Detection and Exploration of Individual Semantic Trajectories Using Social Media Data

Xinyi Liu

Advisor: Qunying Huang

A thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science

(Cartography and Geographic Information Science)

at the

University of Wisconsin-Madison

2018

Acknowledgements

I would like to express my sincere thanks to my advisor, Dr. Qunying Huang. She not only guided me to explore research problems with extreme patience and encouraged me to overcome all difficulties during my study period, but also well supported me in life when I began to live in an unfamiliar country. I was very lucky to work with and study from her. This thesis would never have been completed without her profound expertise, careful guidance and persistent encouragement.

My sincere appreciation also goes to my committee members Dr. Robert Roth and Dr. Song Gao, who inspired me with their creative ideas and provided me with valuable advice and feedback. Besides, I would like to thank Dr. A-Xing Zhu for his inculcation on general academic study. It was my great honor to learn from all of them during the past two years.

I am also grateful to my former and present colleagues, Chris Scheele, Guiming Zhang, Meiliu Wu, Bo Peng, Atlas Guo and Yuhao Kang. They all offered me helpful feedback on this thesis work throughout the two years.

Last but not the least, I would like to thank my dear parents and other lovely friends. I would never achieve anything without their support in my life. I will try to do better in the future with their support and encouragement.

Table of Contents

Abstract

Individual travel trajectories collected from social media platforms (i.e., digital footprints) are often aggregated using methods such as the density-based spatial clustering of applications with noise (DBSCAN) and varying DBSCAN (VDBSCAN) for travel activity (e.g., eating, working, entertainment) identification. However, spatial clusters cannot represent distinct individual representative travel activities. This thesis work first develops a multi-scale spatial clustering method to aggregate digital footprints of a group of users into collective spatial hot-spots (i.e., activity zones), and to identify the activity type (e.g., dwelling, service, transportation and office) of each collective zone by integrating Volunteered Geographic Information (VGI) data, specifically OpenStreetMap (OSM) datasets. Each digital footprint of an individual, represented as a spatiotemporal (ST) point, is then attached with a collective activity zone that either includes or overlaps a buffer zone of the ST point, which is generated by using the point as the centroid and a predefined threshold as the radius. Given an individual's ST points with semantics (i.e., activity type information) derived from the attached collective activity zone, a semantic activity clustering method is then developed to detect daily representative activity clusters of the individual. Next, temporal information of a daily activity cluster, indicating the time period when the individual frequently visits the zone covered by the activity cluster, is detected, and individual representative daily semantic travel trajectory paths (i.e., semantic travel trajectory, defined as chronological travel activity sequences) are constructed between every two subsequent activity clusters. Finally, a geovisual analytical web portal is developed to display individual representative daily travel trajectories and associated activity zone information for better exploring individual and collective semantic travel patterns.

Experiments with the historic geo-tagged tweets collected within Madison, Wisconsin for 49 eligible users reveal that: 1) The proposed multi-scale spatial clustering method can detect

most significant activity zones with accurate zone types identified; 2) The semantic activity clustering method based on the derived activity zones can aggregate individual travel trajectories into activity clusters more efficiently comparing to both DBSCAN and VDBSCAN; 3) Individual semantic travel patterns can be explored and compared through geovisual analytics, and collective semantic travel patterns thus can be unfolded for a group of people with similar individual travel patterns.

Chapter 1. Introduction

Examining people's movement patterns can benefit infrastructure construction and policy making related to transportation, commercial land use planning, disease control, disaster management and so on (González et al. 2008, Song et al. 2010, Cheng et al. 2011, Zhao et al. 2016). It is also paramount for studying a variety of social problems, such as social segregation (Huang and Wong 2016). To study these topics, individual travel data have been traditionally collected by surveys, which is tedious and expensive. Nowadays, the advancement of smart devices significantly increases the opportunities to collect travel data by passively capturing a huge amount of tracking records from cell phones, GPS devices, and social media. Among them, social media data are increasingly used for examining human mobility due to several unique advantages (Huang and Wong 2015), including: 1) public availability, 2) capturing long-term trajectories, 3) a large number of participants, and 4) nearly real-time data available.

Typically, digital footprints collected through social media platforms are recorded as a set of *spatiotemporal* (ST) points that describe individual trajectories with time stamps. Those ST points capture useful information about individual travel patterns. However, the travel patterns are not immediately visible and identifiable from the massive ST points without processing and mining them (Cheng et al. 2011). This is because individual daily movement has been proved to show specific regularity in both temporal and spatial dimensions, which is governed by various factors (e.g., lifestyle, accessibility of transport modes, transportation policy and so on; Asgari et al. 2013). As such, those massive ST points not only capture the regular activities but also random activities (González et al. 2008, Song et al. 2010). In order to identify regular movements, aggregation of individual trajectories in both space and time dimensions become necessary (Giannotti et al. 2011, Andrienko and Andrienko 2012, Zhang et al. 2014), by grouping multiple scattered ST points with proximity in space and time into clusters (Andrienko and Andrienko

2012). Mining groups of individual trajectories then can identify the flows of people's collective movements. To detect those flows, different aggregation models and methods are developed to understand individual patterns (Shaw and Yu 2009, Kang and Yong 2010, Huang and Wong 2015) and their collective characteristics (Wu et al. 2014, Zhang et al. 2014).

The basic aggregation method is to generate statistics such as hourly regularity and radius of gyration via regression (Song et al. 2010). However, statistical methods cannot support mobility analysis in an intuitive, dynamic, and interactive way. Geovisual analytics thus becomes indispensable in unveiling the complexity of millions of individual footprints and in extracting individual and collective movement patterns (Andrienko and Andrienko 2012). Geovisual analytics integrate geographical reasoning processes with interactive visual interfaces (Kraak 2014) and provides a dynamic way to understand and test existing hypothesis and to discover new questions (Roth et al. 2015). For example, Chen et al. (2011) developed a GIS framework that can visualize individual daily activity trajectories through color-symbolized space-time paths to support analysis of activity patterns aggregated from travel diaries. However, trajectory data collected from social media platforms are unstructured, and contain uncertainty in both space and time by capturing both regular and random movements (Huang and Wong 2015). Existing GIS system cannot process and explore such data to mine meaningful travel patterns. Also, existing exploratory systems only provide limited analysis functions by using survey data with certain time stamps (Kwan 2000).

To date, much progress has been made to leverage social media data in mobility study via geovisual analytical methods. For example, Yin and Wang (2016) generated 2D movement flows to show traffic state around Chicago city. However, this study can only generate movement hot-spots without investigating specific individual and collective travel patterns. Huang and Wong (2016) explored individual activity patterns based on inference of home area and workplace of

social media users to reveal the mobility patterns of different *socioeconomic status* (SES) groups. However, their methods cannot infer other important activity locations (e.g., entertainment, education, transportation), producing a rather general analysis of human mobility patterns. Although individual frequent activity zones were derived, it is still unclear what kind of activities people (e.g., eating or shopping) are conducting in a specific ST range (Huang et al. 2016). While several additional activities (e.g., entertainment and education) in an *area of interest* (AOI) could be inferred through incorporating land use data and Google Place Services (Huang et al. 2014), individual representative daily travel trajectory with detailed and accurate activity (e.g., shopping, grocery, eating, going to church etc.) information was not yet produced. Besides, the land use data are published by a local department and need to be collected and processed on a case-by-case basis when studying the mobility patterns of social media users in a different region.

1.1 Research problem

To address the limitations of existing studies, I ask the following research question: How can we detect individual semantic travel patterns from their sparse trajectories collected from social media platforms?

Current studies design multiple schemes to detect individual representative semantic travel trajectories out from traditional survey data or digital records of high density such as cell phone logs. *Origin and destination* (OD) pairs of one trip can be pinpointed from those data with a relatively high resolution, which is the foundation to predict activity information. Different datasets (e.g., activity diary, GeoNames) are introduced to provide geographical contexts of the OD pairs for activity identification (Pinjari and Bhat 2011, Cai et al. 2016).

However, individual frequently-visited locations cannot be detected from digital footprints by the methods applicable for traditional datasets, such as surveys, cell phone records, and GPS trajectories, because of the uncertainties of such footprints in both space and time. First,

traditional surveys provide individuals' exact location and activity information, while cell phone logs and GPS trajectories record individuals' ST footprints in regular intervals, both of which are relatively dense and reliable data sources. In comparison, geo-tagged social media messages could be randomly published at various places depending on users' personal preference and behaviors on using social media, which makes it hard to define and identify users' regular activities from these ST points. While some aggregation algorithms are applied in spatial or ST clustering such as the Spatial/Space-Time Scan Statistics method for aggregating GPS records (Gao et al. 2018, Linton et al. 2014), different methods are required for clustering irregularly and sparsely distributed digital footprints collected from social media platforms. Accordingly, appropriate representation of clusters generated from digital footprints should be further discussed.

These uncertainties make it challenging to detect activity information from clustering results and to identify the sequence of possible activities. This thesis addresses both of these problems. In our work, activity information of individual travel trajectories is denoted as trajectory semantics, which is attached to individuals' daily trajectory points to delineate their everyday activities (Yan et al. 2013). Semantic information can help researchers understand people's travel patterns and travel demands (e.g., the amount of users traveling between any OD pair; Gundlegård et al. 2016). It is identified based on the detection of travel trajectory clusters using current activity zone identification methods (Huang et al. 2016). However, uncertainties of travel trajectory detection can actually be mitigated by integrating semantic information. Therefore, to minimize the uncertainties, semantic information of individual footprints should be accurately identified as much as possible before trajectory aggregation.

It is worth nothing that individual trajectories cannot be aggregated to represent collective travel activity patterns. Previous studies merge trajectories based on their common

ODs (Zheng 2015, Zhi et al. 2014), which denote geometric similarity of the trajectories. For trajectories beyond the common ODs, although similarity scores are calculated via various methods (Gao et al. 2018, Wan et al. 2017, Gong et al. 2016), they are not further applied to group together similar trajectories. While it is reasonable to cluster trajectories with geometric similarities for detection of major events or so on, clustering semantic trajectories based on spatial dimension is meaningless. Semantic similarities are typically analyzed using graphs and statistics, which generate useful dataset while cannot be related to real applications for better exploration. For example, two groups of users living in different communities could go through exact same daily travel trajectories (home –> work; work –> entertainment; entertainment –> home). Such trajectory is typical on campus area where students with similar schedules live far away from each other. Projecting their trajectories on a map and comparing their semantic sequences could hint different community structures. In a word, semantic similarity denotes similar travel patterns, which need to be demonstrated for better understanding citywide travel streams.

The nature of aggregating semantic travel trajectories is to show similarities of their OD pairs since semantic information is only attached to ODs in our scenario. Geovisualization can thus help detect semantic similarities by displaying travel trajectories of different individuals with each of their activity zones identified as one of the predefined activity types. The visualization results could reveal individual travel patterns and suggest future exploration directions. However, current geovisualization schemes either only display geometric aggregation results of collective travel patterns or show separate semantic detection results for different individual trajectories. More advanced geovisualization framework needs to be implemented to correlate individual travel trajectories to study collective travel patterns.

1.2 Research approach

Using geo-tagged tweets as an example, this research proposes a framework for mining social media data, detecting individual semantic travel trajectories, and constructing and visualizing individual representative daily travel trajectory paths.

Within this framework, preprocessing is conducted to clean the tweet sets (i.e., remove noisy data), collected within Madison, Wisconsin (USA). Then collective spatial clusters, defined as activity zones in this paper, are derived through spatial aggregation based on ST points of all users in Madison. Combined with OpenStreetMap (OSM) datasets, scattered spatial clusters are further aggregated into more integrated activity zones. Next, activity type (e.g., dwelling, working or entertainment) of each zone is detected and thus semantics of activity zones can be identified. Then individual travel trajectories are aggregated based on the activity zone distribution derived above to generate a set of representative semantic clusters. This process is thus defined as *semantic activity clustering*. Next, temporal information of each semantic activity zone, indicating the time period the individual frequently visits the activity cluster, is detected. Finally, individual daily semantic travel paths are constructed between every two subsequent activity zones to show the OD pairs and the total daily trips between the OD during the experiment time period.

Finally, a space-time geovisual analytical web portal is developed to display individuals' travel paths and their semantic information. Information such as collective activity zones, individual semantic clusters, and individual representative daily travel trajectory paths, is depicted together in an intuitive manner to explore individual and collective travel patterns. Dynamic geovisual analytical functions are implemented so that individual daily activity patterns could be explored at multiple scales. The effectiveness of aggregation results produced by different clustering methods can also be easily compared. Through this portal, users could

explore individual and collective travel patterns according to their own research interest by customizing clustering parameters and delineating study areas of interest.

1.3 Contribution

Three major contributions are made by this thesis work: 1) a Multi-scale spatial clustering method is proposed to aggregate individual social media travel trajectories into activity zones (hot-spots), with semantic information provided by Volunteered Geographic Information (VGI) data, specifically OSM data. Using the DBSCAN method, eps values (i.e., radius for searching neighbors of each point to form a cluster) at different scales are applied to cluster sparse data points. The eps value at lowest scale (i.e., the smallest value) assures that different zone types can be separated. The next scale eps helps merge collapsed clusters into integrated ones. Additional process is added to keep disparate activity zones (zones of different activities) from merging so that diverse activity information will not be generalized. 2) VGI data are used to detect activity zone types. This study classifies activity zones into ten types, which are further divided into four priorities, including priority I: eating, shopping; priority II: education, work, health; priority III: entertainment, service, dwelling; priority IV: transportation, transportation network. Activities of the lower-level priorities tend to happen at a smaller scale and should be detected earlier; 3) a semantic activity clustering method and the corresponding geovisual analytical functions are developed to aggregate individual representative daily travel trajectories with semantic information. This method achieves better aggregation efficiency and helps explore similarities of individual travel activity patterns.

1.4 Thesis structure

The rest of this thesis is divided into four parts: Chapter 2 introduces related work on recent human mobility study and an important analysis framework – time geography, progresses on digital footprints as new data source for mobility studies and existing methods for activity zone

detection, individual daily activity pattern exploration and geovisual analytics of human mobility. Chapter 3 introduces the workflow of our improved analysis methodology. Each step method is illustrated in a subsection. Chapter 4 demonstrates experiment results of applying our methodology on Twitter data collected at Madison, Wisconsin. Chapter 5 makes conclusions and discusses potential improvement of current methods and vision of future research.

Chapter 2. Related Work

2.1 Time geography and the study of human mobility

The theory of *time geography* was primarily proposed by Hägerstrand (Pred 1977), which introduced time dimension, along with space dimension, to examine human activities (Pred 1977, Corbett 2001, Miller 2005, Shaw and Yu 2009). In time geography, individual activities can be depicted through space-time paths with each representing a possible activity series taken by a specific person (Corbett 2001). While representing individual behaviors in a relatively accurate manner, this framework was not practical due to the lack of visualization techniques to represent the space-time paths two decades ago (Miller 2005). Recently, modern surface monitors (e.g., environment and social network sensors) have become ubiquitous and provide a lot of individual travel data with ST information (Miller 2005), which empowers this framework to tackle many previously understudied problems such as human mobility (Chen et al. 2011).

In human mobility study, time geography has been widely used to investigate individual and collective movement patterns. Specifically, individual trajectories and representative travel paths are depicted and studied in both space and time dimensions under this framework. For example, Chen et al. (2011) developed exploratory GIS approaches that can model and visualize individual daily travel activities in space and time using activity diary data. Shen et al. (2013) built intra-personal commuting ST paths to look at their flexibility with a GPS dataset collected through activity-travel surveys. Similarly, this paper represents individual representative daily activity paths based on this framework.

2.2 Digital footprints as new data source for mobility studies

Individual movements demonstrated a high degree of temporal and spatial regularity (González et al. 2008, Song et al. 2010), which in turn has motivated researchers to develop various methods to explore and understand the trajectory patterns as short- (Shen et al. 2013) or long-term

movements (Huang and Wong 2015). In those studies, three major data sources are exploited, including: 1) traditional travel diaries, 2) GPS tracks, and 3) social media messages. Each method has its advantages and disadvantages. For example, travel diaries include detailed individual activity information, such as start location and time, end location and time, activity content and so on. ST trajectories could be portrayed using those calibrated data (Jiang et al. 2016). However, conducting travel surveys is expensive, ponderous, and time-consuming (Jiang et al. 2016, Huang 2017). Data collected from human location tracking devices such as cellphones (Zhang et al. 2014, Zhao et al. 2016) were used to model individual mobilities (Kang et al. 2010), social networks (Cranshaw et al. 2010), and urban dynamics (Jiang et al. 2015, Gao 2015). Data from other platforms like on-board GPS devices (Giannotti et al. 2011) and smart cards (Liu et al. 2009, Huang and Tan 2014) have also been used. These data are relatively dense and are of higher resolution comparing to social media data.

Technologies and devices with location services generate a huge volume of ST data capturing individual footprints (Hasan et al. 2013, Huang and Wong 2015). Specifically, social media platforms such as Foursquare (Hasan et al. 2013), Twitter (Huang and Wong 2015), and Facebook (Cranshaw et al. 2010) record time-stamped individual locations with their georeferenced messages. The amount of active Twitter users is increasing every year, reaching an average daily *tweet* (message sent on Titter) number of 500 million in the United States during 2017 (Aslam 2018). Massive amounts of tweets, produced by diverse groups of people, are georeferenced or geo-tagged with time stamps with each tweet indicating a user's digital footprints. Besides, the location accuracy of such tweets can reach up to 10m (Jurdak et al. 2015). Additionally, social media data are publicly accessible and thus free of privacy infringement (Yin and Wang 2016, Huang 2017). Also, online platforms are accessible to a relatively large number of user groups with different SES and demographic backgrounds (Yin and Wang 2016).

Moreover, rich information about human behaviors (Lenormand et al. 2015, Preotiuc-Pietro et al. 2015) and social interactions between users (Cranshaw et al. 2010, Gao and Liu 2014) can be captured through social media data and explored for various applications.

Indeed, digital footprints collected from social media platforms include random activities and movements introducing uncertainty while being utilized to understand human daily travel patterns (Huang and Wong 2015). They are "shallow" data that do not contain much information about individual activity. Nonetheless, they can capture movement behaviors over an extended time period and provide insights into representative activities by being pooled according to the spatiotemporal proximity (Huang and Wong 2015). Social media data therefore can be used to derive regular activity patterns.

Different methods were developed to mine individual travel patterns from online platforms. For example, various quantitative and statistical methods (Cheng et al. 2011, Preoţiuc-Pietro and Cohn 2013, Yin and Wang 2016) were used to investigate the ST travel patterns by analyzing digital footprints. The mobility patterns identified from these studies were consistent with the results using other data sources (Cheng et al. 2011). The regular daily movements captured by sparse long-period tweet data allow us to derive individual representative daily space-time paths (Huang and Wong 2015). In addition, because the activity nature of a location or trip purpose (e.g., shopping, working, eating) can be detected by mining the content of messages posted in the location (Preoţiuc-Pietro and Cohn 2013) or by integrating GIS land use data (Huang et al. 2014), social media data can also provide similar travel activity information offered by other data sources.

However, these studies mostly focus on representing, visualizing, and exploring individual travel patterns (Huang and Wong 2015). Most of these patterns are shown with geographic context information simply provided by a map. While some contexts are extracted as

travel semantics (Huang et al. 2014, Huang and Wong 2016), the classification mechanism to differentiate travel activities is neither comprehensive nor accurate enough. A few studies also examined the trajectory data of a group of users. Although travel semantic information is more often to be attached to those clustered collective travel trajectories, the main purpose of those studies is to detect the popular locations of a city based on individual check-ins (Zhang et al. 2014). Individual semantic travel behaviors cannot be further explored. In those studies, individual trajectory data are first aggregated in space and time. For example, Liu et al. (2009) showed the hot-spots of commuting activities along subway lines in Shenzhen during different time periods using smart card records. Because of the limitation to delineate individual semantic travel behavior patterns, such an analysis framework cannot support the analysis of semantic travel behaviors and patterns of a group of users. As such, existing frameworks are not effective to understand the travel patterns of either an individual or a group of people with varying SES and demographic background. Therefore, using tweet data as an example, this study further probes individual travel patterns from digital footprints of a large number of users based on previous work on social media user activity acquisition and mining (Huang et al. 2014). Specifically, it delineates individual representative travel trajectories with detailed activity information and makes an initial effort to explore the semantic similarity of different individual daily travel patterns via geovisual analytics.

2.3 Detection of activity zones and underlying semantics

Geospatial semantics has been studied to solve the problems about publishing, retrieval, reuse, and integration of geo-data (Janowicz et al. 2013), with a focus on the meaning of geospatial objects (Hu 2017). When working with geospatial data from different sources in varying formats (e.g., travel surveys, remote sensing imagery and digital footprints), it is necessary to explain and manage them meaningfully and consistently. Several research areas of geospatial semantics have

been discussed, including ontology, geospatial semantic web, place semantics, and so on (Hu 2017). This thesis is focused on studying place semantics to deal with geospatial data fusion and mining.

Individual travel trajectories denote a series of places people visit along the time. 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 data (Yan et al. 2013, Cai et al. 2016). Traditional survey data directly describe people' activities at certain places, while cost tremendous labors and resources (Hu 2017). GPS data such as taxi logs record exact OD pairs as well as people's stay time along the way, from which semantics are also inferred combined with geographical context data (Yan et al. 2013). Research has been done to understand the activity sequences indicated by either individual or collective spatiotemporal travel trajectories using those dense data. Different models are proposed for trajectory mining and activity inference, including location categorization, frequent region detection, and so on (Njoo et al. 2015). A typical method to match a location or region with a known activity type is to detect stay points and stay intervals of trajectories and to find geographical context of these stay occurrences, which has also been applied to data collected from social media platforms (Furtado et al. 2013, Njoo et al. 2015, Aurelio Beber et al. 2016, Beber et al. 2017).

Aside from travel histories represented in chronological geo-coordinate series, increasing usage of social media platforms provides extra context information sources such as message contents, uploaded images, and so on (Hu 2017). Besides, digital footprints can expose broader area where people travel around (Cai et al. 2016). Despite these advantages, limited progress has been made to mine semantics from digital footprints. Specifically, detection of stay points and their intervals could be inaccurate using social media data because of data sparsity. Then activity

types identified by those stay points can be either false or missing. For example, although an individual usually visits a restaurant during the noon (e.g., 12 – 1 pm), the individual only tweets at around 12 pm while waiting for lunch, which obscures the exact temporal interval of the visit and could cause the failure of detecting the activity.

Huang et al. (2014) defines the notion of activity zone to detect activity types from digital footprints. In this method, individual travel trajectories are aggregated at first using spatial clustering method such as DBSCAN. Then produced clusters are classified based on a regional land use maps and Google Places application programming interface (API). Such land use data are only published at specific places, such as the state cartography office's website at UW-Madison. Researchers need to search for those data based on their study area. Moreover, while major land use maps can be searched for large areas such as the whole United States, detailed land use data for statewide or citywide areas are made in diverse standards, which adds extra work to classify activity zones consistently. Besides, Google Places API is a service that Google opened for developers and will return information about a place, given the place location (e.g., address or GPS coordinates), in the search request. However, API keys need to be generated before people can use these interfaces and each user can only make a limited number of free-charged requests every day (i.e., 1,000 requests per 24 hours period). In sum, previous methods to detect activity zone types using social media data are not sufficient and can hardly achieve data fusion. Comparing to the high cost of using officially published dataset, emerging VGI data offer another choice.

VGI data have been increasingly used for land use classification (e.g., Fonte et al. 2015, Jiang et al. 2015, Estima and Painho 2013). Previous research detected semantic information of individual trajectories by classifying all land pieces of the research area and matching classification results with their spatially overlapped trajectories (Cai et al. 2016). For sparse data

points collected from social media platforms, it is inefficient to try to identify every land use type since only a small portion of land is covered by enough points. Therefore, this thesis proposes an approach to detect activity zones by first aggregating individual travel trajectories and then using OSM data to determine activity zone types, which produces *semantic hot-spots* (i.e., activity zones) of experimental social media users' travel trajectories. Individual representative travel activities are thus detected by finding the overlapped or surrounding activity zones of the individual travel trajectories.

2.4 Spatial semantic clustering methods

Individual representative travel activities are typically represented as semantic clusters, which can be generated from individuals' travel trajectories using spatiotemporal aggregation methods (Lu et al. 2011). In this way, travel activity information is not provided by relevant semantic descriptions but is extracted from spatiotemporal travel patterns. For example, a semantic region where an individual always stays in the morning with a rather small travel speed could be classified as home. Relatively dense ST points are required to generate the stay points and to calculate the travel speed. Therefore, this method cannot be applied to aggregate sparse digital footprints.

Travel semantics minded from message content are also used to measure the similarity of travel trajectories in order to group together similar ones for more accurate representation of travel activities. For example, Huang et al. (2016) conducted topic modeling to identify similar activities represented by different spatiotemporal clusters and then connected those clusters to identify representative activity types. Still, this method is designed to detect collective travel hot-spots, which cannot be applied to cluster individual travel trajectories because all the message content from a single individual cannot provide enough information for activity detection (i.e., data sparsity). Steiger et al. (2015) detected spatiotemporal and semantic clusters of Twitter data

using unsupervised neural networks, which can compute the similarity between three information layers: message content, temporal information and geographic location. However, this method is only designed for mining collective travel activities.

Aside from attached message content, travel semantics are also obtained from underlying geographic context of digital footprints to improve activity clustering accuracy. Du et al. (2016) imposed geographical background constraints on density-based clustering to yield more appropriate clustering results, which would separate certain activity clusters which should not be merged as one, such as the clusters covering stream segments. However, this study aims to improve spatial clustering while did not establish a complete activity classification mechanism for semantic clustering.

2.5 Geovisual analytics of human mobility

Traditionally, regression methods and simple graphic tools (e.g., charts and histograms) were developed to show travel properties, such as average travel distances over both short- and long-term time periods for individuals or different user groups (Zhang et al. 2012, González et al. 2008, Kang et al. 2010). Especially, radius of mobility gyrations is measured to characterize individual travel patterns (González et al. 2008). However, these quantitative and statistical methods cannot effectively present and support the interactive visual explorations of individual and group activities that are needed to gain in-depth understanding about the user's trajectories (e.g., where and when the users travel).

In order to understand human mobility, geovisualization methods have been developed to explore individual and collective travel patterns (Andrienko et al. 2007). Correspondingly, some studies have integrated 2D maps (e.g., density map, direction distributions, origin-destination flows) to understand individual or collective movement behaviors in a better manner. For example, Yuan and Martin (2014) looked at centroids of most frequent stops using density

maps and diverse direction distributions of several clusters. Zheng et al. (2008) compared semantic sequences of different people groups. Yin and Wang (2016) depicted travel flows between origins and destinations across multiple spatial-temporal scales.

However, 2D static maps lack analysis functions in both space and time dimensions. A series of maps are typically applied to animate the human mobility dynamics during varying time periods. Therefore, analysis systems that integrate temporal analysis, along with the 2D space, under a time geography framework is necessary for studying individual and collective mobility patterns. For example, Shaw and Yu (2009) depicted individual trajectories as space-time paths using travel diary data.

However, those systems are based on commercial software components. They are also unable to handle newly emerging big data. On the other hand, existing online analysis systems for mobility studies are short of interactive functions. For example, users cannot change analysis parameters according to their requirements and interest, which should be better supported to help detect hidden patterns (Kraak 2014).

Both semantic flows (Zheng et al. 2008, Zhang et al. 2012) and historical trajectory shapes (Yuan and Martin 2014) were taken into consideration by previous studies in order to aggregate collective travel patterns. However, only statistical analysis of either of their similarities cannot describe individual travel patterns in an efficient way. Effective geovisual analysis methods need to be developed to explore similarity analysis of individual semantic travel patterns by integrating semantic information. To fill in the technology gap, this study will develop a geovisual system to support generating and exploring individual daily travel patterns with semantics and visually comparing the semantic travel patterns of different individuals.

Chapter 3. Methods

The workflow of this study includes three integrated components illustrated as Figure 1.

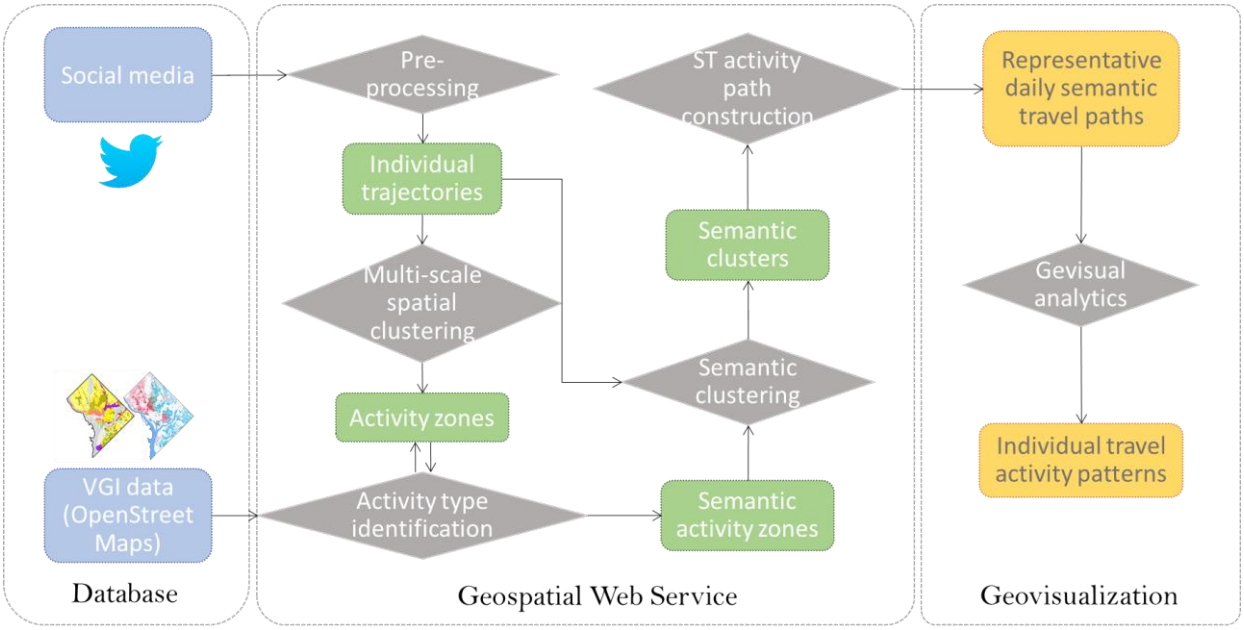


Figure 1. Workflow

1) Data Preprocessing: Preprocess tweet records to remove inactive users and abnormal users (e.g., users representing a company with account being shared by many employees); Represent users' tweet records as a set of ST points with each point representing individual footprint (i.e., presence or visit) in a location at a specific time within a day (24 hours).
2) Multi-scale spatial clustering: Aggregate ST points of all experiment individuals into representative activity zones (i.e., spatial hot spots) using the multi-scale spatial clustering method we propose and derive the centroid of each activity zone as a representative activity location where corresponding activities performed.
3) Activity type identification: Identify the activity type (e.g., residential, service, transportation, and office) of each representative zone by integrating multiple datasets from OSM and assign the activity type to each representative location accordingly.

4) Semantic activity clustering: Aggregate individual ST points into representative activity clusters based on their correlation with the activity zones detected above. Representative activity type of each cluster is thus also detected and appended. Next, the travel semantics (e.g., going to office or taking the metro bus) will be assigned based on activity types of the origin and destination of each ST path segment.
5) ST activity path construction: Connect every two sequential representative activity locations chronologically to derive individual representative daily ST travel paths with semantics.
6) Geovisual analytics: ST paths with semantics indicating various travel activities are depicted in different colors to support geovisual exploration of the individual travel activity patterns. As a result, individual representative travel paths with dwelling areas located at different neighborhoods can be compared visually to characterize travel patterns of these neighborhoods with different SES and demographic background.

The following chapters elaborate these components in detail.

3.1 Data preprocessing

Using Twitter streaming API, users within Madison are identified and their historic data are collected. There are two criteria to select eligible users for this study: 1) First of all, only users with enough geo-tagged tweets (> 50) are selected to generate their representative daily travel paths; and 2) Not all tweets are posted by human beings and one Twitter account can represent an organization or company and be shared by many individuals. Therefore, the users with a relocated speed of two subsequent geo-tagged tweets over 240m/s are discarded (Yin and Wang 2016).

After identifying eligible users, the following step removes invalid geo-tagged tweets from these users. Since this study focuses on local travel patterns, long-distance travels (e.g.,

travels to other cities far way or country) are not taken into consideration. Specifically, a box boundary is predefined to remove invalid tweets which are located outside the target area, and to exclude corresponding long-distance travels. The remaining geo-tweets are then processed and organized as a set of ST points under time geography framework to indicate individual daily trajectories within 24 hours. Each point is represented as:

$$p = <x, y, t>$$

where x and y represent geographical coordinates and t represents timestamp where a social media message was posted.

3.2 Multi-scale spatial clustering

A single ST point can either show the location of a regular activity, which captures frequent individual travel activities while scattering around the frequently visited area or record the location of a random activity. While random ones should be discarded, regular ST points should be aggregated to derive daily mobility patterns. Via aggregating ST points representing individual travel trajectories, ST clusters are generated in previous research (Huang and Wong 2015). However, because of data sparsity, only a few clusters can be detected for most individuals and each cluster usually takes a small area, the geometry of which rarely provides enough context information to identify related activities. For example, a person may regularly tweet at one fixed location of a restaurant so the spatial distribution of these tweets does not cover the geographic feature (e.g., a point of interest, namely POI on a map) that represents the restaurant. The geometry of his tweeting area does not overlap with any other eating activity zone either. Hereby, semantic information of the tweeting locations is therefore missing. Corresponding activity types of these clusters can thus hardly be inferred unless referring to other resources such as message content. To detect semantics indicated by geographic features of these clusters, this thesis develops a method to first aggregate collective travel trajectory hot-spots as representative

activity zones, then infer the type of each activity zone, and finally attach any individual travel trajectory cluster to an activity zone type. When more individuals' footprints are being aggregated, larger clusters are generated to reveal the underlying information of people's regularly visited places.

To date, there are a variety of density-based clustering algorithms for aggregating spatial data points with noise, including DBSCAN (Ester et al. 1996, Bäcklund et al. 2011), ST-DBSCAN (Birant and Kut 2007), varying DBSCAN (VDBSCAN; Daniel et al. 2014), and so on. Comparing to other clustering methods such as k-means and k-medoid (Bäcklund et al. 2011), these density-based clustering methods can detect arbitrary cluster shapes and need less predetermined input parameters that rely on domain knowledge. Among those methods, DBSCAN is widely used in human mobility studies to cluster massive digital footprints (Kang et al. 2010, Huang et al. 2014, Huang and Wong 2015). While using DBSCAN, two parameters should be supplied, including *eps* (neighborhood radius) and *minPts* (the minimum neighbors to consider a point as core point). Given a set of ST points for clustering and minPts value as 4, any point having at least three other points within the distance of eps are considered as a core point, those points reachable from a core point are border points, and the rest are noisy points (Bäcklund et al. 2011).

DBSCAN clustering results vary with different parameter combination of eps and minPts. While 25 meters as eps is applied to cluster dense GPS trajectories (Gong et al. 2015), one single eps fails to detect clusters with different densities from sparse digital footprints (Liu et al. 2007). Several modified algorithms based on DBSCAN have been designed to make up for this defect such as varying DBSCAN (VDBSCAN; Liu et al. 2007), OPTICS (Ankerst et al. 1999), and HDBSCAN (Campello et al. 2015). These methods detect clusters of varying densities based on the distribution of spatial points. However, the clustering relationship of digital footprints is not only dependent on their spatial distribution. A large spatial cluster may be separated into two

smaller ones as it covers two diverse areas which are close to each other. For example, a dining hall close to a teaching building could be clustered into the academic area as people's digital footprints centered on both of them and cannot be spatially separated into two clusters.

Since clustering that is only based on spatial distribution cannot aggregate digital footprints into activity zones accurately, it is meaningless to further identify typical activity type for each zone. In reality, geographical context information can be referred to for aggregation to improve clustering accuracy. Cai et al. (2016) detected similar semantic trajectories from online geo-tagged photos using an optimized OPTICS clustering method based on significant semantic fields extracted from GeoNames and weather observation databases. This grid-based method can produce semantic clusters of varied densities and arbitrary shapes. However, 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 in order to detect arbitrary shapes (Hio et al. 2013). Besides, this method is very sensitive to parameters such as minPts. To experiment with different parameterizations is time-consuming and actually redundant with grid subdivisions (Cai et al. 2016). Instead of grid-based aggregation, this thesis proposes a multi-scale spatial clustering method to mine accurate activity zones more efficiently.

In this thesis, three distinct eps values are used to detect hot-spots: 50m, 100m, and 200m. While 50m can detect rather small clusters, eps as 200 m is reported to be able to identify most areas of interest in cities using geo-tagged Flickr photos (Hu et al. 2015). Therefore, a set of clusters (C1, and C2, and C3) are generated by applying DBSCAN on the ST points representing the digital footprints of all users, at three different scales (Algorithm 1 line 1). Theoretically, a smaller eps separates spatial clusters while a larger eps merges them.

Then, a series of activity zones (Z1, and Z2, and Z3) are generated (Algorithm 1 line 2). Specifically, a convex hull is derived from all the points in each cluster (Huang et al. 2016) to

reveal the geometrical features (e.g., location, shape and range) of each hot-spot. The convex hulls are stored as polygon features which are defined as activity zones in this thesis. At least three different ST points are required to generate such polygons, and representative centroids are thus used to denote activity zones that contain less than three different locations. Within Huang et al. (2014)'s work, a geometric center is calculated as representative centroid of an activity zone and used to represent the activity location of that zone. However, geometric center may be a fake point that does not represent any footprint. Instead, this thesis calculates geometric median to represent activity zone by finding the real ST point that has the largest amount of neighbor points.

Algorithm 1: Multi-scale spatial clustering with geographical context

Input: ST point set
Output: activity zone set (Z1) labeled by zone type (T)
1: generate cluster set C1 (eps: 50m), C2 (eps: 100m) and C3 (eps: 200m) using DBSCAN (minPts: 4)
2: generate activity zone set Z1, Z2 and Z3 by getting convex hull & centroid of each C in C1, C2 and C3
3: identify activity zone type (T) of each Z in Z1, Z2 and Z3 using **Algorithm 2**
4: for each Z in Z1 and each Z' in Z2
5: if T' = T and Z' overlaps Z
6: replace Z with Z'
7: if T' != 'Others' and Z' not overlaps Z
8: add Z' to Z1
9: for each Z in Z1 and each Z' in Z3
10: if T = T1 and Z overlaps Z'
11: replace Z with Z'
12: if T' != 'Others' and Z' not overlaps Z
13: add Z' to Z1
14: return Z1

Next, the associated zone type of each activity zone is identified with Algorithm 2 (Algorithm 1 line 3; Section 3.3). Then smaller activity zones detected with the eps value at the first scale are replaced by a larger zone which is detected with the eps value at the second scale, and shares the same zone type as well as overlaps with all of the smaller zones (e.g., Figure 2a; Algorithm1 lines 4 – 6). Larger activity zones with zone type identified without overlapping with any smaller activity zone are also kept (Algorithm1 lines 7 – 8). Specifically, this multi-scale spatial clustering algorithm will first check the type of each activity zone detected with the eps

value at the first scale, and then compare them with the surrounding activity zone detected with the eps at the next scale. If they have different activity zone types (Figure 2b and 2c), the algorithm will not merge them. This is because mergence could improperly aggregate collapsed activity zones into one with a generalized zone type, which may not be able to represent diverse activity zones within the merged area (Figure 2b and 2c). As a result, important activity zones (i.e., cluster A in Figure 2b; cluster A, and B in Figure 2c) are merged as part of the large surrounding activity zone detected with next scale eps value (i.e., zone Z') and therefore their associated activity zone types are generalized (e.g., Type II in Figure 2b, Type II and III in Figure 2c) and cannot be accurately identified. Next, the algorithm will continue to merge the remaining activity zones based on the activity zones detected with the eps value of the largest scale (Lines 9-13).

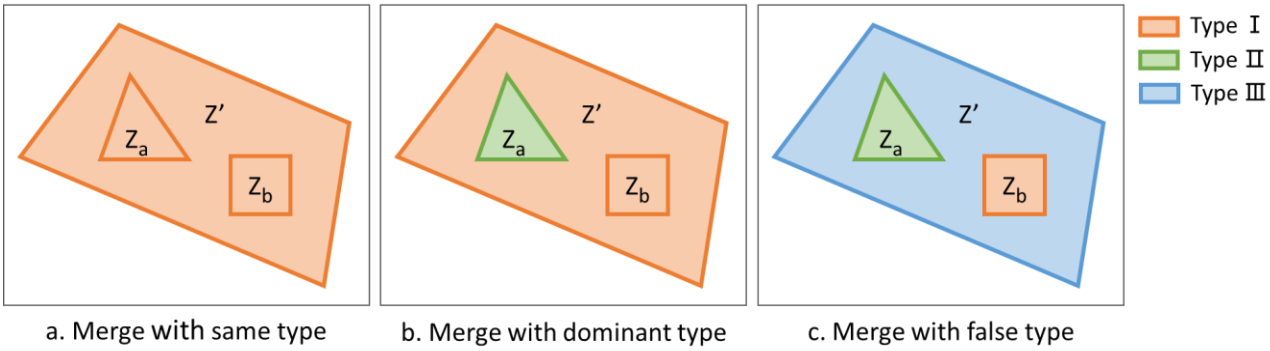


Figure 2. Activity zone type detection at different scales: zone Z_a, $Z_b \in Zi$ (i = 1, 2), detected with a smaller scale eps value; zone Z' $\in$ Zi +1, detected with the next scale eps value

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 corresponding T of F (T = t) referring to table 2
4: set default zone type of Z as t
1: for each other OSM feature F within the target area
2: if F overlaps Z
3: get corresponding T of F referring to table 3
4: if T = "unknown"
5: T = t
4: increase the count of T by 1 in potential zone type map TM = (T, count)

5: sort entries in TM based on the count of T in descending order
6: pick the first T in TM
7: return T

3.3 Activity type identification

Cai et al. (2016) identifies semantic hot-spots using GeoNames and a weather database. GeoNames, a large and one of the most often used gazetteer, contains large amounts of qualified data for geographical context identification (Ahlers 2013). Comparatively, OSM is also an open data source which increasingly contributes location collections. However, the most recent OSM datasets provide more comprehensive location sets with detailed information such as location type and name. For example, more than 20 OSM records could be detected within an activity zone while less than 5 from GeoNames based on our experiment. Besides, aside from point data (e.g., POI dataset), OSM 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), which improves detection resolution as more point coordinates become observed. Therefore, this thesis utilizes OSM data to detect semantics of each spatial cluster. OSM provides datasets with various feature geometries including point, polygon and polyline (Table 1). For simplicity, each record of all the datasets are denoted as POI in this thesis.

Table 1. OSM Dataset Description

Dataset	Feature Geometry	Distinct feature class
landuse	polygon	residential, scrub, park, …, retail, heath, cemetery, industrial
pois	point	dentist, college, cinema, …, restaurant, bank, hotel, laundry, florist, track, camera_surveillance
pois_a	polygon	
pofw	point	27nferred, jewish, hindu, …, 27nferred27, muslim, 27nferr
pofw_a	polygon	
natural	point	cliff, beach, peak, spring, …, glacier, volcano, tree
natural_a	polygon	
traffic	point	waterfall, dam, fuel, stop, …, street_lamp, traffic_signals
traffic_a	polygon	
transport	point	railway_halt, tram_stop, …, ferry_terminal, taxi, bus_stop
transport_a	polygon	

buildings	polygon	cafeteria, clinic, storage, …, university, auditorium, apartment, gym, cinema, mall, station, bridge, barn, offices, brewery, empty
water	polygon	water, wetland, dock, reservoir, river
waterways	polygon	river, stream, drain, canal

The zone type of each activity zone is detected using OSM datasets as described above. Specifically, previous work (Huang et al. 2014) defined eight possible activity types, such as home, office, education, etc., and leveraged GIS land use maps of the study area to infer the activity types for each zone. GIS land use data indicate the types of public or private usage of local lands. Such data are available through online resources such as USGS website and can be used to generate land use maps, with each polygon feature on the map representing a piece of land and being indexed by one of the land use types. Similarly, each polygon feature in the OSM land use dataset is indexed by one of the land use types. While GIS land use types typically include residential, commercial, and service area, etc., different cities or districts could design different land use classification schemes based on their own standard and thus produce different land use type subdivisions. Conversely, OSM classifies all data in a consistent standard.

In this work, ten activity types are defined, including eating, shopping, education, work, health, entertainment, service, dwelling, transportation, and transportation network (Table 3) to capture the majority of activities in an urban area. A spatial join operation is first performed on all activity zones and land use dataset to identify the land use type of each zone, which in turn can indicate whether an activity type is considered as dwelling (e.g., the activity is performed within the area of the residential land use type) or other seven general types (i.e., work and education). Specifically, each OSM land use feature class is projected to a land use type, as is illustrated in Table 2. Given the polygon geometry of an activity zone or the coordinates of a centroid of the activity zone, its overlapped feature class and the projected land use type are returned (e.g., “retail” and “shopping”).

Table 2. Mapping of Land Use Type 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

Given an activity zone, its associated land use type can mostly determine the activity zone type. However, three activity zone types, including eating, transportation, and transportation network cannot be inferred in land use type. Besides, commercial land use type cannot be directly mapped to an activity type and many land use types could include large area with mixed activities (e.g., dwelling area may include eating activities). Therefore, more spatial datasets should be included to further improve the type reference for the activity zones.

Other OSM datasets are thus also spatially joint with activity zones to indicate relevant activities, including poi, buildings, and water datasets (Table 1). Each distinct feature class of these data records (all denoted as POI in this thesis) is projected to an activity zone type (Table 3). As is illustrated in Figure 3, for any POI with an unknown feature class (e.g., blank or meaningless class name), the land use type identified for the activity zone will be considered as the POI's representative zone type. Besides, activity zones with no POIs included will be assigned its land use type if applicable (i.e., the land use type is the same as 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 of its included OSM POIs (Figure 3). However, the most frequently appearing POI activity zone type is not necessarily selected as the type of the activity zone. Specifically, ten activity zone types are divided into four priorities (Table 3). Activities of lower-level types are more specific and could be distributed at smaller scales (i.e., small sized areas),

which could be conducted in activity zones of higher-level priorities. For example, there can be restaurants within the campus and people eat there instead of performing education related 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 choosing others of lower priorities, in case that finely divided zones are misclassified as broader types.

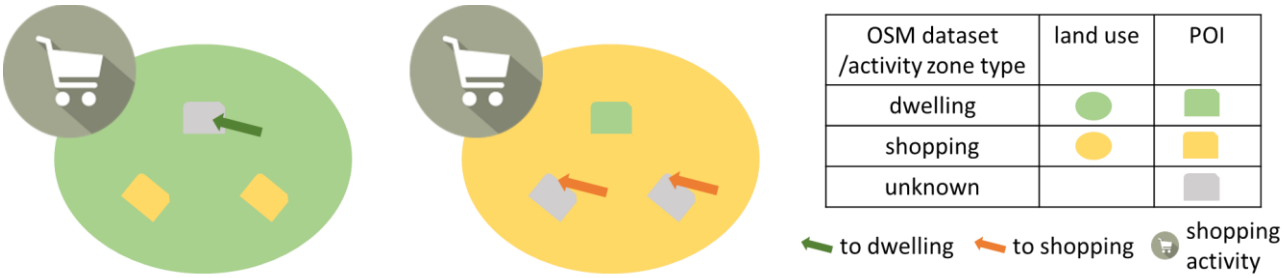


Figure 3. Activity zone type identification in different scenarios

Table 3. Mapping of Activity Zone Types and Other OSM Feature Classes

Activity Zone type	Priority	Distinct feature class
eating	I	bakery, cafe, fast_food, …, restaurant, food_court, caboose, cafeteria
shopping		beauty_shop, bookshop, …, clothes, beverages, bicycle_shop, butcher, computer_shop, florist
education	II	college, kindergarten, …, library, school, academic, university, academic_building
work		embassy, courthouse, …, bank, community_centre, fire_station, police, factory, barn, industrial
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

3.4 Semantic activity clustering

Next, in order to cluster individual travel trajectories, I designed and implemented an improved spatial clustering method, which is defined as semantic activity clustering. Before this work, modified density-based clustering algorithms such as VDBSCAN (Liu et al. 2007) and K-DBSCAN (Debnath et al. 2015) are conducted to identify hot-spots of individual travel trajectories with each hot-spot indicating a regular activity in space. These methods can detect clusters of arbitrary shapes and varied densities. They are designed to aggregate experiment points using only one attribute. For example, most studies cluster people's GPS or digital footprints based on their spatial distribution (Sabarish et al. 2018, Fu et al. 2016, Huang 2017). However, as explained in section 3.2, it is not appropriate to represent individual activities simply using spatial clusters.

After detecting the semantics of the collective activity zones in section 3.2 and 3.3 based on collective travel trajectory hot-spots and OSM data, each individual digital footprint (i.e., ST point) can thus be attached to its overlapped or surrounding collective activity zone, which indicate possible activity of this individual at a specific location. Specifically, a buffer zone of 200m (the same as the largest eps to detect activity zones) for each ST point is generated and used to find the overlapped activity zone of the ST point as individual activity zones. Then most ST points are attached to a collective activity zone that overlaps its buffer zone. ST points without any activity zone attached are considered as noise and discarded. ST points attached to the same activity zone are grouped together as a distinct cluster, which is labeled with the corresponding activity zone type. Particularly, neighboring clusters of the same activity zone type should be merged as current semantic activity clustering method cannot distinct them, although their underlying activity zones fail to be merged during the process of clustering (Section 3.2) because

of inconsistent classification using different eps values. In contrast, a point should not be reclassified if it is already attached to an activity zone.

Each derived cluster indicates an activity zone that an individual visit, which is thus defined as an activity cluster. However, not every activity cluster represents a regular travel activity of the individual. In fact, many clusters capture random activities. Specifically, if an activity cluster includes a large number of ST points, it indicates that an individual frequently visits the cluster area in a typical day (Huang et al. 2014). In order to differentiate regular clusters and random clusters, the minimum number of points for each cluster, similar as minPts parameter in DBSCAN, is defined to remove random ones. As discussed in previous research (Huang and Wong 2015), we use 4 as the minimum number. After removing insignificant clusters, representative spatial activity clusters and their attached activity zones are derived.

3.5 ST activity path construction

Density-based spatial clustering methods cannot easily deal with the time dimension while detecting clusters using ST points (Birant and Kut 2007). In order to handle the time dimension, previous methods divide 24 hours of daily time into consecutive temporal windows and then conducts spatial clustering within each window (Huang and Wong 2015). This approach is successfully applied in clustering individual trajectories with relatively dense ST points though at least four points are required to form an ST cluster. For rather sparse ST points, temporal division further decreases the amount of ST points included at each activity zone, resulting in the detection of many insignificant activity clusters, which will be removed and eventually cause missing of semantic information implied by the trajectory. Besides, arbitrary division of temporal windows introduces a modifiable temporal unit problem (MTUP, Liu et al. 2017), which results in inconsistency of detecting different semantic series using varied parameterizations. The impact could be prominent as every possible activity zone is important for identifying semantic

individual travel trajectory patterns. Other aggregation algorithm such as ST-DBSCAN (Birant and Kut 2007) is also developed. This method can detect ST clusters by integrating space and time dimensions while not well applied for sparse data points collected from social media platforms.

For activity type detection of every individual travel trajectory cluster, Jiang et al. (2015) explores individual activities at the local level using disaggregated land use data referred from widely collected VGI POIs, especially those from Yahoo!. POI types are classified using a machine learning model. Their methods can classify land use types at a city block level. However, disaggregated method is only applied on dense POI datasets. Also, to merge POI data from different data sources is very time-consuming. It is unnecessary to detect activity zone types for every city block as only areas where individual travels around are studied.

On temporal side, it is more important to identify activity sequence instead of specific temporal windows. As individual activity clusters indicate people's regular activities and their spatial locations, time stamps of each record in a cluster provide temporal information for the corresponding activity. A representative temporal interval could be identified for each activity zone, which represents the time slice when most individual footprints appear at the zone (Huang et al. 2014). Using this method, a threshold needs to be defined to find a specific amount of proportions of all ST points within a cluster. However, this threshold is arbitrary and needs further study to formulate selection guidelines.

Similar as Huang and Wong (2015), this study connects the centroids of two sequential representative ST activity zones as a segment of the ST path indicating the OD of travel flows. Differently, after removing tweet records published on weekends, all ST points of an individual trajectory are sorted based on their daily time stamps. Then the points are looped over to count transition frequencies between every activity cluster pair. The centroids of each cluster pair with

non-zero transition frequencies are connected to construct a representative daily travel activity path. The travel semantics (e.g., going to office from home or taking the metro bus to school) of an ST path are eventually visualized at multiple hierarchies using the geovisualization system introduced in the next section.

3.6 Geovisual analytics for individual pattern explorations

Individual daily travel patterns can be explored through directly manipulating (Roth et al. 2015) and visually examining the ST paths, and comparing the paths produced with different model parameterizations. Specifically, a spatial web portal is developed based on Google Maps API to support geovisual analytics in a map context.

As is showed in Figure 4, cartographic representation (Table 4; Roth 2011) schemas are designed to display semantic activity zones generated in section 3.2 - 3.3, individual semantic activity clusters detected in section 3.4 and individual representative ST semantic activity paths derived in section 3.5. Cartographic interaction (Table 4; Roth 2011) functions are also developed to complement static geovisualization for support of user-defined geovisual analytics. This system includes three major components: input widgets to choose specific user id (if applicable) and clustering methods for aggregating individual or collective travel trajectories (e.g., 'HotSpot' for clustering collective travel trajectories using multi-scale spatial clustering, and 'SemanticCluster' for clustering individual travel trajectories), visualization of collective travel hot-spots and individual representative daily travel patterns at a map context, and customized widgets for manually identifying and labeling activity zone types as ground true datasets for experimental validation.

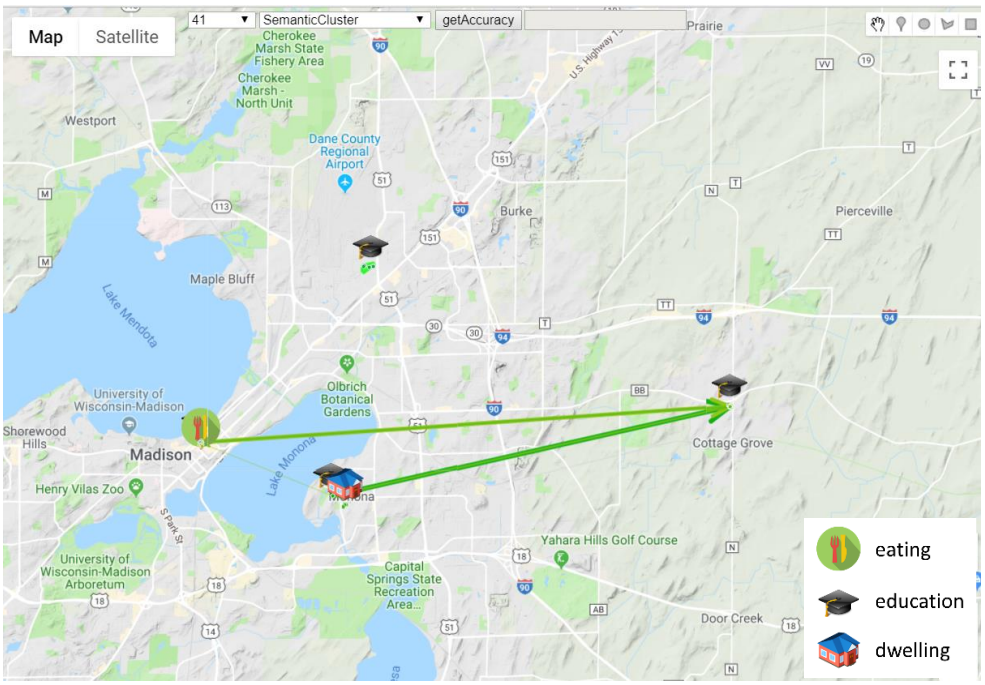


Figure 4. Overview of the geovisual analytical system to show individual representative daily travel ST paths with Google maps context

Table 4. Cartographic Representation and Interaction

Representation		
1	**Basemap**	World map provided by Google Maps API
2	**Activity zones**	Polygon features representing the convex hulls of collective travel trajectory hot-spots shown in different color hues
3	**Activity zone centroids**	Point features located at the centroids of collective travel trajectory hot-spots represented in different icons to indicate the correspondent activity zone type
3	**Activity clusters**	Groups of spatial points with each group indicating a distinct individual activity, represented in different color hues
4	**Routes**	The individual representative ST semantic travel path connecting every OD pair represented by an arrow line of different thickness scaled to the amount of daily transitions between the OD pair. Color will be used to indicate the numbering of the origin cluster.
5	**Legend**	Visual description of the point-symbolizers
Interaction		
1	**Individual Selection**	Retrieve: Objects. Get attached activity zones, zone centroids, activity clusters and routes for a specific individual
2	**Cluster Method Selection**	Retrieve: Objects. Get attached activity zones, zone centroids, activity clusters and routes via a specific cluster method
3	**Pan and Zoom**	Pan and Zoom on the basemap

4	**Retrieve**	Request details of each ST point when clicking it, including message content, timestamp and the cluster it belongs to

To signify the zone types of an activity zone, an image symbol representing one of the ten zone types or a question mark is placed at the geographical median centroid of every activity zone. Eleven different icons are selected to represent ten activity zone types and the unknown type based on their semantic meanings. For example, the icon of a house is used to represent dwelling zones and the question mark is used to represent unknown zones. The shape and color of these icons should be identifiable enough to help relate them with correspondent activities. The size of these icons should be relatively large so that people can immediately differentiate them when observing the map.

Individual daily frequent travel trajectories are visualized as ST paths on the map. Each path segment connects the centroids of two activity clusters for an individual, which denote OD locations respectively. In this thesis work, an arrow line is used to represent the path segment with the arrow end representing destination and the other end representing origin. A path segment is displayed in the same color as the clustered points located at its origin area. A distinct color is assigned to every activity cluster for each individual. After drawing ST path segments between activity clusters, not only are shown spatial locations of corresponding OD pairs, but their related semantic information is also displayed (i.e., traveling types between activity zones).

Chapter 4. Results and Discussion

This study uses publicly accessible digital footprints, collected as geo-tagged tweets via Twitter's streaming API, to identify individual daily activity zones and the corresponding representative semantic travel trajectories. The tweets were posted within the geographic boundary of Madison, Wisconsin from September 2013 to June 2015 were archived. Next, 49 unique users were selected for our experiments in this study with each user having published more than 50 tweet records in total, since these users would have adequate trajectory points to unfold their activity patterns.

4.1 Detection of semantic activity zone

After aggregating all ST points representing these users' geo-tagged tweet records using the multi-scale spatial clustering method discussed in Section 3.2, 401 activity zones are detected and 363 of them are identified with a zone type (Figure 5). Compared with the base map provided by Google Maps API, most activity zones are classified as the zone types that share the same context information as Google Maps. An area around University of Wisconsin-Madison (UW-Madison) campus (Figure 6) is selected to demonstrate the consistency, which indicates that our classification mechanism based on DBSCAN and only VGI (i.e., OSM datasets) are feasible and effective.

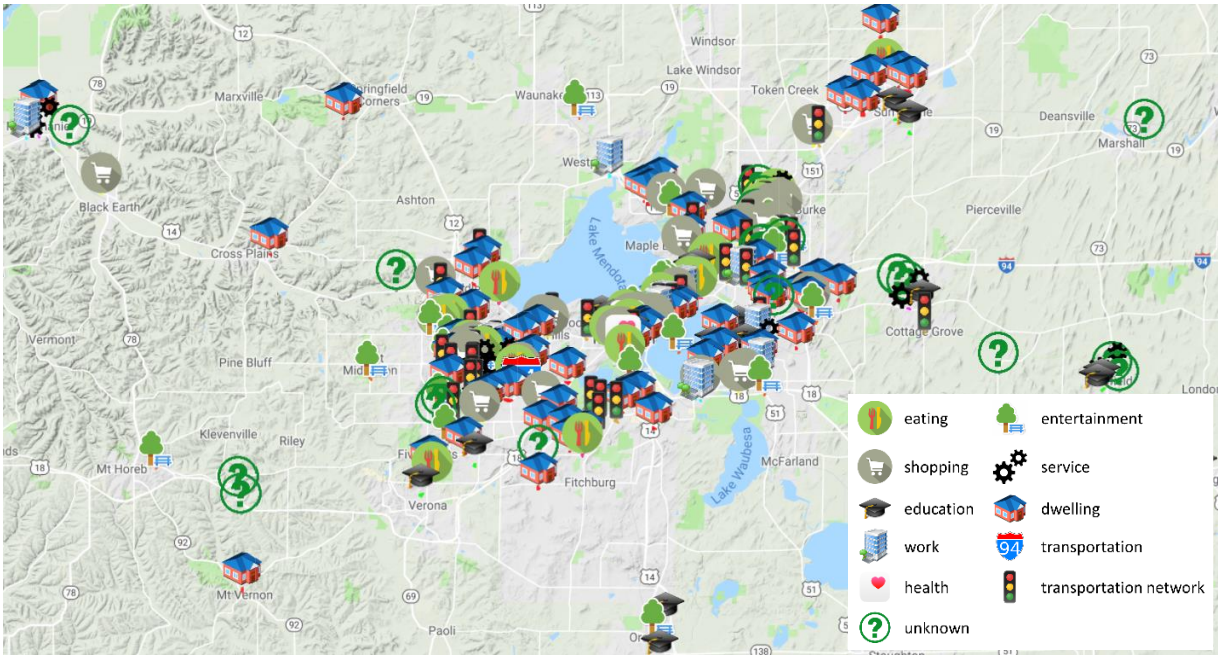


Figure 5. Detection of semantic activity zones at Madison, Wisconsin

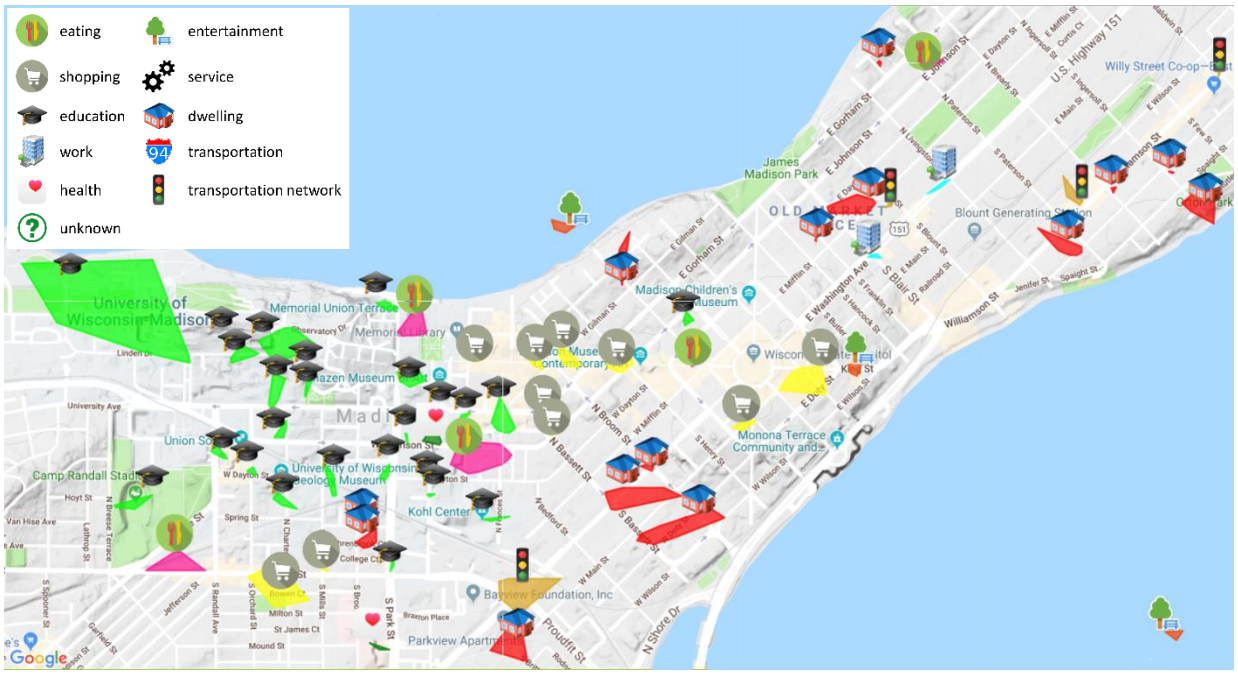
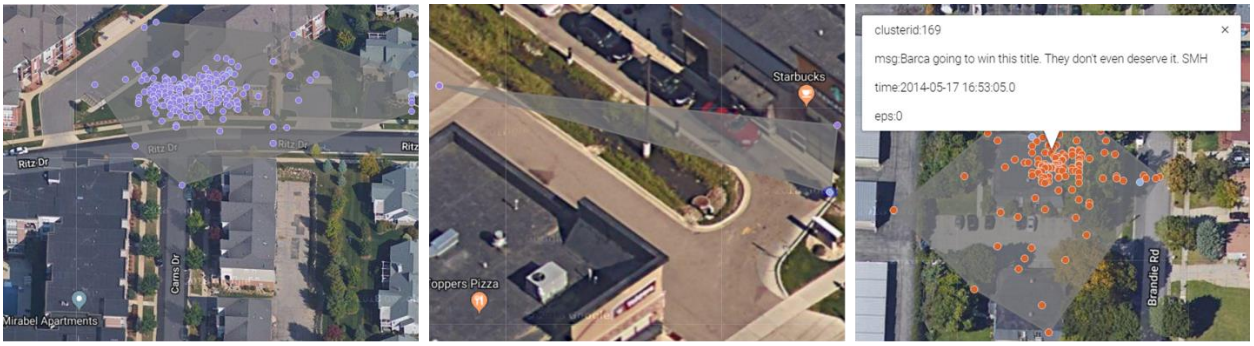


Figure 6. Semantic activity zones around the east campus of UW-Madison

Activity zones symbolized with a question mark ('?') indicate the failure of detecting zone types. As such, we further explore the reasons by manually investigating individual daily footprints within these clusters. It is found that most unclassified points are located at areas with

straightforward geographical contexts provided by Google Maps, such as residential area (Figure 7a) or eating area (Figure 7b). Besides, messages with content of watching sports games are published in some of the unknown clusters (Figure 7c), which also indicate that these clusters are located at residential area and is consistent with the indication of Google Maps. Therefore, failure to detect types of those activity zones are caused by the missing of OSM data at related areas and should be overcome by collecting other open source datasets or by examining social media messages in future work.



a. Unknown zone 1 (residential) b. Unknown zone 2 (eating) c. Unknown zone 3 (residential)

Figure 7. Semantics indicated by Google Maps or online messages for activity zones with an unknown type

4.2 Detection of semantic cluster

Individual travel trajectories are clustered based on their travels among activity zones. Figure 8a shows four clusters (cluster A, B, C, D) detected for a typical user using the proposed semantic activity clustering method, with each cluster displayed in a distinct color. Among these clusters, cluster C and D are visible spatial clusters, which are also detected by either DBSCAN (Figure 8b and 8c) or VDBSCAN (Figure 8d) which is designed to find clusters of varied densities. However, cluster A and B are spatially scattered and are not detected by the VDBSCAN method. Similarly, they cannot be detected by DBSCAN until eps value is increased to 700m, which is very large and actually integrates noise into many clusters. The semantic activity clustering method is thus a feasible approach to detect individual trajectory clusters based on the spatial

distribution of collective semantic activity zones instead of spatial distribution of trajectory footprints.

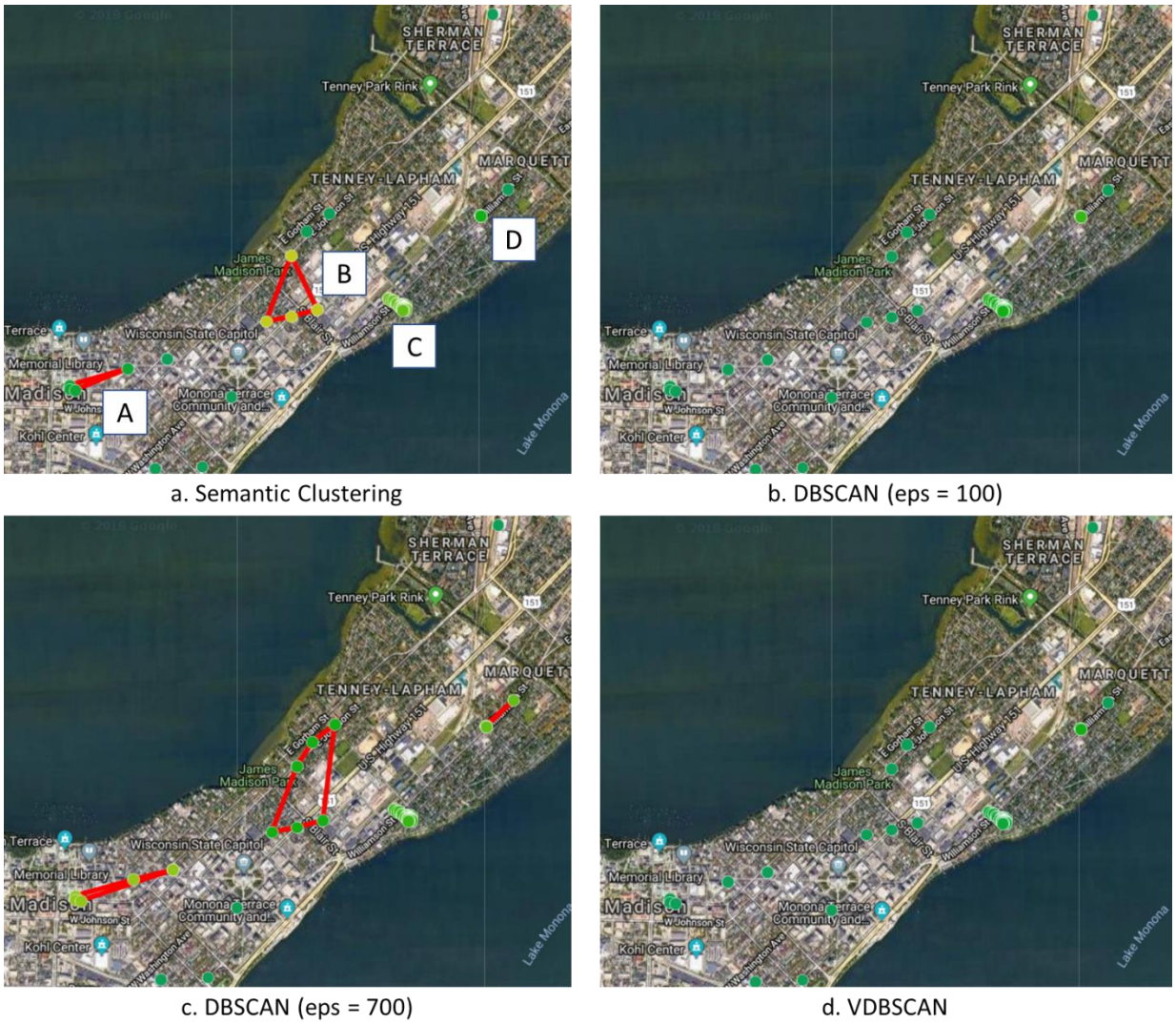


Figure 8. Detection of activity zones using semantic activity clustering, DBSCAN with different eps values and VDBSCAN for selected Twitter user 1

Figure 9 shows the clustering results of another selected Twitter user. Similar with the last example, VDBSCAN (Figure 9d) misses most meaningful clusters as most points appear to be noise according to their spatial distribution. DBSCAN with eps value as 100m (Figure 9b) can only detect some small clusters that are clearly agminate. While DBSCAN with eps as 300m (Figure 9c) detects three clusters, it still trades by merging neighboring smaller clusters such as

cluster B and C. Besides, a best eps value needs to be found for clustering different individual trajectories, which is very tedious and hard to achieve automatic aggregation. Therefore, the semantic activity clustering method can detect spatial clusters from individual digital footprints in a more effective way (Figure 9a).

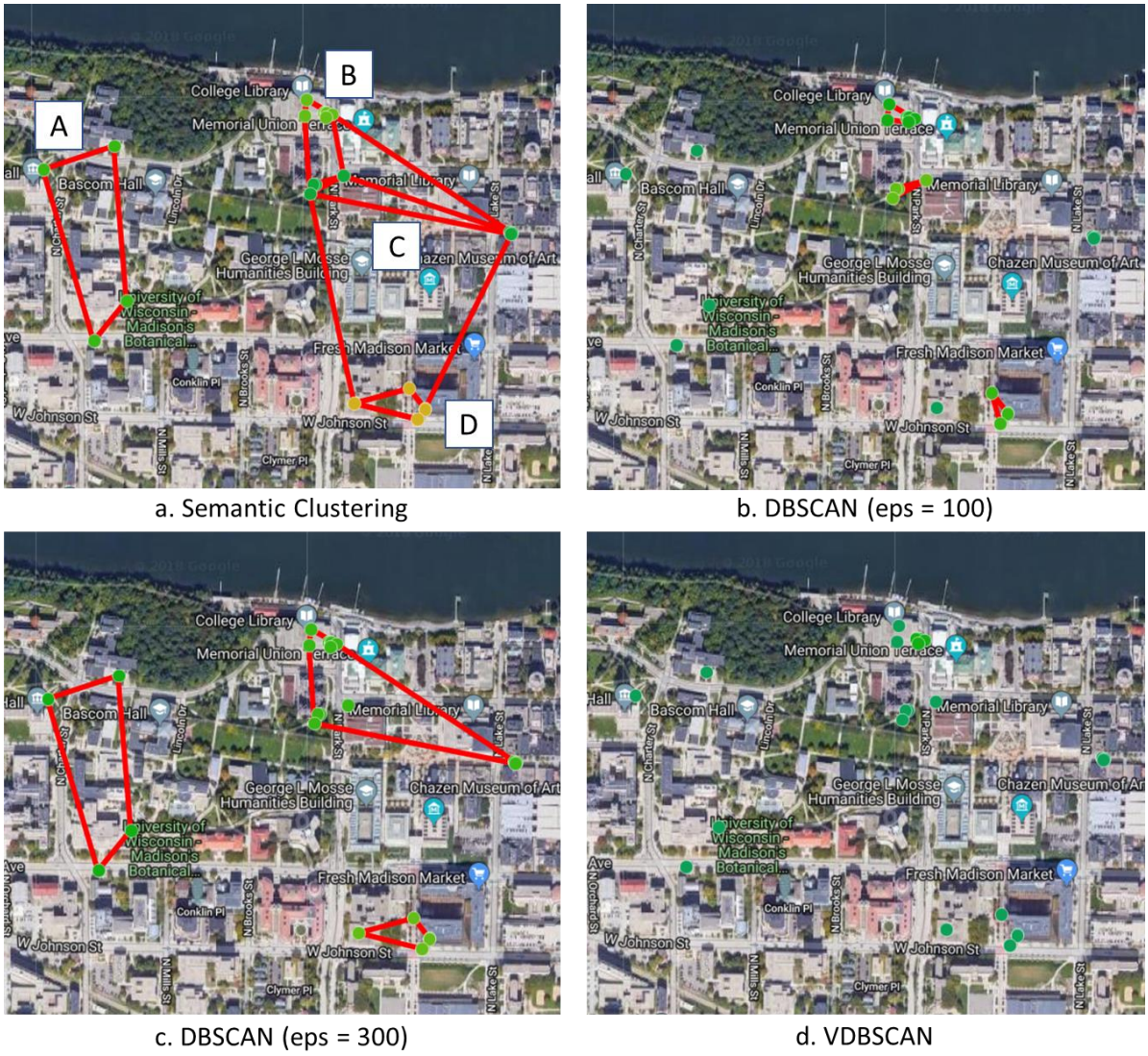


Figure 9. Detection of activity zones using semantic activity clustering, DBSCAN with different eps values and VDBSCAN for a selected Twitter user 2

However, there exist two problems in the multi-scale spatial clustering method. Firstly, we used three increasing eps values: 50m, 100m and 200m, which is based on spatial scales of common geographical objects such as street buildings and large urban areas. eps values beyond

this scope should also be tested to achieve more systematical aggregation. Secondly, overlapped activity zones of different eps values could belong to different zone types. In that case, OSM data within the nonoverlapping parts of zones should be further investigated to detect zone types of these parts.

4.3 Individual semantic daily travel pattern

Four users are selected to demonstrate visualization results of individual representative daily semantic travel trajectory paths (Figure 10). The thickness of each path segment represents transition frequency between two activity clusters it connects, with thicker segments denoting more transitions from its OD in total during a day. For the individual A in Figure 10a, four activity clusters are detected, with four activity zones of three distinct zone types (dwelling, education and entertainment) shown. However, only three travel trajectory path segments are detected out from sixteen possible ones. Missing segment between any two activity clusters indicates that no temporally consecutive tweets belong to the two clusters within one day respectively. Although missing or very thin segments cannot inform us the exact number of corresponding transitions, they can reasonably signify the low probability that the user once traveled from an origin to a destination.

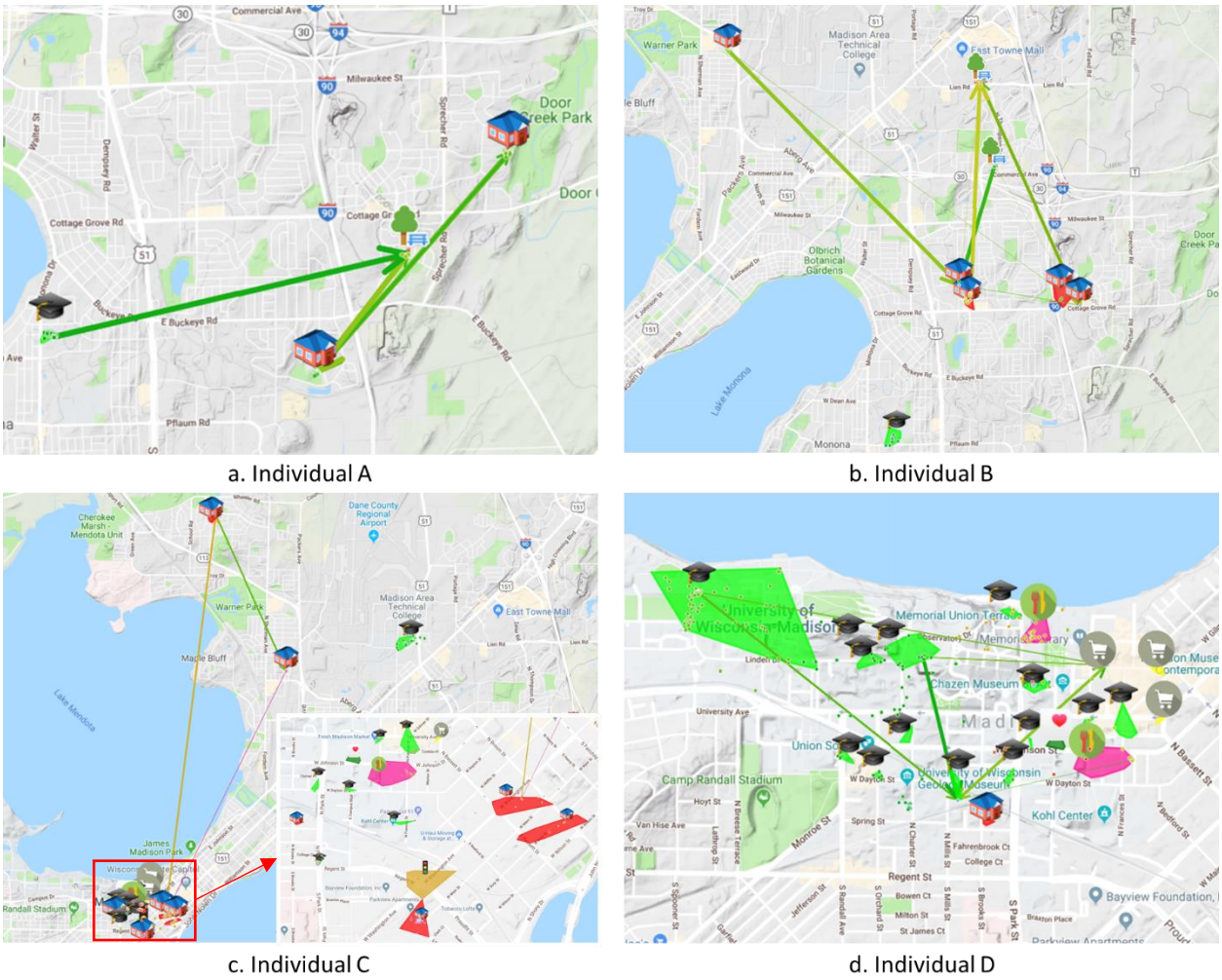


Figure 10. Delineation of the individual representative daily semantic travel trajectories for four selected Twitter users

Conducting mutual comparations among the four individual representative daily travel trajectories shown in Figure 10, it can be observed that the trajectory patterns of the individual A and B are much simpler than the ones of the individual C and D. While only three activity zone types are detected for A and B, there are at least five types being detected for C and D. Through further investigation of the education zones, it can be found that the education zones traveled by A and B are both middle schools, while the ones for C and D are university/college campus areas. Therefore, two different traveler profiles can be defined based on their featured travel patterns:

middle school students and university/college students. It shows that the latter conduct more diverse activities, which relate well with our understanding of these two groups of people.

Chapter 5. Conclusion and Future Research

With the explosive growth of social media data, existing GIS systems do not provide sufficient support for exploring individual travel activity patterns and mining collective activity patterns from sparse but valuable footprint records collected from social media platforms. This study focuses on developing an extensible system for efficient spatiotemporal aggregation of individual ST trajectories with travel activity type identification and effective visual analytics for activity pattern exploration. In future, further efforts could be devoted to improving this work elaborated as below.

- Activity zone detection and mining: In this work, multi-scale spatial clustering is conducted to aggregate digital footprints into collective hot-spots as activity zones. Only three scales of clustering are applied, with eps values as 50, 100, and 200m respectively (Section 3.2). In future, more scales could be further explored to identify the optimal number of scales. Besides, different eps values at different scale levels will be tested to identify optimal eps values for each scale. Furthermore, larger activity zones, detected with eps at a higher scale, at the collective level using the digital footprints of all users, are deleted if their zone types are not the same with their overlapped smaller zones with eps at a lower scale, and therefore are not used for identifying potential activity zone types of ST points that are within these zone areas, at the individual level. In future work, given such a larger activity zone, a post-processing step will be developed to first cut the overlap area of the larger activity zone with smaller zones, and to detect the activity type of the remaining area, which in turn can be used for the subsequent semantic activity clustering process (Section 3.4).
- Activity zone type identification: In this work, the type of each activity zone is identified by counting the activity zone types mapped by all the POIs which overlap the activity zone and then taking the type which appears the most frequently (i.e., maximum votes). Instead of

associating a single type to each activity zone, there may be multiple kinds of activities conducted there. For example, eating and education-related activities may both happen in a student center. Since only location information cannot point to an exact activity, multiple activities will be labeled for the activity zones of certain finesorted types (e.g., student center; Carvalho and Freitas, 2009).

- Individual trajectory clustering: To infer the semantics of individual travel trajectories based on digital footprints, 200m is selected as the radius of buffer zone to search the activity zones surrounding each footprint point (Section 3.4). In future, different buffer zone radius should be assigned to each activity zone based on its zone type. For example, eating zones could be assigned smaller radius than dwelling zones as the former tend to have smaller areas. Additionally, aside from visual investigation of the clustering result, statistical analysis should be conducted to validate the effectiveness of semantic activity clustering method and its superiority over other methods. Specifically, a certain number of users should be held back and their ST trajectory data points will be manually classified based on their underlying geographic context and the investigation of their typical message content. Besides, the sensibility (i.e., optional choice) of several parameters will be tested, such as the number of activity zone types, minPts values for DBSCAN and the minimum point number threshold for eligible semantic activity clusters. Considering the uncertainties of using semantic clusters to represent travel activities (e.g., uncertainty of cluster boundaries and uncertainty of activity types), different classification schemas and manual label standards could be established based on the specific needs of system users. For example, eating, shopping and entertainment (versus education and dwelling) activities should maintain higher priorities when identifying activity zone types and testing the classification results for tourist users.

- Geovisual analytics: This paper designed cartographic representation and implemented key interactive functions for the geovisualization system to display collective travel hot-spots and individual representative daily semantic travel trajectories. Both spatial and temporal travel patterns could be explored in this system. Some existing cartographic representation will be adjusted and more interactive functions will be developed (Table 5) for future research. Particularly, semantic travel trajectories could be aggregated for different week day and weekend, as temporal analysis, to explore travel pattern variations through a week.

Table 5. Cartographic Representation and Interaction to be improved

Representation (to be adjusted)		
1	**Legend**	Add legend of other cartographic elements representing activity clusters, ST points and representative ST travel paths
2	**Routes**	Use curving flows instead of arrow lines to represent representative ST travel paths
Interaction (to be added)		
3	**Filter Panel**	Filter: Objects and time. Adjust the initial query parameters (e.g., years, months, day of week). Search: individual daily transitions
4	**Overlay**	Add animation to better represent the origin and destination connected by individual representative ST travel paths

- Collective travel pattern exploration: As individual representative daily semantic travel trajectory paths are constructed, trajectory similarities could be calculated and quantified to obtain collective semantic travel patterns. While this thesis did not complete this step yet, this goal will be achieved based on our analysis framework integrated with new spatial aggregation and activity zone type detection methods.

References

AHLERS, D., 2013. Assessment of the accuracy of GeoNames gazetteer data. Proceedings of the 7th Workshop on Geographic Information Retrieval. Orlando, Florida: ACM, 74-81.

ANDRIENKO, G., et al. 2007. Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science, 21(8), 839-857.

ANDRIENKO, N. and ANDRIENKO, G. 2012. Visual analytics of movement: An overview of methods, tools and procedures. Information Visualization, 12(1), 3-24.

ANKERST, M., et al., 1999. OPTICS: ordering points to identify the clustering structure. Proceedings of the 1999 ACM SIGMOD international conference on Management of data. Philadelphia, Pennsylvania, USA: ACM, 49-60.

ASGARI, F., GAUTHIER, V. and BECKER, M. 2013. A survey on Human Mobility and its applications. CoRR, abs/1307.0814.

ASLAM, S., 2018. Twitter by the Numbers: Stats, Demographics & Fun Facts.

AURELIO BEBER, M., et al., 2016. TOWARDS ACTIVITY RECOGNITION IN MOVING OBJECT TRAJECTORIES FROM TWITTER DATA.

BÄCKLUND, H., HEDBLOM, A. and NEIJMAN, N. 2011. DBSCAN: A Density-Based Spatial Clustering of Application with Noise. Linköpings Universitet – ITN.

BEBER, M. A., et al. 2017. Individual and Group Activity Recognition in Moving Object Trajectories. JIDM, 8, 50-66.

BIRANT, D. and KUT, A. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering, 60(1), 208-221.

CAI, G., LEE, K. and LEE, I., 2016. Mining Semantic Sequential Patterns from Geo-Tagged Photos. 2016 49th Hawaii International Conference on System Sciences (HICSS). Australia, 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.

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., Exploring Millions of Footprints in Location Sharing Services. ed. ICWSM, 2011.

CORBETT, J., 2001. Torsten Hägerstrand, Time Geography. CSISS Classics. UC Santa Barbara: Center for Spatially Integrated Social Science.

CRANSHAW, J., et al., 2010. Bridging the gap between physical location and online social networks. Proceedings of the 12th ACM international conference on Ubiquitous computing. Copenhagen, Denmark: ACM, 119-128.

DANIEL, G. P., VALÊNCIO, C. R. and RODRIGUES, R. C. C., 2014. VDBSCAN*: An efficient and effective spatial data mining algorithm using GPU.

DEBNATH, M., TRIPATHI, P. K. and ELMASRI, R. 2015. K-DBSCAN: Identifying Spatial Clusters with Differing Density Levels. 2015 International Workshop on Data Mining with Industrial Applications (DMIA), 51-60.

DU, Q., et al. 2016. Density-Based Clustering with Geographical Background Constraints Using a Semantic Expression Model. 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. Proceedings of

the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press, 226-231.

ESTIMA, J. and PAINHO, M., 2013. Exploratory analysis of OpenStreetMap for land use classification. Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information. Orlando, Florida: ACM, 39-46.

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.

FU, Z., et al., 2016. A Two-Step Clustering Approach to Extract Locations from Individual GPS Trajectory Data.

FURTADO, A. S., FILETO, R. and RENSO, C. 2013. Assessing the Attractiveness of Places with Movement Data. JOURNAL OF INFORMATION AND DATA MANAGEMENT, 4(2).

GAO, H. and LIU, H., 2014. Data Analysis on Location-Based Social Networks. 165-194.

GAO, S. 2015. Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age. Spatial Cognition & Computation, 15(2), 86-114.

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.

GIANNOTTI, F., et al. 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal, 20, 695-719.

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.

GONG, L., YAMAMOTO, T. and MORIKAWA, T., 2016. Comparison of Activity Type Identification from Mobile Phone GPS Data Using Various Machine Learning Methods.

GONZÁLEZ, M. C., HIDALGO, C. A. and BARABÁSI, A.-L. 2008. Understanding individual human mobility patterns. Nature, 453, 779.

GUNDLEGÅRD, D., et al. 2016. Travel demand estimation and network assignment based on cellular network data. Computer Communications, 95, 29-42.

HASAN, S., ZHAN, X. and UKKUSURI, S. V., 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. Chicago, Illinois: ACM, 1-8.

HIO, C., et al., 2013. A Hybrid Grid-based Method for Mining Arbitrary Regions-of-Interest from Trajectories. Proceedings of Workshop on Machine Learning for Sensory Data Analysis. Dunedin, New Zealand: ACM, 5-12.

HU, Y. 2017. Geospatial Semantics. Comprehensive Geographic Information Systems.

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. 2017. Mining online footprints to predict user's next location. International Journal of Geographical Information Science, 31(3), 523-541.

HUANG, Q., CAO, G. and 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., 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. and 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. and WONG, D. W. S. 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, X. and TAN, J., 2014. Understanding spatio-temporal mobility patterns for seniors, child/student and adult using smart card data.

JANOWICZ, K., SCHEIDER, S. and ADAMS, B., 2013. A geo-semantics flyby. Proceedings of the 9th international conference on Reasoning Web: semantic technologies for intelligent data access. Mannheim, Germany: Springer-Verlag, 230-250.

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.

JIANG, S., et al. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. Proceedings of the National Academy of Sciences, 113(37), E5370.

JURDAK, R., et al. 2015. Understanding Human Mobility from Twitter. PLoS ONE, 10(7), e0131469.

KANG, C., et al., Analyzing and geo-visualizing individual human mobility patterns using mobile call records. ed. 2010 18th International Conference on Geoinformatics, 18-20 June 2010 2010, 1-7.

KANG, J. and YONG, H.-S. 2010. Mining Spatio-Temporal Patterns in Trajectory Data. JIPS, 6, 521-536.

KRAAK, M.-J., 2014. Introduction to Geovisual Analytics.

KWAN, M.-P. 2000. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. Transportation Research Part C: Emerging Technologies, 8(1), 185-203.

LENORMAND, M., et al. 2015. Corrigendum: Influence of sociodemographic characteristics on human mobility. Scientific Reports, 5, 12188.

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, L., et al., Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. ed. 2009 12th International IEEE Conference on Intelligent Transportation Systems, 4-7 Oct. 2009 2009, 1-6.

LIU, P., ZHOU, D. and WU, N., VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise. ed. 2007 International Conference on Service Systems and Service Management, 9-11 June 2007 2007, 1-4.

LIU, X., et al., 2017. The impact of MTUP to explore online trajectories for human mobility studies. Proceedings of the 1st ACM SIGSPATIAL Workshop on Prediction of Human Mobility. Redondo Beach, CA, USA: ACM, 1-9.

LU, C.-T., et al., 2011. A Framework of Mining Semantic Regions from Trajectories. International Conference on Database Systems for Advanced Applications. Berlin, Heidelberg: Springer, 193-207.

MILLER, H. J. 2005. A Measurement Theory for Time Geography. 37(1), 17-45.

NJOO, G., et al., 2015. A fusion-based approach for user activities recognition on smart phones.

PINJARI, A. R. and BHAT, C. R., 2011. Activity-based Travel Demand Analysis. A Handbook of Transport Economics. Edward Elgar Publishing.

PRED, A., 1977. A Choreography of Existence—Comments On Hägerstrand's Time-Geography and Its Usefulness.

PREOŢIUC-PIETRO, D. and COHN, T., A temporal model of text periodicities using Gaussian Processes. ed. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013 Seattle, Washington, USA, 977-988.

PREOTIUC-PIETRO, D., LAMPOS, V. and ALETRAS, N., An analysis of the user occupational class through Twitter content. ed. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, July 26-31, 2015 2015 Beijing, China, 1754–1764.

ROTH, R., 2011. Interacting With Maps: The Science And Practice Of Cartographic Interaction.

ROTH, R., S. ROSS, K. and MACEACHREN, A., 2015. User-Centered Design for Interactive Maps: A Case Study in Crime Analysis.

SABARISH, B., RAMESH, K. and GIREESHKUMAR, T., 2018. Clustering of Trajectory Data Using Hierarchical Approaches. 215-226.

SHAW, S.-L. and YU, H. 2009. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical–virtual space. Journal of Transport Geography, 17(2), 141-149.

SHEN, Y., KWAN, M.-P. and CHAI, Y., 2013. Investigating commuting flexibility with GPS data and 3D geovisualization: A case study of Beijing, China.

SONG, C., et al. 2010. Limits of Predictability in Human Mobility. Science, 327(5968), 1018.

STEIGER, E., RESCH, B. and ZIPF, A., 2015. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks.

WAN, Y., ZHOU, C. and PEI, T. 2017. Semantic-Geographic Trajectory Pattern Mining Based on a New Similarity Measurement. 6(7), 212.

WU, L., et al. 2014. Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. PLoS ONE [Internet], 9(5), e97010.

YAN, Z., et al., 2013. Semantic Trajectories: Mobility Data Computation and Annotation. [online]. Available from: http://infoscience.epfl.ch/record/177359/files/ACM-TIST.pdf [Accessed 2012].

YIN, D. and WANG, S., 2016. Sensing Spatial Structures through Large-scale Social Media. The Association of American Geographers 112nd Annual Meeting. San Francisco, California, USA.

YUAN, Y. and MARTIN, R., 2014. Measuring similarity of mobile phone user trajectories– a Spatio-temporal Edit Distance method.

ZHANG, D., et al., 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. Proceedings of the 20th annual international conference on Mobile computing and networking. Maui, Hawaii, USA: ACM, 201-212.

ZHANG, W., LI, S. and PAN, G., Mining the semantics of origin-destination flows using taxi traces. ed. UbiComp, 2012.

ZHAO, K., et al., Urban human mobility data mining: An overview. ed. 2016 IEEE International Conference on Big Data (Big Data), 5-8 Dec. 2016 2016, 1911-1920.

ZHENG, Y. 2015. Trajectory Data Mining: An Overview. ACM Transaction on Intelligent Systems and Technology.

ZHENG, Y., et al. 2008. Mining user similarity based on location history.

ZHI, Y., et al. 2014. Urban spatial-temporal activity structures: a New Approach to Inferring the Intra-urban Functional Regions via Social Media Check-In Data. CoRR, abs/1412.7253.