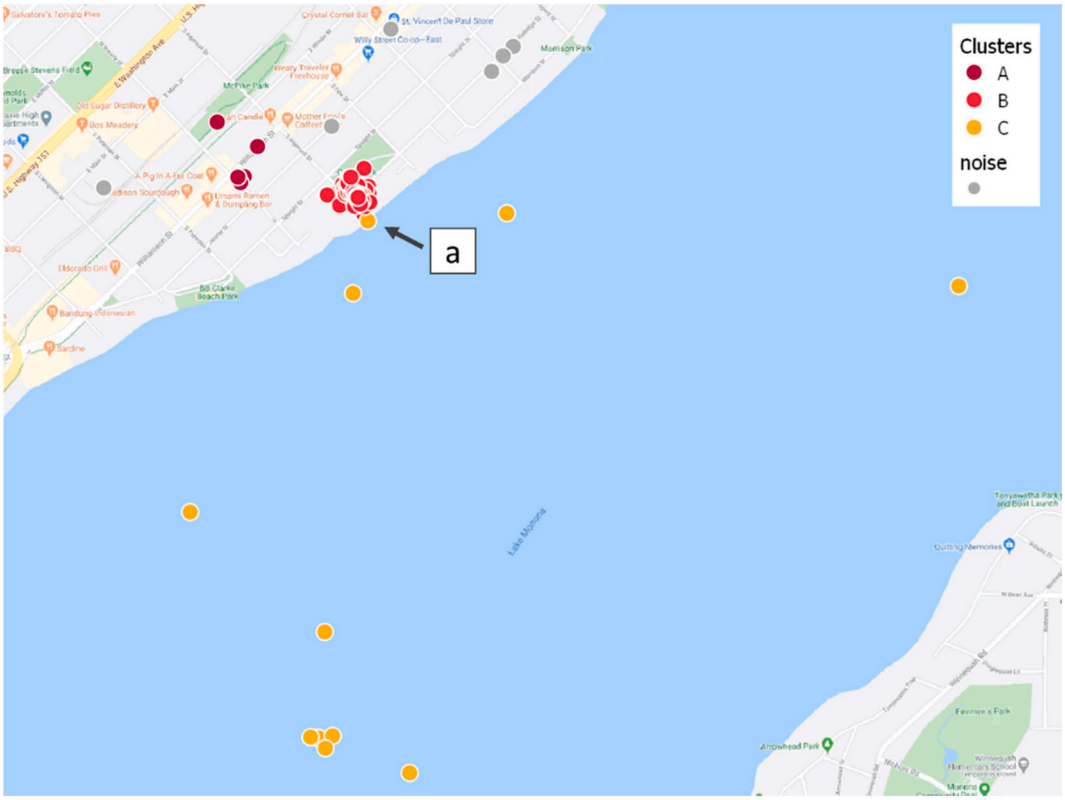


Fig. 11. Demonstration of manually detected activity clusters.

5.1. Clustering evaluation with rand index and adjusted Rand index

RI (Rand, 1971) and ARI (Hubert & Arabie, 1985) are used to evaluate the effectiveness of the proposed location-aware clustering method. Both RI and ARI can validate clustering results by measuring the agreement between object pairs in two partitions. Specifically, if a pair of objects are assigned to the same class or different classes using different partition methods, the assignment is counted as an agreement (A), or otherwise a disagreement (D). The similarity of two partitions can thus be estimated by measuring the overlap of A versus D (Walde, 2006). ARI can estimate the similarity between two partitions with different numbers of clusters (Santos & Embrechts, 2009; Yeung & Ruzzo, 2001). To measure the similarity between manually labeled outcomes (C) and the predicted results (M) produced by different clustering methods, a $RI(C,M)$ and an $ARI(C,M)$ are generated for each selected user. Table 4 shows the averaged values ($\overline{RI}(C,M)$ and $\overline{ARI}(C,M)$) for all six users using different clustering methods (Table 4).

As shown in the table, location-aware clustering achieves high $\overline{RI}(C,M)$ and $\overline{ARI}(C,M)$ values (over 0.98), which means it can aggregate digital footprints into activity clusters in agreement with manually labeled results. In contrast, M-DBSCAN receives lower values for both indexes, and DBSCAN gets lower values for $\overline{RI}(C,M)$, but higher values for $\overline{ARI}(C,M)$ with *eps* less than or equal to 100 m. This is because DBSCAN with small *eps* may collapse a cluster (i.e., M) into smaller clusters (i.e., C), decreasing the agreement measure for $\overline{RI}(C,M)$ while increasing the measure for $\overline{ARI}(C,M)$. As *eps* increases, both index values increase first and then drop for DBSCAN. When *eps* equal to 100 m, both indexes become nearly the same as the indexes for location-aware clustering. Despite the high indexes, DBSCAN with a single *eps* value cannot detect clusters of different spatial densities as illustrated in Section 4.2. Points in sparse clusters are misclassified as noise without evidently decreasing both indexes. Therefore, the performance of location-aware clustering shows an improvement while detecting distinct activities.

5.2. Activity zone type identification

We evaluate the performance of activity zone type identification at two levels: the point level and the activity zone (cluster) level. At the point level, activity zone type (i.e., one out of ten predefined categories) is manually identified and predicted for each trajectory point. Each point is thus considered as a classification instance and gets evaluated. At the cluster level, the predicted activity type of each manually identified cluster is defined as the predicted type of the majority points within the cluster. Therefore, each manually identified cluster (without noise) is considered as a classification instance.

At both levels, given a predefined activity zone type, a binary label (i.e., true or false) is assigned to each classification instance indicating whether it is classified as this type. For each instance, there are four measurement indexes: true positive (*TP*), false negative (*FN*), true

Table 4
Rand index and adjusted Rand index analysis using different clustering methods.

Cluster Method	Location-aware Clustering	M-DBSCAN	DBSCAN				
eps	N/A	N/A	20 m	50 m	100 m	200 m	300 m
$\overline{RI}(C,M)$	0.984	0.971	0.950	0.979	0.983	0.976	0.955
$\overline{ARI}(C,M)$	0.984	0.978	0.990	0.992	0.985	0.948	0.896

Table 5
Activity zone detection using manual labeling and location-aware clustering (L: label; P: prediction).

	Point				Activity Cluster													
	Clustered		Noise		Dwelling		Education		Work		Eating		Entertainment		Shopping		Transportation	
	L	P	L	P	L	P	L	P	L	P	L	P	L	P	L	P	L	P
U1	406	403	48	51	1	2	1	1	0	0	0	1	3	1	0	0	0	0
U2	324	335	79	68	2	3	0	0	1	2	2	1	4	4	3	3	2	0
U3	389	417	91	63	2	8	1	1	0	1	2	2	7	3	0	0	0	0
U4	605	605	14	14	1	3	0	0	0	0	0	0	2	0	1	0	0	0
U5	219	216	46	49	1	2	2	2	0	1	0	0	2	1	1	0	0	0
U6	343	335	13	21	1	2	0	0	1	0	0	0	1	1	1	0	0	0

Table 6
Performance analysis at point/cluster level for activity type identification using location-aware clustering.

	Overall accuracy		Overall sensitivity		Overall F1 score	
	Point	Cluster	Point	Cluster	Point	Cluster
Dwelling	0.932	0.746	1.000	1.000	0.955	0.633
Education	0.973	0.944	0.860	0.833	0.912	0.889
Work	0.884	0.839	0.500	0.500	0.865	0.667
Eating	0.994	0.964	0.939	0.750	0.968	0.833
Entertainment	0.956	0.729	0.511	0.478	0.714	0.672
Shopping	0.932	0.790	0.000	0.000	N/A	N/A
Transportation	0.987	0.929	0.000	0.000	N/A	N/A

Table 7
Number of detected collective activities of different types using variant combinations of *eps* values for multi-level spatial clustering (please refer to Table 3 in appendix for the counts of a complete set of combinations).

Activity Zone Type	*eps* (m)					*eps* (m) combination		
	20	50	100	200	300	20, 50, 100	50, 100, 200	50, 200, 300
Eating	70	40	36	29	33	97	41	48
Work	13	26	22	10	6	27	26	24
Shopping	56	28	24	30	28	75	35	37
Education	59	46	23	20	24	79	50	49
Health	1	3	2	2	2	3	3	3
Entertainment	41	83	82	52	43	75	85	78
Service	2	2	1	2	1	3	3	4
Dwelling	37	56	55	29	25	47	59	65
Transportation	1	1	1	0	0	2	0	0
Transportation network	20	22	18	13	9	33	28	24
Others	22	8	8	22	13	8	15	12
Sum	322	315	272	209	183	449	345	344

negative (*TN*), and false positive (*FP*) (Fawcett, 2006). *TP* indicates the number of instances our method successfully classifies, whereas *TN* counts the instances that are misclassified and do not belong to a certain type. Three quality measures for each classifier are then defined below,

- $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$: the ratio of the amount of correctly classified instances to the amount of all classification instances. It measures the general accuracy. A higher value means more instances are correctly classified.
- $sensitivity\ (recall) = \frac{TP}{TP+FN}$: the ratio of the number of instances which are correctly identified as a specific type to the number of instances labeled as this type. A higher value means a better capability to detect such activity type.
- $F1\ score = \frac{2*precision*recall}{precision+recall}$: a combined measure of precision and recall, where $precision = \frac{TP}{TP+FP}$. It measures a balanced performance, with a higher value indicating the more effectiveness of activity detection.

Table 5 lists the activity zones of each activity zone type detected for every test user and Table 6 shows the averaged quality measure values for all test users. Generally, the detected activities of *Dwelling* or *Work* are more than the manually labeled ones. Meanwhile, less *Entertainment* or *Shopping* activities are detected. The predicted *Education* and *Eating* activities are almost the same as the manually labeled ones. Accordingly, *Education* and *Eating* activities receive relatively high F1 scores at the cluster level. Three activity zone types, including *Health*, *Service*, and *Transportation network*, are neither detected by the algorithm nor manually identified, which are thus not included in Table 5 and Table 6.

Since *TN* index is always high (i.e., most points are classified as not belonging to the activity zone type), all activity type predications receive high accuracy at the point level. Most predictions also receive high accuracy at the cluster level except for *Dwelling* and *Entertainment* activities. This is because many *Entertainment* activities are misclassified as *Dwelling* ones as they are located at *Dwelling* areas, which might be a friend's home instead indicated by Twitter message content. Other misclassifications are also related to the mismatch of people's real activities and geographic-context-indicated activities. For example, an *Entertainment* activity at a gym located at a shopping mall is misclassified as a *Shopping* activity, and an *Eating* activity near a workspace is misclassified as a *Work* activity.

All manually labeled *Dwelling* activities can be successfully detected by the proposed location-aware clustering method. Therefore, the overall sensitivity for *Dwelling* becomes 1.0. Most *Education* and *Eating* activities are also successfully detected with a high sensitivity measure. Furthermore, nearly half of *Entertainment* and *Work* activities can be correctly identified. *Shopping* activities are rarely detected as they are mostly located at the "Commercial" area, which is not classified as any specific activity type in this study.

5.3. Discussion

This section discusses potential limitations in our work. First, the *eps* at different levels for multi-level spatial clustering are selected statically and by referencing the findings from previous studies (Hu et al., 2015; Liu, Huang, & Gao, 2019; Hwang et al., 2018). Based on the principles of DBSCAN, the smaller the *eps*, the more spatial clusters of high densities will be produced. In contrast, a larger *eps* will detect more spatial clusters of low densities. To better understand the impact of their various combinations, we generate collective activity zones using three selective *eps* out of an extensive set of values (i.e., 20 m, 50 m, 100 m, 200 m, 300 m) and summarize the number of detected activity zones for each activity type (Table 7). The outcome shows that when reducing the *eps* of the first level to 20 m, more activity zones, especially those with compact space (e.g., *Eating*, *Shopping*, and *Education* zones), are detected. Meanwhile, increasing the *eps* at the third level enables the detection of activity zones consisting of very sparse spatial points, especially for *Dwelling* and some other *Eating* activities. However, 20 m as *eps* could break activity zones detected with larger values (e.g., 50 m) and result in many individual frequent activities undetected, because each broken zone mostly contains a limited number of footprints of an individual. Meanwhile, 300 m is too large to distinguish spatially adjacent activities

of the same type. Therefore, this study selected a combination of 50 m, 100 m, and 200 m. However, the most appropriate parameters are variant depending on different activity types and should be further investigated in future research.

Moreover, a uniform buffer zone radius of 200 m is used to search for individual footprints for individual activity identification, while varying radius might be used for different activity types. Furthermore, the priority of activity zone types in multi-level spatial clustering is determined via extensive experiments to boost the overall activity identification accuracy, which is explained by analyzing the scales of diverse activity zones. However, the priority order can be changed to give dominance to different activity types. A universal activity identification scheme will be explored by establishing machine learning models in future study. Finally, the proposed multi-level spatial clustering method is not compared with other spatial clustering methods (e.g., graph-based and grid-based approaches) through evaluation experiments. While other spatial clustering methods can also be modified to infer travel semantics by integrating geographic context data, our study focuses on developing the two-step framework which enables semantic inference and can detect individual frequent travel activities from sparse digital footprints.

6. Conclusion and future research

With the explosive growth of social media data, existing spatiotemporal and semantic aggregation methods do not provide sufficient support for exploring collective or individual travel activity patterns using sparse digital footprints. This study focuses on developing a two-step AKD framework for efficient detection of travel activities at both collective and individual levels. First, collective activity zones (i.e., spatial clusters) are generated from digital footprints of all users collected at the target area, using the multi-level spatial clustering method by considering the underlying activities (i.e., geographic context) of the clusters, which are identified by integrating OSM datasets. Second, individual digital footprints overlapped with the same collective activity zone are clustered and represent one distinct activity at the individual level with location-aware clustering. The experiment results show that by integrating activity information drawn from OSM datasets, the proposed framework can better aggregate digital footprints into clusters at both collective and individual levels and the individual clusters are identified as different activities in high accuracy.

Future efforts could be devoted to improving the performance of activity type identification. Specifically, parameters in multi-level clustering, such as *eps* values and buffer zone size, should be determined based on different activity zone types. Additionally, more features associated with the tweet points can be utilized for prediction, such as message content and temporal information. Further, computation models with more robust performance, such as machine learning models, should be developed to fully leverage all data resources.

Declarations of Competing Interest

None.

Funding

This work was supported by the University of Wisconsin – Madison office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2020.101551.

References

Ahlers, D. (2013). Assessment of the accuracy of GeoNames gazetteer data. In *Proceedings of the 7th workshop on geographic information retrieval* (pp. 74–81). Orlando, Florida: ACM.

Alvares, L. O., et al. (2007). *Towards semantic trajectory knowledge discovery*. Limbourg, Belgium: Hasselt University.

Ankerst, M., et al. (1999). OPTICS: ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD international conference on management of data* (pp. 49–60). Philadelphia, Pennsylvania, USA: ACM.

Aurelio Beber, M., et al. (2016). Towards activity recognition in moving object trajectories from twitter data. In *Proceeding XVII GEOINFO, November 27–30, 2016*. Campos do Jordao, Brazil.

Bäcklund, H., Hedblom, A., & Neijman, N. (2011). *DBSCAN: A density-based spatial clustering of application with noise*. Linköpings Universitet – ITN.

Batty, M., et al. (2012). Smart cities of the future. *The European Physical Journal Special Topics, 214*, 481–518.

Blanford, J. I., et al. (2015). Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PloS One*, 10(6).

Cai, G., Lee, K., & Lee, I. (2016). Mining semantic sequential patterns from geo-tagged photos. In *2016 49th Hawaii International Conference on System Sciences (HICSS). Australia* (pp. 2187–2196).

Campello, R. J. G. B., et al. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *J ACM Trans. Knowl. Discov. Data, 10*(1), 1–51.

Cao, G., et al. (2014). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems, 51*. https://doi.org/10.1016/j.compenvurbsys.2015.01.002.

Chaniotakis, E., et al. (2017). Inferring activities from social media data. *Transportation Research Record, 2666*(1), 29–37.

Chen, J., et al. (2011). Exploratory data analysis of activity diary data: A space–time GIS approach. *Journal of Transport Geography, 19*(3), 394–404.

Cheng, Z., et al. (2011). *Exploring millions of footprints in location sharing services*. ed. ICWSM.

Comito, C., Falcone, D., & Talia, D. (2016). Mining human mobility patterns from social geo-tagged data. *Pervasive and Mobile Computing, 33*, 91–107.

Du, Q., et al. (2016). Density-based clustering with geographical background constraints using a semantic expression model. *ISPRS International Journal of Geo-Information, 5*(5), 72.

Ester, M., et al. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining* (pp. 226–231). Portland, Oregon: AAAI Press.

Estima, J., & Painho, M. (2013). Exploratory analysis of OpenStreetMap for land use classification. In *Proceedings of the Second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information* (pp. 39–46). Orlando, Florida: ACM.

Estivill-Castro, V., & Lee, I. (2000). Autoclust: Automatic clustering via boundary extraction for mining massive point-data sets. In *Proceedings of the 5th international conference on geocomputation*.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Fonte, C. C., et al. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science, 29*(7), 1269–1291.

Gao, H., & Liu, H. (2014). Data analysis on location-based social networks. In A. Chin, & D. Zhang (Eds.), *Mobile social networking*. New York, NY: Computational Social Sciences. Springer.

Gao, Y., et al. (2018). A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science, 32*(7), 1304–1325.

Gong, L., et al. (2015). Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. *Journal of Modern Transportation, 23*(3), 202–213.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779.

Hasan, S., & Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies, 44*, 363–381.

Hio, C., et al. (2013). A hybrid grid-based method for mining arbitrary regions-of-interest from trajectories. In *Proceedings of workshop on machine learning for sensory data analysis* (pp. 5–12). Dunedin, New Zealand: ACM.

Hu, Y., et al. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems, 54*, 240–254.

Huang, Q., Cao, G., & Wang, C. (2014). From where do tweets originate?: a GIS approach for user location inference. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. Dallas/Fort Worth, Texas: ACM*, 1–8.

Huang, Q., Li, Z., et al. (2016). Mining frequent trajectory patterns from online footprints. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoStreaming. Burlingame, California: ACM*, 1–7.

Huang, Q., & Wong, D. W. S. (2015). Modeling and visualizing regular human mobility patterns with uncertainty: An example using twitter data. *Annals of the Association of American Geographers, 105*(6), 1179–1197.

Huang, Q., & Wong, D. W. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science, 30*(9), 1873–1898.

Huang, W., & Li, S. (2018). An approach for understanding urban human activity patterns with the motivations behind. *International Journal of Geographical Information Science.*. https://doi.org/10.1080/13658816.2018.1530354.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification, 2*(1), 193–218.
Hwang, S., Hanke, T., & Evans, C. (2018). Frontiers in information systems. In A. C. Teodoro (Ed.), *Frontiers in information systems* (pp. 184–207).
Jiang, S., et al. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems, 53*, 36–46.
Kang, C., et al. (2010). Analyzing and geo-visualizing individual human mobility patterns using mobile call records. ed. In *2010 18th international conference on geoinformatics* (pp. 1–7), 18–20 June.
Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer, 32*(8), 68–75.
Linton, S. L., et al. (2014). Application of space-time scan statistics to describe geographic and temporal clustering of visible drug activity. *Journal of Urban Health: Bulletin of the New York Academy of Medicine, 91*(5), 940–956.
Liu, P., Zhou, D., & Wu, N. (2007). VDBSCAN: varied density based spatial clustering of applications with noise. ed. In *2007 international conference on service systems and service management* (pp. 1–4), 9–11 June.
Liu, X., Huang, Q., & Gao, S. (2019). Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN. *International Journal of Geographical Information Science, 33*(6), 1196–1223.
Lu, C.-T., et al. (2011). A framework of mining semantic regions from trajectories. In *International Conference on Database Systems for Advanced Applications* (pp. 193–207). Berlin, Heidelberg: Springer.
Maciąg, P. S. (2017). A survey on data mining methods for clustering complex spatiotemporal data. In *International conference: beyond databases, architectures and structures* (pp. 115–126). Cham: Springer.
Marques, F. J. N. (2014). *A constraint-based clustering algorithm for detection of meaningful places* (Master's thesis).
Njoo, G., et al. (2015). A fusion-based approach for user activities recognition on smart phones. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, 2015* (pp. 1–10).
Preoţiuc-Pietro, D., & Cohn, T. (2013). A temporal model of text periodicities using Gaussian Processes. ed. In *Proceedings of the 2013 conference on empirical methods in natural language processing, 2013 Seattle, Washington, USA* (pp. 977–988).
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association, 66*(336), 846–850.
Santos, J. M., & Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. ed.. *International Conference on Artificial Neural Networks, September 14-17 2009 Limassol, Cyprus* (pp. 175–184).
Song, C., et al. (2010). Limits of predictability in human mobility. *Science, 327*(5968), 1018.
Steiger, E., Resch, B., & Zipf, A. (2015). *Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks*.
Tu, W., et al. (2017). Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science, 31*(12), 2331–2358.
Velmurugan, T., & Santhanam, T. (2011). A survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal, 10*, 478–484.
Walde, S. S.i. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational linguistics, 32*(2), 159–194.
Wu, L., et al. (2014). Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PLoS One, 9*(5), Article e97010.
Xu, Y., et al. (2018). *Human mobility and socioeconomic Status: Analysis of Singapore and Boston. Computers environment and urban systems.* https://doi.org/10.1016/j.compenvurbsys.2018.04.001.
Yan, Z., et al. (2013). Semantic trajectories: Mobility data computation and annotation. [online]. Retrieved from http://infoscience.epfl.ch/record/177359/files/ACM-TIST.pdf [Accessed 2012].
Yang, F., et al. (2014). Dynamic origin-destination travel demand estimation using location based social networking data. In *TRB 2014 annual meeting*.
Yeung, K. Y., & Ruzzo, W. L. (2001). Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics, 17*(9), 763–774.
Yin, D., & Wang, S. (2016). Sensing spatial structures through large-scale social media. In *The Association of American Geographers 112nd annual meeting*. San Francisco, California, USA.
Yin, Z., et al. (2011). Diversified trajectory pattern ranking in geo-tagged social media. In *Proceedings of the 11th SIAM international conference on data mining* (pp. 980–991). https://doi.org/10.1137/1.9781611972818.84. SDM 2011.
Yuan, Y., Liu, Y., & Wei, G. (2017). Exploring inter-country connection in mass media: A case study of China. *Computers, Environment and Urban Systems, 62*, 86–96.
Zhang, D., et al. (2014). Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking* (pp. 201–212). Maui, Hawaii, USA: ACM.
Zhao, K., et al. (2016). *Urban human mobility data mining: An overview*. ed. 2016 IEEE International Conference on Big Data (Big Data), 5–8 Dec. 2016 (pp. 1911–1920).
Zhao, P., et al. (2017). A trajectory clustering approach based on decision graph and data field for detecting hotspots. *International Journal of Geographical Information Science, 31*(6), 1101–1127.
Zheng, Y., Zha, Z., & Chua, T. (2012). Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology, 56*, 1–18. https://doi.org/10.1145/2168752.2168770, 3 3.
Zhou, X., & Zhang, L. (2016). Crowdsourcing functions of the living city from twitter and foursquare data. *Cartography and Geographic Information Science, 43*(5), 393–404.
Zielstra, D. and Zipf, A., A comparative study of proprietary geodata and volunteered geographic information for Germany. 13th AGILE international conference on geographic information science, vol. 2010.