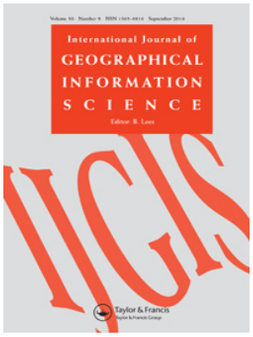

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?

Qunying Huang & David W. S. Wong

To cite this article: Qunying Huang & David W. S. Wong (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, DOI: 10.1080/13658816.2016.1145225

To link to this article: https://doi.org/10.1080/13658816.2016.1145225

Published online: 22 Feb 2016.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
VOL. 30, NO. 9, 1873–1898
http://dx.doi.org/10.1080/13658816.2016.1145225

Taylor & Francis
Taylor & Francis Group

Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?

Qunying Huang [a] and David W. S. Wong [b,c]

[a]Department of Geography, University of Wisconsin-Madison, Madison, WI, USA; [b]Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA; [c]Department of Geography, University of Hong Kong, Pokfulam, Hong Kong

ABSTRACT

Individual activity patterns are influenced by a wide variety of factors. The more important ones include socioeconomic status (SES) and urban spatial structure. While most previous studies relied heavily on the expensive travel-diary type data, the feasibility of using social media data to support activity pattern analysis has not been evaluated. Despite the various appealing aspects of social media data, including low acquisition cost and relatively wide geographical and international coverage, these data also have many limitations, including the lack of background information of users, such as home locations and SES. A major objective of this study is to explore the extent that Twitter data can be used to support activity pattern analysis. We introduce an approach to determine users' home and work locations in order to examine the activity patterns of individuals. To infer the SES of individuals, we incorporate the American Community Survey (ACS) data. Using Twitter data for Washington, DC, we analyzed the activity patterns of Twitter users with different SESs. The study clearly demonstrates that while SES is highly important, the urban spatial structure, particularly where jobs are mainly found and the geographical layout of the region, plays a critical role in affecting the variation in activity patterns between users from different communities.

ARTICLE HISTORY
Received 7 July 2015
Accepted 15 January 2016

KEYWORDS
Socioeconomic status; urban spatial structure; activity zones; Twitter; spatial clustering; social networks

1. Introduction

Understanding activity patterns has significant implications in many planning processes and decision supports (e.g., Kwan 1998; National Research Council 2002). An important aspect of analyzing activity patterns is to determine if individuals intersect spatiotemporally (e.g., Shaw and Yu 2009; Yu 2007). The lack of interaction is believed to induce prejudice and therefore may enhance segregation (e.g., Massey and Denton 1993; Pettigrew 2008). Evaluating segregation should consider the activity spaces of individuals (e.g., Wong and Shaw 2011). In addition, interaction among population groups and segregation are believed to have significant implications in various societal aspects, including public health (e.g., Williams and Collins 2001; Kwan 2013; MacPherson and Gushulak 2001; Matthews and Yang 2013). Thus, augmenting our capabilities to

CONTACT Qunying Huang qhuang46@wisc.edu Department of Geography, University of Wisconsin-Madison, Madison, Wisconsin, 53706

describe, analyze and explain differences in activity patterns among population groups is significant to address various societal issues.

Studies of human activity patterns rely on data tracking individual movements. A common source of such data is travel diary (e.g., Kwan 2008; Chen *et al*. 2011). But gathering travel diary data is usually expensive and laborious. These days GPS devices are commonly used for gathering travel diary-type data. Recently, cell phone data, which are not intended for tracking, have been used to study human activities (e.g., Järv *et al*. 2014; Silm and Ahas 2014). Accessing this type of data derived from the use of information and communications technology (ICT) is privileged, but the number of subjects included can be large. The data may also cover a relatively long period.

Another potential source of data to support activity pattern studies is social media data. Among many popular social media services, data provided by Twitter have been used often. Most geographical studies using Twitter data focused on aggregated spatial patterns, such as the movements of a population group participating in specific events (e.g., Stefanidis *et al*. 2013) or the dispersion of ideas (e.g., Tsou *et al*. 2014). While using Twitter data to study individual activity patterns has to deal with the sparseness and irregularity of sampling points over space and time, identifying the regular locations where an individual performs various daily activities is not straightforward (Huang and Wong 2015). In addition, Twitter data usually do not reveal much information about the socioeconomic status (SES) and demographic characteristics of users unless the textual tweets are mined (Mislove *et al*. 2011), but success is not guaranteed. Therefore, using tweets or similar types of social media data to study individual activity patterns can be challenging. However, acquiring Twitter data is inexpensive, if it costs at all. The data gathering period may last for months or even years. In addition, access to Twitter service is quite ubiquitous globally. Thus, Twitter data may reflect more comprehensive activity patterns, including activities beyond the intra-urban scale, an area that previous studies have not explored thoroughly.

Therefore, *a major objective of this study is to explore the extent that Twitter data can be used to support activity pattern analysis*. Specifically, the study is intended to explore how Twitter data may be combined with other large scale survey data to shed lights on the variability of activity patterns between population groups of different SESs. After a systematic literature review, Steiger *et al*. (2015) concluded that using twitter data for studies in urban planning and management is very limited, if they exist (p. 19). They also pointed that there exists a gap in the literature of 'leveraging other data sources … for data fusing …' (Steiger *et al*. 2015, p. 21). Our work reported here fills these gaps in the literature. Although the literature has thoroughly investigated the variations in commuting distances across different SES groups, the results have not been overwhelmingly conclusive. Therefore, another objective of this study explores if additional evidence can be provided by Twitter data. *This study will assess the roles of urban structure in affecting the activity patterns of residents, particularly, if the residential locations of the disadvantaged population group are of any significance*. These results may assist the formulations of future social programs and policies to level the playing field. Twitter data provide a geographical scale dimension to the study of activity patterns by including locations beyond the region where the users resided. We will assess the variations of activity patterns at three geographical scale levels (international, national, and regional)

between the rich and poor, a simplified societal context, as revealed by the geo-tagged tweets posted by users who likely belonged to these groups.

2. Activity patterns, segregation and social media data

2.1. *Activity space, interaction–segregation and socioeconomic status*

Activity pattern analysis is critical in understanding the potential interaction between population groups of different SES–demographic characteristics and racial–ethnic relationship. Lack of interaction between population groups may induce prejudice and misunderstanding (see, e.g., Pettigrew 2008) and inter-group interaction can reduce some aspects of segregation (Pettigrew and Tropp 2008). Segregation may restrict the disadvantaged groups from accessing critical resources and life-changing opportunities. Thus, these restrictions can be translated into differentials in many economic, social and health-related outcomes across population groups (Williams and Collins 2001; Echenique and Fryer 2007).

While most studies of the place-based approach to segregation relied on aggregated or ecological data and focused on the lack of access to opportunities and resources, people-based approach, which focused on interaction, relies on individual level data (Matthews 2011). Nevertheless, an underlying premise of both approaches is that individuals visit places not only to seek opportunities and resources, but also to interact with people of the same or different groups. Thus, understanding the activity patterns at the individual level can provide great insight about improving the potential well-being of the disadvantaged population groups.

Using travel diary data collected from surveys of three selected types of neighborhood in Beijing, Wang *et al*. (2012) compared the respondent's activity spaces among two types of gated communities and ordinary neighborhoods in terms of their geographical extensiveness (extensity), the intensity of visiting certain locations, diversity of visited places, and isolation or exclusiveness of visited places (exclusivity). Results reflected different activity patterns due to SES, but the roles of neighborhoods, conditioning the influences of SES, were not assessed. Using household survey data to recall visited locations, Jones and Pebley (2014) analyzed the relationships between individual's socio-demographic characteristics and characteristics of activity spaces. One of their conclusions was that individuals were exposed to more heterogeneous social environments outside of their residential space. Thus, exploring activity patterns outside of residential space should be critical in understanding the social–spatial relationships between population groups.

Recent people-based studies along the racial–ethnic lines have exploited cell phone data. Using cell phone data, Silm and Ahas (2014) were able to locate Estonia- and Russian-speaking residents spatiotemporally, and to evaluate their segregation levels at different times (windows) of a day, extending the measure of segregation beyond the residential space (Wong and Shaw 2011). Also using the Estonia cell phone data, Järv *et al*. (2014) created the activity spaces of individuals over different temporal frames. They suggested that the differences in the characteristics of activity spaces (sizes and numbers of locations) between the Estonia- and Russian-speaking populations might be associated with their personal demographic factors and work and residence locations.

Farber *et al*. (2012) also adopted the activity space concept to evaluate the isolation–exposure between two linguistic groups in the French-Canadian region.

All these studies point to the direction that studying activity patterns is more comprehensive than what is offered by the 'sedentarist' approach that focuses solely on residential locations (Järv *et al*. 2014; Sheller and Urry 2006). These studies also attempted to explain variations in activity patterns among population groups due to a single factor. In the Beijing study by Wang *et al*. (2012), differences in activity patterns were attributable to SES, but other potential factors such as family status were not considered (e.g., Johnston 1976; Davies and Murdie 1993).

On the other hand, Järv *et al*. (2014) attempted to explain the variation of activity patterns through a number of ethnic, demographic, workplace and phone usage variables. Although the language variable which represents ethnicity was significant most of the time, other variables also partially explain the differences in particular situations. Similarly, Jones and Pebley (2014) found that besides race, education, age, working condition and residing in certain areas could explain activity-space variation. The relationships between activity space and various factors (SES, travel preference, resident's image of urban spatial structure and home characteristics) have been thoroughly discussed and theorized (e.g., Horton and Reynolds 1971). The importance of urban structure, specifically, locations of various activities and c variables have been highlighted (Hanson 1982; Pas 1984; Ewing and Cervero 2001; Buliung and Kanaroglou 2006b).

Many of these studies were able to explain variations in activity patterns by a variety of factors. These studies were supported by data recording the movement of individuals and reflecting their SES, demographic and racial–ethnic characteristics. This paper explores the use of Twitter data in activity space analysis. Like most social media data, Twitter data are 'thin' as they do not provide much information about the background of the users (Huang and Wong 2015). Therefore, we likely cannot disentangle the interaction of the multi-faceted and multi-level factors in affecting activity patterns using Twitter data, but we attempt to evaluate the utilities of social media data in supporting this type of studies.

2.2. *Social media data and activity pattern analysis*

Studies on activity space, including those mentioned before, have relied on data from three major sources: surveys from respondents in the form of travel diaries (e.g., Wong and Shaw 2011; Farber *et al*. 2012; Wang *et al*. 2012; Jones and Pebley 2014), cell phone data recording where and when calls were made (Järv *et al*. 2014; Silm and Ahas 2014), and movement data collected from GPS tracking devices (Shen *et al*. 2013). However, all these data are privileged and expensive to acquire. Travel diary data require recruiting subjects. They have their strengths and weaknesses in supporting activity pattern analysis (Palmer *et al*. 2013; Huang and Wong 2015).

An alternate data source that may be used in mobility studies is social media data. Twitter and other social media data are not intended for tracking movements, and are gathered in a passive manner (data users cannot control when data are gathered). Thus, using Twitter data to conduct activity pattern analysis has to overcome many challenges, partly due to their sparsity and irregularity spatiotemporally (Huang and Wong 2015).

An obvious problem with social media data is biasness. First of all, users of social media networks constitute a group biased toward the younger and tech-savvy individuals, excluding the economically disadvantaged and older populations. Not all social media data have location information. For instance, approximately 1% of Twitter users are willing to share their location information to the public. In addition, only 1% of Twitter data is publicly accessible through the Twitter application program interface (API) (Morstatter *et al*. 2013). Similar to the cell phone data, the social media data are relatively sparse. Location information is recorded only when they use the media. Therefore, compared to GPS tracking data or travel diary data, which are used to reconstruct the spatial trajectories entirely with great detail, using social media data has to deal with 'hit or miss' situations. People also use social media more likely in certain places and times, such as taking public transportations and in the evening. These using habits add to the temporal and geographical biases of the data.

Despite these undesirable characteristics and serious challenges of using social media data to analyze activity patterns, social media data have several advantages. Instead of recruiting several dozens or maybe the most over hundreds of subjects to record travel diaries, social media data may include a much larger numbers of 'subjects', depending on location (urban versus rural) and geographical coverage (a small town versus a metropolitan region). Studies in social or behavioral research are highly constrained by not just data availability, but also the types of data available, such as 'surface data' that are about many and 'deep data' that are about a few (Manovich 2011). Twitter data provide snapshots of human daily trajectory at a macro scale (more than 500 million registered users publishing 400 million tweets per day in 2013; Morstatter *et al*. 2013), a coverage that could only be dreamt of using the traditional survey approach. Using even a small percentage of these data to study may yield meaningful spatial mobility patterns, but the data are 'thin' in the sense that besides location information, they do not offer much information about the individual's SES background in explaining spatial behavior and patterns. Acquiring these data is relatively inexpensive, especially for the technically savvy researchers.

Whether the sample provided by the Twitter's Streaming API is a sufficient representation of activity on Twitter as a whole depends largely on the coverage and the type of analysis that the researcher would like to perform. Morstatter *et al*. (2013) also found that 'the Streaming API almost returns the complete set of the geo-tagged tweets despite sampling while gathering geo-tagged tweets using the geographic boundary box'. Although the number of geo-tagged tweets is still very small in general (~1%), we 'can be confident that they work with an almost complete sample of Twitter data when geographic boundary boxes are used for data collection' (Morstatter *et al*. 2013, p. 8).

Leveraging different mining techniques, information extracted from tweet contents may be fused with geographical information and profile data to reveal a great deal of information about the users (Xu *et al*. 2013; Longley *et al*. 2015). Studies have shown promising results on detecting users' information (Cheng *et al*. 2010; Chandra *et al*. 2011; Mahmud *et al*. 2012; Huang *et al*. 2014), for instance, to estimate user locations at the city level purely based on tweet content without using any external information, such as gazetteers, IP information, or geo-tags. These approaches are therefore known as 'text based approach' and do not consider the interaction between users. In addition to tweet content, many studies also integrated information from online social profiles (including

addresses) and the spatial behaviors and interactions of users and friends to estimate user locations (Li *et al*. 2012; Mahmud *et al*. 2012). Huang *et al*. (2014) successfully inferred individual daily activity zones to a spatial resolution of city block level. Even higher resolution level is possible when more spatial information (e.g., geo-tagged tweets) is available, together with GIS data and the use of spatiotemporal (ST) clustering methods. The current study infers Twitter user's home and work locations based on this method (Huang *et al*. 2014).

Once the users' home and work locations are inferred, we need to infer the SES (e.g., race, income or age) of individuals by incorporating survey data provided by the U.S. Census Bureau, specifically the American Community Survey (ACS) data that describe the population characteristics within census units. Therefore, we experiment using geo-tagged tweets to study the variation of activity patterns of Twitter users across neighborhoods of different SESs.

3. Hypotheses, data and methodology

Related to the general objective to assess the relationship between activity patterns and SES using Twitter data, our inquiries may be formulated into several hypotheses, covering both substantive questions and methodological issues:

(1) The literature has documented the importance of SES in affecting intra-urban travel behavior, but the specific relationships can be place-dependent. Therefore, the hypothesis is that SES of Twitter users should have some relationships with the sizes of their activity spaces, indicating their mobility levels.
(2) Extending the above hypothesis to activities at the national and international levels, we expect that population of higher SES are relatively more active at these scales and have wider geographical coverage than the lower SES groups.
(3) However, under the umbrellas of the spatial mismatch hypothesis and segregation studies (e.g., Kain 1992; McLafferty and Preston 1996; Preston and McLafferty 1999; Gobillon *et al*. 2007), the disadvantaged population tend to have longer commutes than the more affluent population, although studies have reported inconsistent results occasionally (e.g., Taylor and Ong 1995; Cohn and Fossett 1996; Wyly 1996). We expect that the current study using Twitter data will support the spatial mismatch hypothesis.
(4) Related to the spatial mismatch hypothesis, commuting patterns are also influenced by the geographical context of the neighborhoods, their locations in reference to job centers and the geographical layout of the city (Horton and Reynolds 1971; Weinberg 2000; Downey 2003; Stoll and Covington 2012).

Despite the relatively low and irregular ST resolution of Twitter data, these data can still be used to depict the activity patterns of users to a certain degree (e.g., Huang and Wong 2015). However, to explain the activity patterns of users, these data have to be fused with other data in order to provide richer context for interpretations (e.g., Longley *et al*. 2015). To address these hypotheses, characteristics of individuals and their respective neighborhoods are needed, but these pieces of information are not directly available from Twitter data. Population and socioeconomic data from surveys pertaining to

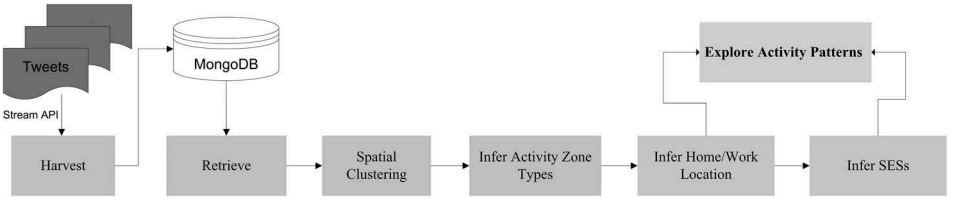

Figure 1. The workflow to retrieve, process and analyze Twitter data to assess the relationships between mobility patterns and socioeconomic status of users.

the neighborhoods are needed to describe the neighborhood environment and to infer the SES of individuals. While using and interpreting these survey data are relatively straightforward, processing and analyzing Twitter data to answer our research questions involve multiple processes and require various analytical techniques. Figure 1 describes the workflow from identifying a sample of Twitter users, inferring their various activity zones, to analyzing their activity patterns.

3.1. *Data collection*

Washington, DC (DC) was chosen as a case for illustrative purposes and three additional reasons. To highlight the SES relationships to activity patterns, the studied population should have sufficient variation in SES status. As population in poor-minority status is often under-represented, this population needs to be 'oversampled'. First, DC has over 50% of African-American (AA) according to the 2010 Census. Thus, it should include enough poor-minority population. Second, different sections in DC have relatively distinctive neighborhood characteristics: the northwest affluent quadrant and the southeast deprived quadrant. Therefore, such clearly fragmented urban landscape facilitates the analysis of neighborhood's roles in affecting the activity patterns of individuals. Third, DC has a relatively young population, expecting a high rate of adopting ICT, including the use of Twitter. Thus, we hoped to be able to identify sufficient Twitter users for our case study by using DC. Only tweets with a geo-tag are used. Based upon the geo-tag locations, we selected Twitter users who were likely working or living in DC (determining if a user lived or worked in the area will be discussed later).

To collect the appropriate Twitter data, we first used the Twitter's streaming API with the bounding box of DC as the geo-tag filter parameter such that we could harvest tweets posted within DC. From the massive tweets collected for 5 months (from January 2014 to March 2014 and from September to November, 2015), we identified 14,066 unique users who used 'Washington, DC' in their profiles. These were very likely DC residents and were our potential subjects.

For the selected users, we used their unique identification numbers to retrieve their tweets in the archive, with a limit of 3200 tweets per user per harvest. Among the 14,066 users, tweets of 3212 users were not available for download because their accounts were 'protected'. Many users sent less than 100 geo-tagged tweets (Figure 2). If a user posted very few geo-tagged tweets (in fact, these are ST points or locations), the data will not be sufficient to reveal the user's activity pattern. While the literature does not provide consistent guidance on the minimum number of locations that is required to identify

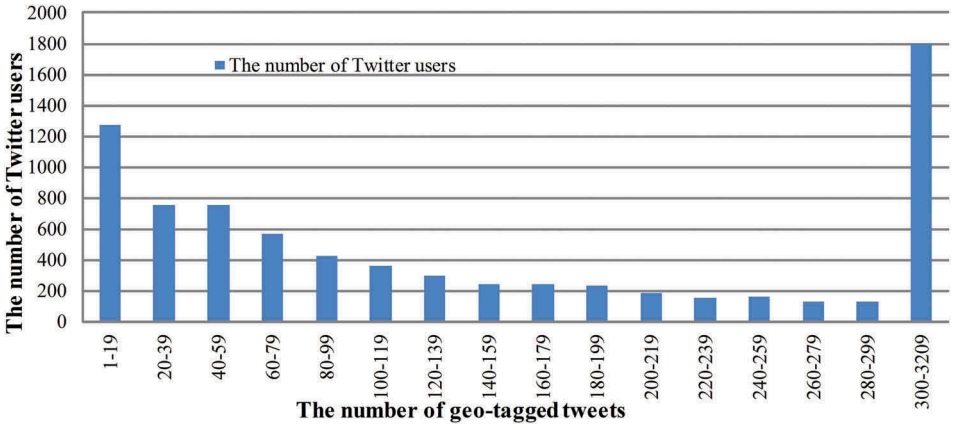

Figure 2. The distribution of Twitter users posting geo-tagged tweets.

regular activity patterns, we chose, 40, a relatively low number of geo-tagged tweets as the cut off to select eligible users. Given that our study needs to determine the locations for regular activities based upon the clusters of frequently visited locations, and assuming that these locations may include home, work, school, church, grocery, etc., using a lower threshold of tweets will include more users, but unlikely reveal activity locations. Using a higher threshold on the number of geo-tagged tweets could identify frequently visited locations with more certainty but with fewer users included. Using 40 as the threshold, 3194 (about 29.4 %) of the remaining 10,854 users were removed, leaving 7660 users in our study. Each of these users posted more than 40 geo-tagged tweets throughout the approximately 2-year period.

3.2. *Inferring home and work sites*

Putting aside the content of the tweets, Twitter data offer mainly ST points of individuals when they tweet, but not why the users were there (i.e., the types of activities that the users were engaged at the respective locations). Although it is possible to mine the tweets to infer the nature of engaged activities, text mining is not the approach we adopted here. We assume that sites that are frequently visited are major activity sites (such as home, work, school, etc.), and be able to identify them provides the basis of activity pattern analysis (Horton and Reynolds 1971). After determining the locations of majority activity sites, the nature of the activities will be inferred. Using Twitter data to determine activity locations needs to be aware of two phenomena. First, users tweet anywhere, not necessarily at the activity sites. Second, people do not follow the exact activity routine every day, and deviations from the norms and ad hoc events may also be captured in Twitter data. Therefore, some tweet locations may not coincide with major activity sites or even within the vicinity of major sites. As they introduce noises to the data depicting the regular activity patterns and sites, they need to be removed.

We used the approach developed by Huang *et al*. (2014) that relies on a spatial clustering method to determine major activity sites. As these major sites are frequently visited, the ST points indicated by tweets are likely clustered around these sites. Using

the density-based spatial clustering of applications with noise (DBSCAN) algorithm to cluster locations reported in the geo-tags of tweets (Huang *et al*. 2014; Huang and Wong 2015), the method produces a set of clusters *R* and these clusters are zones for different types of activities where the user frequently visited (and posted tweets). While some tweet locations are included to form the clusters, dispersed point locations associated with rare events or activities that deviated from the regular patterns are excluded from these clusters or activity zones. The center point of each activity zone can be used to present each zone.

To infer the specific types of activities conducted in the zones, GIS data (e.g., land use image or land planning information) within the vicinity of zone centers are considered. Using St. Louis, Missouri as an example, Huang *et al.* (2014) classify activity zone into nine categories: residential, office, transportation, education, eating, health, shopping, entertainment and service to capture the activities that an individual may engage in an urban setting. In this study (focusing on Washington, DC area), we classified activity zones into a less detailed scheme as our major concerns are about the work and residential locations. As a result, four zone types are used: residential, work space (or office), open space (e.g., green space and river area) and service which includes transportation, eating, health, shopping and entertainment, etc. (Figure 3).

Depending upon the number of tweets and their locations, each individual user is associated with a number of zones, which are assigned to different types based upon the land use characteristics in the vicinity, but multiple zones may be assigned to the same activity type (e.g., residential or work space). We assume that most people have only one home location and one work location. Therefore, for those activities which are likely conducted in a single location, the zone for that activity has to be determined. Several assumptions are adopted to determine which residential zone includes the individual's residence. Within a day, we assume that an individual moves in and out of

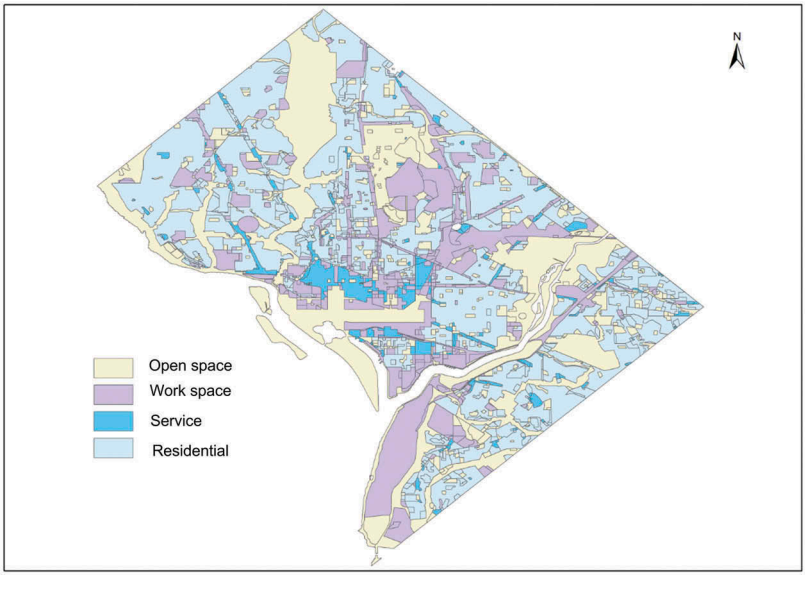

Figure 3. The distribution of activity zones classified from urban land planning map in Washington, DC (http://opendata.dc.gov).

different activity zones, labeling these moves as transitions, and these zones can be of the same type or different types. The transition between the two zones i and j is defined as

$$T_{i,j} = \langle R_i, R_j \rangle,$$

where R_i and R_j are the zones or regions defined by the spatial clustering method. In other words, $T_{i,\ j}$ represents that the user moved from R_i to R_j. We assume that an individual leaves from and returns to home more frequently than from and to other places. Therefore, the residential zone with the largest number of transition should be the home location, while other residential zones associated with this individual could be the homes of friends or relatives. Similarly, among all work zones associated with an individual, the zone with the largest number of transitions is chosen as the office location.

3.3. *Exploring activity patterns*

Many studies have proposed and applied various methods, including geovisualization techniques, to study activity locations, which are often represented by ST points (e.g., Delmelle *et al*. 2013; Beecham *et al*. 2014). Computational movement analysis also suggests many techniques to analyze spatial trajectories (e.g., Buchin *et al*. 2010; Laube and Purves 2006, 2014). Unfortunately, most of these methods may not be applicable to social media data, which capture locations in irregular ST intervals and sometime sparsely.

A variety of tools have been developed to study activity space (e.g., Chaix *et al*. 2012). A standard implementation of the activity space concept is to use ellipses (e.g., Schönfelder and Axhausen 2003; Järv *et al*. 2014). When data density was too sparse to determine ellipses reliably, Jones and Pebley (2014) used minimum convex polygons to define activity spaces. Using density kernels has also been suggested and applied (e.g., Wang *et al*. 2012). On the other hand, the space–time (S–T) path approach explicitly describes and visualizes the temporal dimension in the 2D geographical space environment, and many methods have been proposed to compare S–T paths (e.g., Ren and Kwan 2007; Shaw *et al*. 2008; Chen *et al*. 2011). These methods require clearly trajectories of individuals, but Twitter data we used are temporally too sparse to construct representative S–T paths with reasonable confidence levels (Huang and Wong 2015). Instead, the following set of measures is proposed. Although these measures are not exactly the same as those used in previous studies (e.g., Hanson 1982; Wang *et al*. 2012), they are conceptually similar, but are modified to suit the current data situation.

3.3.1. *Number of the activity zones (representative locations)*

After processing the tweet data, different activity zones were identified and types of activities associated with these zones are also determined. To compare and differentiate various activity patterns, the variety of locations that a user visited regularly can be considered. The variety can simply be translated into the number of activity zones associated with each individual. Would be of interest is the variation of activity diversity among individuals across different SESs or neighborhoods.

3.3.2. *Distance between home and activity zones*

The distances from home to all activity zones associated with an individual can indicate the geographical extent of one's activity space. Various statistics based upon these zonal distances for each individual and population group can be computed. In addition, zonal distances between home and specific types of zones can be compared.

3.3.3. *Standard deviational ellipse (SDE)*

SDE is a centrographic measure to summarize the spatial distribution of a set of point locations. Besides providing the overall location of the set of points by the center of the ellipse, the ellipse also describes the spread of the point locations along the two orthogonal axes of minimum and maximum spreads, which are the lengths of the two axes (Wong and Lee 2005). Through the angle of rotation, the ellipse also indicates if the set of locations exhibits any specific orientation or directional bias.

SDE has been used to summarize activity space in mobility analyses (e.g., Järv *et al*. 2014). Activity spaces depicted by ellipses can be compared by size (the area of an ellipse) and shape (deviations along the two axes). In general, bigger ellipses indicate larger spatial extents of the activity spaces. Different ellipses can be derived from different types of point locations. Among all tweet locations, some represent ad hoc events that were not part of the regular activity patterns. A subset of all tweet locations was used to form the activity zones representing the regular activity spaces based upon the spatial clustering results. Therefore, the SDE for the following sets of ST points were computed:

(1) Overall activity space (OAS): SDEs derived from all ST points.
(2) Regular activity space (RAS): SDEs derived only from ST points forming those activity zones, excluding those ST points not being a member of any cluster.

3.4. *Visualizing activity patterns*

Much effort in the past decade has been devoted to exploit computing and GIS technologies supporting interactive 2D or 3D geo-visualization of movement data (Kwan 2000; Shen *et al*. 2013). Using travel diary and GPS tracking data, 3D rendering of S–T paths and other visualization methods have been used to explore human activity patterns (Kwan 2000; Buliung and Kanaroglou 2006a; Shen *et al*. 2013). These techniques mostly rely on relatively dense movement data, incompatible to the Twitter data with sparse ST points of irregular resolutions. Using clustering method, we were able to identify significant or representative activity zones. Centers of these activity zones were derived for each Twitter user as the activity locations.

To show the mobility patterns of users, we adopt the flight route map design to show the spatial relationship between home locations (communities) and activity zones, locations that users frequently tweeted. These maps are simple but effective to reveal differences between communities/population groups in terms of the locations of their activity zones and distances of travel. They can be compiled for activity zones at different geographical scales.

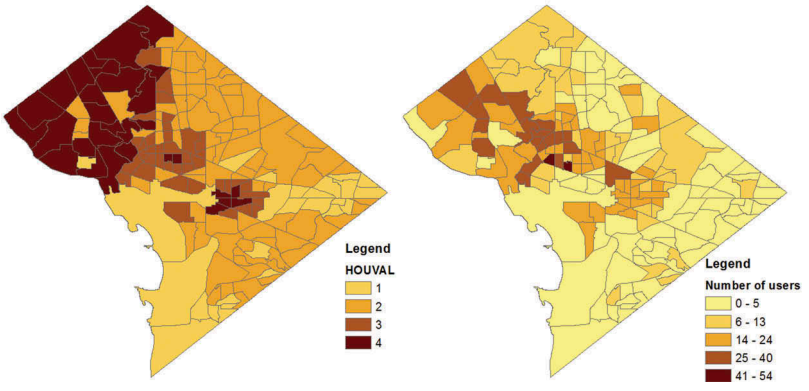

Figure 4. Distribution of social groups classified based on median house value (left) and the number of twitter users in each tract (right).

4. Demonstration

In this demonstration, we classified census tracts in DC into four social groups based on the median house value[1] reported in the 2009–2013 5-year ACS (Figure 4, left). These groups were created using the mean of the median house value (MHV) minus one standard deviation, the mean and mean plus one standard deviation as the class breaks. While tracts with MHV less than 30, 730 are considered as poor communities (Group_1 in Figure 4, left), tracts with MHV more than 110, 456 are considered as rich communities (Group_4). Rich communities are concentrated in northwest DC while the poor communities are mostly in the southeast (Figure 4, left).

We selected four different communities in DC for detailed analysis later (Figure 5). These communities include two relatively poor neighborhoods in the southeast (seven

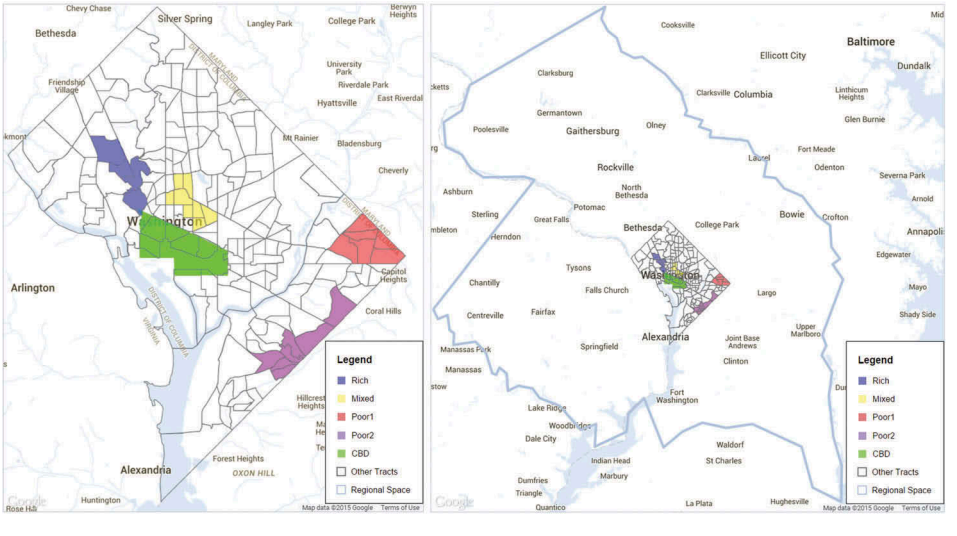

Figure 5. Distribution of selected communities and boundary of regional space at different scale; regional space is defined as the Washington metropolitan area centered around DC, and surrounded by two counties in Maryland (MD; Montgomery, Prince George's) and five counties (and independent cities) in Virginia (VA; Arlington, Fairfax, Falls Church, Fairfax City, Alexandria).

census tracts) and east (six tracts), a relatively rich neighborhood (four tracts) in the northwest, and a mixed neighborhood (eight tracts) around the center of the city. The median house value and median household income of these tracts are shown in Figure 6. ACS also reports the 90% margin of error (MOE) of each estimate. The confidence bounds of these estimates (i.e., estimates ± the MOEs) are plotted in the figures. If the confidence bounds of two estimates do not overlap, the two estimates are statistically different.[2] The figures show that according to the two economic variables, majority of the tracts are statistically different between the poor and rich communities at the 90% confidence level. While the income variable cannot distinguish the mixed community with the other three communities clearly, the house value variable can reasonable discriminate among the three types of communities. However, in terms of racial composition, the average percentages of African–American in the rich, the mixed and two poor communities were 3.79%, 45.72%, 95.42% and 95.71%, respectively, and these

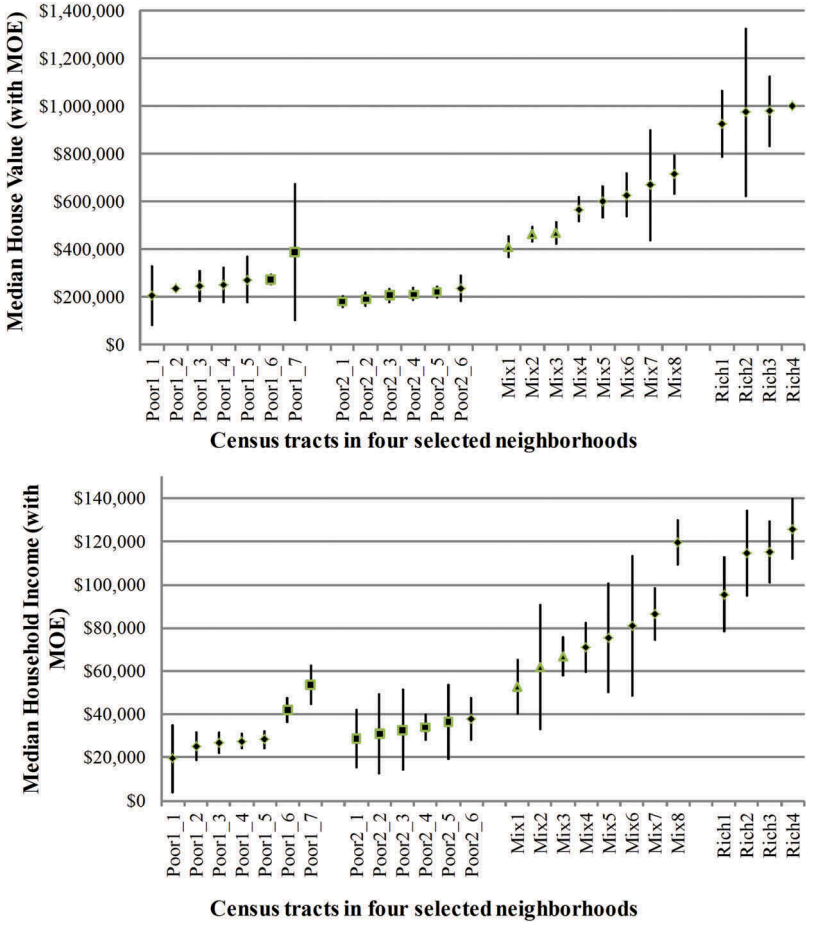

Figure 6. Estimates of median house value (upper) and median household income (lower) of census tracts in the three Washington, DC communities, using 2009–2013 American Community Survey data. The confidence bounds were constructed using the 90% margin of error reported in ACS.

percentages among the three types of community are all different statistically at 95% confidence level.

Using ACS data to infer the SES of individual users has some limitations. Similar to typical census data, ACS data are ecological or aggregated data. Inferring the SES of individual users from ACS data likely commits ecological fallacy. To minimize the effects of ecological fallacy, ACS data of smaller areal units may be used. While census-tract level ACS data are reasonably reliable, many block group level data have much large error, making the use of smaller unit data highly undesirable. Therefore, ACS data cannot help improve the 'spatial granularity or demographic breakdown'.

4.1. *Activity zones detection*

Each geo-tagged tweet has a pair of coordinates (latitude and longitude) revealing the specific location where it was posted. We applied spatial clustering to determine the locations from where an individual frequently tweeted. We choose DBSCAN as our clustering algorithm, which needs to specify the radius of a cluster (*eps*) and the minimum number of points (*MinPts*) for a cluster as inputs. Similar to Zhou *et al*.'s (2004) study, we set the *eps* value to 20 meters, which approximates the uncertainty in GPS readings. Ester *et al*. (1996) show that using a *MinPts* value smaller than four may misclassify random points as clusters while a *MinPts* value of four or larger will unlikely produce clusters of varying results. Recommendation offered by Ester *et al*. (1996) is likely data dependent (we did test that the number of clusters are different using different *MinPts* values). Using a relatively small value of four, the result should be rather conservative, likely including locations that are not part of the regular activity patterns. However, using a larger *MinPts* value is not necessary in this step of the analysis due to the use of additional criteria of selecting representative clusters in later steps, which will be discussed below. Therefore, we use a *MinPts* value of four. We also tested the sensitivity of using different *eps* values, but we did not find significant differences in our results.

To reduce the very large number of activity zones (or clusters) and the transitions extracted between places, less frequently visited locations need to be removed to be visualized without cluttering. Therefore, a cluster with ST points less than a threshold *l* (e.g., 20) may be considered as noise, and was therefore discarded. However, some users may only have a small number of geo-tagged tweets but are highly clustered. Given their location consistency, they could be used to derive the represented clusters. If the cluster has less points than the point threshold *l*, but the proportion of ST points in the cluster over all ST points is larger than a threshold *k* (5% in this study), the cluster can be considered as a representative cluster as well. As mentioned earlier, 7660 qualified users active in DC posted more than 40 geo-tagged tweets. In addition, when a user has only one detected cluster, this single activity zone cannot offer meaningful activity patterns, and therefore these users (647 in total) were discarded. Among the remaining users (7013), at least one representative activity zone can be determined for 5216 of them.

We then applied the home inference model to derive their home locations from the representative locations. Home locations of 2014 users were derived in all DC tracts (Figure 4 right). Among them, 323 users were located in the four selected communities (Figure 5) with 21, 27, 170 and 95 users in poor1, poor2, mixed and rich zones,

Table 1. Distribution of the representative activity zone numbers and the zone number with at least 5% of the entire ST points or more than 20 tweets for the selected users with home location detected in different socioeconomic communities.

Community	Total number of users	Average number of retrieved tweets	Average number of retrieved geo-tagged tweets	AZone[a]	BZone[b]
Group_1	70	2992	492	12.2	3.69
Group_2	648	2633	363	9.76	2.88
Group_3	788	2499	335	9.79	2.85
Group_4	508	2315	317	9.19	2.85

[a]AZone indicate the average count of all activity zones.
[b]BZone indicates the average count of representative activity zones with more than 5% of the entire ST points or 20 points.

respectively (Table 1). Noted that the two poor communities (poor1 and poor2) were members of Group_1 and Group_2, the mixed community was a member of Group 2 and Group 3, and the rich community belongs to Groups 4. Quite interesting is that many fewer users were detected in the poor communities (Figure 4 (right) and Table 1). Witte and Mannon (2010) claimed that marginalized populations often lack of or have limited internet access in their households, making access to Twitter a socially stratified practice. Livingstone and Helsper (2007) found that inequalities by age, gender and socioeconomic status are related to the quality of access to and the use of the internet. Our result confirms previous findings that socioeconomic differences contribute to the disparity in social media usage.

However, more general clusters and representative clusters were detected for the users in poor communities and these users on average tweet more and were more likely to include the geo-tags in their messages. On the other hand, relatively less geo-tagged tweets were from the rich communities.

4.2. *Variations of trip distances*

To compare the collective movement behaviors of users from different communities, we focus on the distances from home to different types of activity zones. We classified destinations into work, regional, national, and international, and associated travel distances are calculated as below.

- Work destinations: The distance from home to work destinations of all users. From the 2014 users with home locations, we further applied the office location detection procedure, and 319 users were detected to have at least one work zone (Table 2). Note that our program can only detect locations inside DC. Users working outside of DC may be discarded. As we stated earlier, among the clusters detected as work zones, the one with the largest number of transitions were assigned as the office location.
- Non-work regional destinations: The average travel distance from home to all non-work activity zones within the regional space (Figure 5, right). The study would be more informative by evaluating activity patterns beyond work place and residences. Unfortunately, data for our current study could not support the analysis of activity patterns beyond the home and work locations.

Table 2. Average travel distances of the users in different communities for different types of destinations.

Types of destinations	Work		Non-work regional		National		International	
	AveDis (km)	UserNum	AveDis (km)	UserNum	AveDis (km)	UserNum	AveDis (km)	UserNum
Group_1	4.85	10	7.13	66	578.55	34	3864.49	3
Group_2	3.37	107	5.27	589	937.36	354	6390.73	95
Group_3	2.45	122	3.95	718	1117.14	466	6454.29	137
Group_4	3.31	80	5.07	463	1100.80	309	6536.71	91

Note: AveDis: averaged distance in km; UserNum: Number of users.

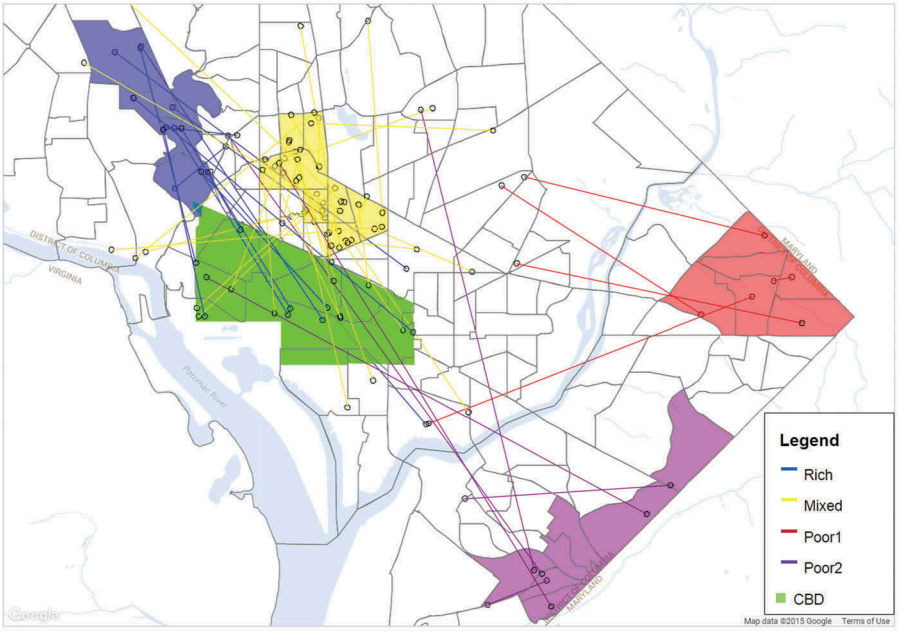

Figure 7. Work destinations of users in the four selected communities.

- National destinations: The average travel distance from home to activity zones in the United States, but outside the regional space.
- International destination: The average travel distance from home to activity zones outside of the United States.

Table 2 shows the average distances to different types of destination across different communities. While we expected that SES should correlate with travel distances, results indicate that the poorest population needed to travel (both for work and regional destinations) the longest distance, the mid-income people or the people in the mixed neighborhoods (Group 3) have the shortest travel distances, not the richest population (Group 4). This finding is contrary to the literature which put the poor and rich at the two ends of the spectrum in terms of travel distances. To explain this unusual result, we have to investigate the urban structure of DC. Figure 7 shows the work destinations of users in the four selected neighborhoods. Most of the destinations are around the center of DC, part of which was designated by the city government as the central business

district where most of the jobs are located.[3] As the poor communities (poor1 and poor2) were relatively far away from the city center, residents in these communities have to travel relatively long distances to work. On the other hand, the mixed community was closest to the city center, resulting in the shortest work distances.

We also calculated the *t*-statistics to test the differences between these averaged distances across the four types of communities (Table 3). Most comparisons for work and non-work destinations are significant at 90% or higher. One exception is the comparisons between Groups 2 and 4 for non-work trip. The travel distances for non-work trips for populations in the moderately poor or mixed communities and rich communities were not that different, partly because both groups were located somewhat along the periphery of the city (Figure 8). While the indifference between Groups 1 and 2 for work trips is not surprising, the most interesting result is the indifference in work trip distances between Groups 1 and 4 and between Groups 2 and 4. Work trip distances between the rich and poor or moderately poor communities were not significantly different. For travel at the national scale, all group comparisons yield significant differences, except between Groups 3 and 4, the mixed and rich communities. At the international level, population in the poor communities was really different from all other groups. These results indicate that relationships between SES and travel distances are quite complicated. They do not exhibit simple correlations (such as low SES with short distances and high SES with long distances) and their relationships vary by types of travel (Tables 2 and 3).

Table 2 shows that users in the poor communities (Group 1) had relatively longer regional travel distances than those in the rich (Group 4) and mixed (Groups 2 and 3) communities (and the differences are significant – Table 3). Again, such patterns may be governed by the city spatial structure. In addition, user's social network can play a significant role, as these destinations can be locations of extended family members and friends. Unfortunately, without mining the tweets, we cannot ascertain the purposes of these trips and future research may include such effort.

For destinations at the national scale, users of the Group_3 had the longest distance, closely followed by Group 4 (Table 2). Based on the map for the selected communities (Figure 9), many destinations for users of the rich communities were far from DC in the west coast of the United States, and in the south. Users of the mixed community have diverse national destinations as well. However, destinations of users in the poor communities were mostly in the east coast of the United States. Only a few from this group had destinations relatively far from DC.

For the international travel, difference in patterns between the rich and poor communities is quite distinguishable: relatively large proportions of users from Groups 3 and 4 had international destinations, while only a small proportion of users (3 out of 70) in

Table 3. Comparing averaged distances to different types of destinations between different communities.

Group numbers	Non-work	Work	National	International
1 vs. 2	**	#	**	**
1 vs. 3	**	**	**	**
1 vs. 4	**	#	**	**
2 vs. 3	**	**	**	#
2 vs. 4	#	#	**	#
3 vs. 4	**	*	#	#

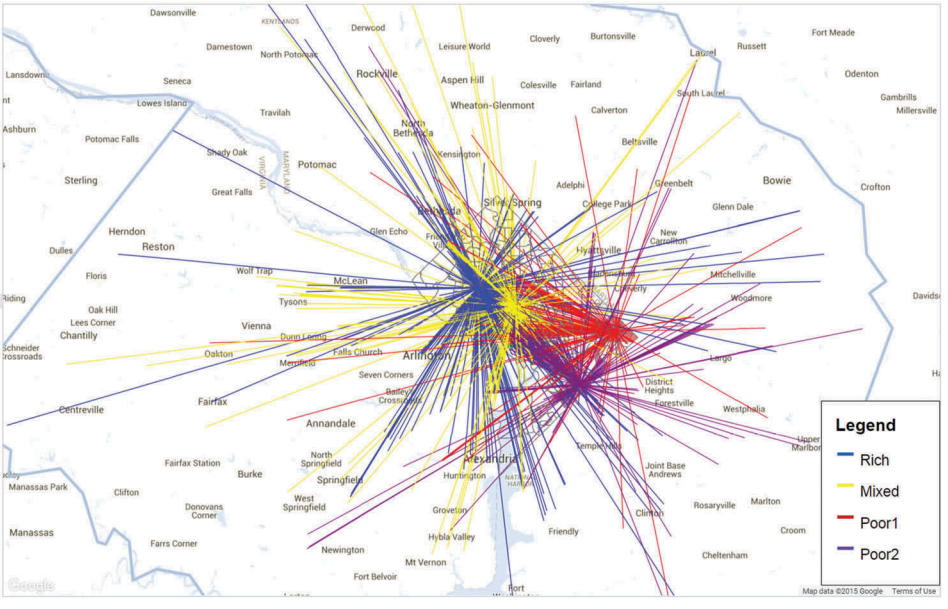

Figure 8. Regional 'routes' from different communities.

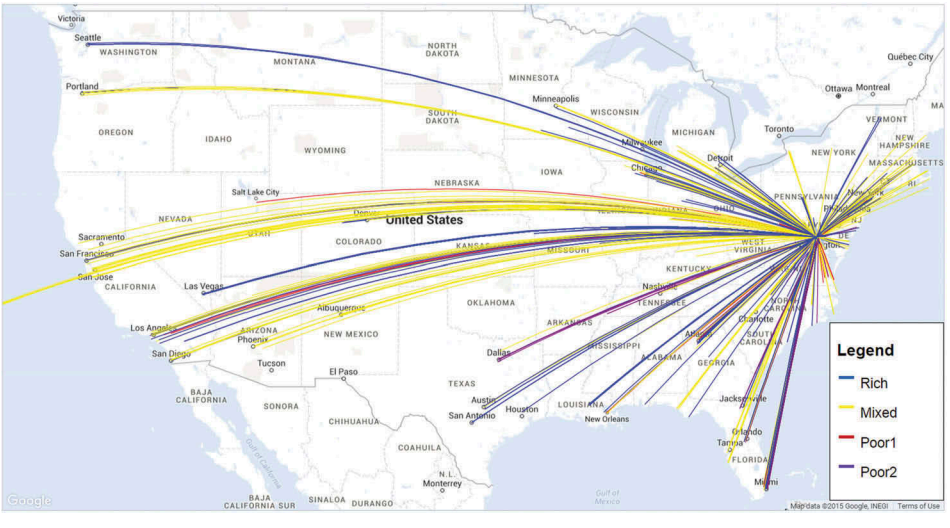

Figure 9. National 'routes' from different communities.

Group 1 had an international destination. The route map (Figure 10) shows that the destinations of users from the selected rich community were found in different continents, including Asian, Europe and Africa. Destinations of users from the mixed community were mostly in Europe, South America and Canada.

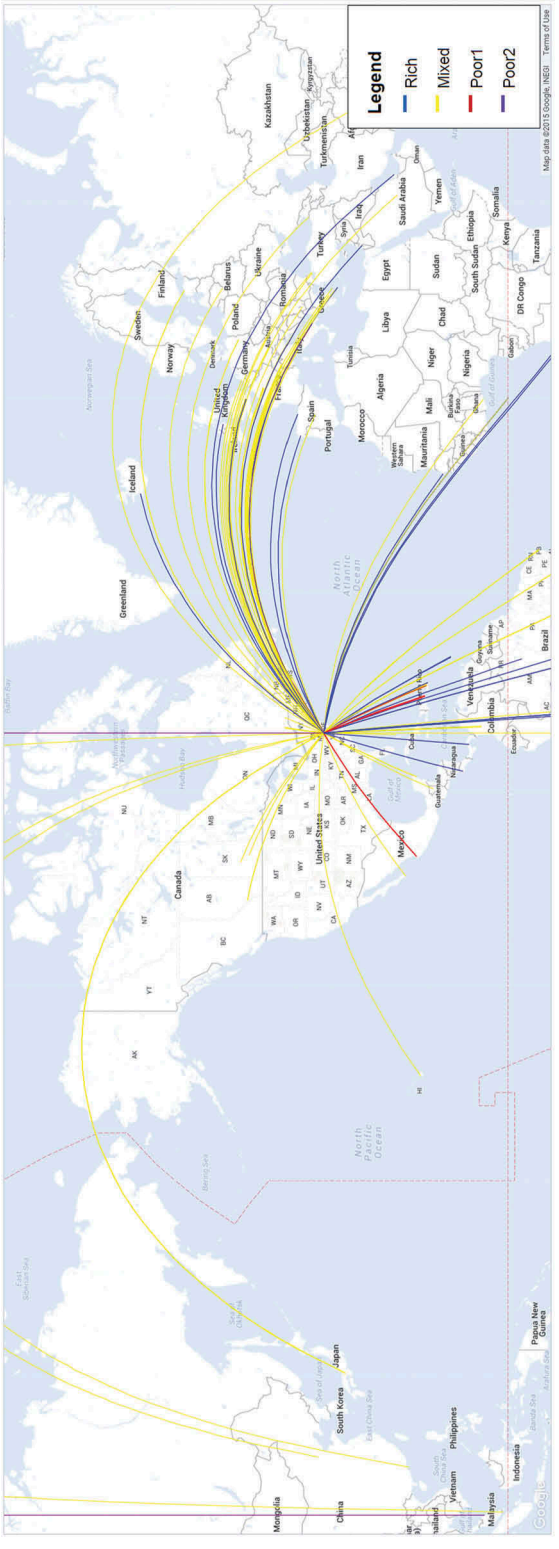

Figure 10. International 'routes' from different communities.

4.3. *Standard deviational ellipse*

As stated earlier, two types of SDE-based measures, OAS and RAS, are designed to explore the spatial extent of individual activity spaces. To calculate OAS, all ST points within the regional space are used. To derive RAS, only points in the representative zones of the regional space are used.

Table 4 shows the average shape and size of OAS and RAS for users in different communities. Users in Group 1 have the largest overall activity spaces, followed by those in the Group_2, Group_4, and then the Group_3. The results that users of poor communities had larger activity spaces than the other groups seem to be at odds with the general expectation that poor people are less mobile. But given that these activity spaces also consider work locations, and users of the poor communities had relatively long work commutes, these results are reasonably consistent with other findings reported in this article. Also expected is that the sizes of RAS are smaller than those for OAS, given that RAS consider only a subset of locations constituting OAS. If the two axes have very different lengths, it is strong indication that the ellipse has a strong directional bias. Table 4 shows that the proportional differences between the two axes for the RAS are much larger than the corresponding values for OAS. Therefore, ellipses for RAS are very narrow or elongated, reflecting the clear directional bias of the corresponding destinations.

5. Conclusion and discussion

Many geographical studies have exploited social media data, including Twitter. As Steiger *et al*. (2015) pointed out, very few studies, if any, are related to urban planning and management (p. 19). One of the main reasons that social media data do not offer a great deal of support for socioeconomic analysis is the lack of demographic and SESs information of the users. The SES of the social media users may be inferred when data are fused with other survey data (e.g., Longley *et al*. 2015). This article demonstrates how social media data can be fused with residential-based data (ACS, the primary source of socioeconomic data in the United States.) to support socioeconomic research geographically, filling a major gap in the literature as pointed out by Steiger *et al*. (2015, p. 21). Thus, the activity patterns of Twitter users were analyzed in respect to their SESs and the geographical settings of their neighborhoods. This general approach may be used to analyze other social media data to answer social-geographical questions, while the risk of committing ecological fallacy should be aware.

Despite various limitations, such as their sparseness spatiotemporally, of using social media data like Twitter for geographical research, this study shows a high degree of success in using Twitter data for activity pattern analysis. Our methodology on deriving and mining meaningful travel patterns from sparse and long-term ST point data collected through social media provides a blue print for geographers and spatial social scientists to leverage these online ST trajectory data.

Besides methodological contributions, our study also shed lights on the spatial mismatch hypothesis. Previous studies claimed that the economic disadvantaged people have to travel longer distance as works are located far from their residences while this mismatch is a lesser problem for the more affluent population. Our study

Table 4. Average overall activity space (OAS) and regular activity space (RAS) of the users in different socioeconomic groups.

	OAS					RAS				
Community	UserNum	AvgSDEx (km)	AvgSDEy (km)	AvgSDEx + AvGSDEy (km)	Avgsize (km^2)	UserNum	AvgSDEx (km)	AvgSDEy (km)	AvgSDEx+ AvgSDEy (km)	Avgsize (km^2)
Group_1	70	2.36	4.23	6.59	39.58	60	0.83	2.83	3.66	13.13
Group _2	648	2.04	3.31	5.36	27.75	548	0.43	1.82	2.25	5.52
Group _3	788	1.75	3.01	4.76	21.42	651	0.37	1.62	1.98	4.6
Group _4	508	1.9	3.15	5.05	23.43	445	0.36	1.67	2.03	3.59

Note: UserNum: number of users; AvgSDEx: averaged deviation along the x-axis in km; AvgSDEy: averaged deviation along the y-axis in km; AvgSDEx + AvgSDEy: the sum of the two deviations, Avgsize: the averaged area/size of the ellipse in km^2.

shows that in the case of DC, the urban spatial structure of the city is the key issue. Although the poor has to travel the longest distance in general, the mid-income group may have the shortest travel, not the most affluent group. Thus, from a policy perspective, improving the efficient and quality of public transportation services to the poorer communities should be the priority and such improvement can reduce social disparities.

In the context of international travel, results are clear that the more affluent residents were more internationally oriented than the economically disadvantaged residents. The two groups were also quite different in their international orientations. The clear destination biases at the regional scale between the two groups may only be partly explained by the differences in economic status (e.g., Hong *et al*. 1996) as the racial–ethnic backgrounds of individuals could be a significant factor (e.g., Philipp 1993, 1994). Future effort should focus on determining the racial–ethnic identities of social media users in order to address issues related to race–ethnicity.

Notes

1. We also considered using median household income to delineate neighborhoods, but using median house value produce more geographically coherent neighborhoods than using income.
2. Note that the opposite is not true, i.e., estimates are not statistically different if their confidence bounds overlap although many people assume that this is the case.
3. http://ddot.dc.gov/sites/default/files/dc/sites/ddot/publication/attachments/dc_central_business_district_bikes_0.pdf

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Qunying Huang http://orcid.org/0000-0003-3499-7294
David W. S. Wong http://orcid.org/0000-0002-0525-0071

References

Beecham, R., Wood, J., and Bowerman, A., 2014. Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47, 5–15. doi:10.1016/j.compenvurbsys.2013.10.007

Buchin, K., Buchin, M., and Gudmundsson, J., 2010. Constrained free space diagrams: a tool for trajectory analysis. *International Journal of Geographical Information Science*, 24 (7), 1101–1125. doi:10.1080/13658810903569598

Buchin, M., Dodge, S., and Speckmann, B., 2014. Similarity of trajectories taking into account geographic context. *Journal of Spatial Information Science*, 9, 101–124.

Buliung, R.N. and Kanaroglou, P.S., 2006a. A GIS toolkit for exploring geographies of household activity/travel behavior. *Journal of Transport Geography*, 14 (1), 35–51. doi:10.1016/j.jtrangeo.2004.10.008

Buliung, R.N. and Kanaroglou, P.S., 2006b. Urban form and household activity-travel behavior. *Growth and Change*, 37 (2), 172–199. doi:10.1111/grow.2006.37.issue-2

Chaix, B., *et al*., 2012. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. *American Journal of Preventive Medicine*, 43 (4), 440–450. doi:10.1016/j.amepre.2012.06.026

Chandra, S., Khan, L., and Muhaya, F.B., 2011. Estimating twitter user location using social interactions–a content based approach. *In*: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011*, Boston, MA. IEEE, 838–843.

Chen, J., *et al*., 2011. Exploratory data analysis of activity diary data: a space–time GIS approach. *Journal of Transport Geography*, 19, 394–404. doi:10.1016/j.jtrangeo.2010.11.002

Cheng, Z., Caverlee, J., and Lee, K., 2010. You are where you tweet: a content-based approach to geo-locating twitter users. *In*: *Proceedings of the 19th ACM international conference on Information and knowledge management*, Toronto, ON. ACM, 759–768.

Cohn, S. and Fossett, M., 1996. What spatial mismatch? The proximity of blacks to employment in Boston and Houston. *Social Forces*, 75 (2), 557–573. doi:10.1093/sf/75.2.557

Davies, W.K.D. and Murdie, R.A., 1993. Measuring the social ecology of cities. *In*: L.S. Bourne and D. F. Ley, eds. *The changing social geography of Canadian cities*. Montreal, Quebec: McGill-Queen's University Press, 52–75.

Delmelle, E., *et al*., 2013. Methods for space-time analysis and modeling: an overview. *International Journal of Applied Geospatial Research*, 4 (4), 1–18. doi:10.4018/IJAGR

Downey, L., 2003. Spatial measurement, geography and urban racial inequality. *Social Forces*, 81 (3), 937–952. doi:10.1353/sof.2003.0031

Echenique, F. and Fryer Jr., R.G., 2007. A measure of segregation based on social interactions. *The Quarterly Journal of Economics*, 122 (2), 441–485. doi:10.1162/qjec.122.2.441

Ester, M., *et al*., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In: the second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 2-4 August Portland, Oregon. California: AAAI Press, 226–231.

Ewing, R. and Cervero, R., 2001. Travel and the built environment: A synthesis. *Transportation Research Record: Journal of the Transportation Research Board*, 1780, 87–114. doi:10.3141/1780-10

Farber, S., Páez, A., and Morency, C., 2012. Activity spaces and the measurement of clustering and exposure: A case study of linguistic groups in Montreal. *Environment and Planning A*, 44 (2), 315–332. doi:10.1068/a44203

Gobillon, L., Selod, H., and Zenou, Y., 2007. The mechanisms of spatial mismatch. *Urban Studies*, 44 (12), 2401–2428. doi:10.1080/00420980701540937

Hanson, S., 1982. The determinants of daily travel-activity patterns: relative location and socio-demographic factors. *Urban Geography*, 3 (3), 179–202. doi:10.2747/0272-3638.3.3.179

Hong, G.-S., Morrison, A.M., and Cai, L.A., 1996. Household expenditure patterns for tourism products and services. *Journal of Travel & Tourism Marketing*, 4 (4), 15–40. doi:10.1300/J073v04n04_02

Horton, F.E. and Reynolds, D.R., 1971. Effects of urban spatial structure on individual behavior. *Economic Geography*, 47 (1), 36–48. doi:10.2307/143224

Huang, Q., Cao, G., and Wang, C., 2014. From where do tweets originate? - A GIS approach for user location inference. *In*: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '14)*, 4–7 November, Dallas, TX, USA. New York, NY: ACM, 1–8.

Huang, Q. and Wong, D., 2015. Modeling and visualizing regular human mobility patterns with uncertainty: an example using Twitter data. *Annals of the Association of American Geographers*, 105, 1179–1197. *forthcoming*. doi:10.1080/00045608.2015.1081120

Järv, O., *et al*., 2014. Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Studies*. doi:10.1177/0042098014550459

Johnston, R.J., 1976. Residential area characteristics: research methods for identifying urban sub-areas – social area analysis and factorial ecology. *In*: D.T. Herbert and R.J. Johnston, eds. *Social areas in cities*. Chichester: Wiley, 193–235.

Jones, M. and Pebley, A.R., 2014. Redefining neighborhoods using common destinations: social characteristics of activity spaces and home census tracts compared. *Demography*, 51, 727–752. doi:10.1007/s13524-014-0283-z

Kain, J.F., 1992. The spatial mismatch hypothesis: three decades later. *Housing Policy Debate*, 3 (2), 371–460. doi:10.1080/10511482.1992.9521100

Kwan, M.-P., 1998. Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. *Geographical Analysis*, 30 (3), 191–216.

Kwan, M.-P., 2000. Interactive geovisualization of activity travel patterns using three-dimensional geographical information systems: a methodological exploration with a large dataset. *Transportation Research Part C: Emerging Technologies*, 8, 185–203. doi:10.1016/S0968-090X(00)00017-6

Kwan, M.-P., 2008. From oral histories to visual narratives: re-presenting the post-September 11 experiences of the Muslim women in the USA. *Social and Cultural Geography*, 9 (6), 653–669.

Kwan, M.-P., 2013. Beyond space (as we knew it): toward temporally integrated geographies of segregation, health, and accessibility. *Annals of the Association of American Geographers*, 103 (5), 1078–1086. doi:10.1080/00045608.2013.792177

Laube, P. and Purves, R.S., 2006. An approach to evaluating motion pattern detection techniques in spatio-temporal data. *Computers, Environment and Urban Systems*, 30 (3), 347–374. doi:10.1016/j.compenvurbsys.2005.09.001

Li, R., *et al*., 2012. Towards social user profiling: unified and discriminative influence model for inferring home locations. *In*: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing. ACM, 1023–1031.

Livingstone, S. and Helsper, E., 2007. Gradations in digital inclusion: children, young people and the digital divide. *New Media & Society*, 9 (4), 671–696. doi:10.1177/1461444807080335

Longley, P.A., Adnan, M., and Lansley, G., 2015. The geotemporal demographics of Twitter usage. *Environment and Planning A*, 47 (2), 465–484. doi:10.1068/a130122p

MacPherson, D. W., and Gushulak, B. D., 2001. Human mobility and population health: new approaches in a globalizing world. *Perspectives in Biology and Medicine*, 44 (3), 390–401.

Mahmud, J., Nichols, J., and Drews, C., 2012. Where is this tweet from? Inferring home locations of Twitter users. *ICWSM*, 12, 511–514.

Manovich, L., 2011. Trending: the promises and the challenges of big social data. *Debates in the Digital Humanities*, 2 (2011), 460–475.

Matthews, S. A. and Yang, T.-C., 2013. Spatial polygamy and contextual exposures (SPACEs): promoting activity space approaches in research on place and health. *American Behavioral Scientist*, 57 (8), 1057–1081.

Massey, D.S. and Denton, N.A., 1993. *The American apartheid: Segregation and the making of the underclass*. Cambridge, MA: Harvard University Press.

Matthews, S.A., 2011. Spatial polygamy and the heterogeneity of place: studying people and place via egocentric methods. *In*: L.M. Burton, *et al*., eds. *Communities, neighborhoods, and health: expanding the boundaries of place*. New York, NY: Springer, 35–55.

McLafferty, S. and Preston, V., 1996. Spatial mismatch and employment in a decade of restructuring. *The Professional Geographer*, 48 (4), 420–431. doi:10.1111/j.0033-0124.1996.00420.x

Mislove, A., *et al*., 2011. Understanding the demographics of Twitter users. *In*: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 17–21 July Barcelona, Catalonia, Spain. Menlo Park, CA: AAAI Press, 554–557. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

Morstatter, F., *et al*., 2013. Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose. *arXiv preprint* arXiv:1306.5204.

National Research Council, 2002. *Community and quality of life: data needs for informed decision making*. Washington, DC: National Academy Press.

Palmer, J.R.B., *et al*., 2013. New approaches to human mobility: using mobile phones for demographic research. *Demography*, 50, 1105–1128. doi:10.1007/s13524-012-0175-z

Pas, E.I., 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environment and Planning A*, 16 (5), 571–581. doi:10.1068/a160571

Pettigrew, T.F., 2008. Future directions for intergroup contact theory and research. *International Journal of Intercultural Relations*, 32, 187–199. doi:10.1016/j.ijintrel.2007.12.002

Pettigrew, T.F. and Tropp, L.R., 2008. How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, 38, 922–934. doi:10.1002/ejsp.v38:6

Philipp, S.F., 1993. Racial differences in the perceived attractiveness of tourism destinations, interests, and cultural resources. *Journal of Leisure Research*, 25 (3), 290–304.

Philipp, S.F., 1994. Race and tourism choice: A legacy of discrimination? *Annals of Tourism Research*, 21 (3), 479–488. doi:10.1016/0160-7383(94)90115-5

Preston, V. and McLafferty, S., 1999. Spatial mismatch research in the 1990s: progress and potential. *Papers in Regional Science*, 78, 387–402. doi:10.1007/s101100050033

Ren, F. and Kwan, M.-P., 2007. Geovisualization of human hybrid activity-travel patterns. *Transactions in GIS*, 11 (5), 721–744. doi:10.1111/tgis.2007.11.issue-5

Schönfelder, S. and Axhausen, K.W., 2003. Activity spaces: measures of social exclusion? *Transport Policy*, 10, 273–286. doi:10.1016/j.tranpol.2003.07.002

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.

Shaw, S.-L., Yu, H., and Bombom, L.S., 2008. A space–time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS*, 12 (4), 425–441. doi:10.1111/tgis.2008.12.issue-4

Sheller, M. and Urry, J., 2006. The new mobilities paradigm. *Environment and Planning* A, 38 (2), 207–226.

Shen, Y., Kwan, M.-P., and Chai., Y., 2013. Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. *Journal of Transport Geography*, 32, 1–11. doi:10.1016/j.jtrangeo.2013.07.007

Silm, S. and Ahas, R., 2014. Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers*, 104 (3), 542–559. doi:10.1080/00045608.2014.892362

Stefanidis, A., Crooks, A., and Radzikowski, J., 2013. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78 (2), 319–338. doi:10.1007/s10708-011-9438-2

Steiger, E., De Albuquerque, J.P., and Zipf, A., 2015. An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Transactions in GIS*, 19, 809–834. doi:10.1111/tgis.12132

Stoll, M.A. and Covington, K., 2012. Explaining racial/ethnic gaps in spatial mismatch in the US: the primacy of racial segregation. *Urban Studies*, 49 (11), 2501–2521. doi:10.1177/0042098011427180

Taylor, B.D. and Ong, P.M., 1995. Spatial mismatch or automobile mismatch? an examination of race, residence and commuting in US metropolitan areas. *Urban Studies*, 32 (9), 1453–1473. doi:10.1080/00420989550012348

Tsou, M.-H., *et al*., 2014. Mapping ideas from cyberspace to Realspace: visualizing the spatial context of keywords from web page search results. *International Journal of Digital Earth*, 7 (4), 316–335. doi:10.1080/17538947.2013.781240

Wang, D., Li, F., and Chai, Y., 2012. Activity spaces and sociospatial segregation in Beijing. *Urban Geography*, 33 (2), 256–277. doi:10.2747/0272-3638.33.2.256

Weinberg, B.A., 2000. Black residential centralization and the spatial mismatch hypothesis. *Journal of Urban Economics*, 48, 110–134. doi:10.1006/juec.1999.2159

Williams, D.R. and Collins, C., 2001. Racial residential segregation: a fundamental cause of racial disparities in health. *Public Health Reports*, 116, 404–416. doi:10.1016/S0033-3549(04)50068-7

Witte, J.C. and Mannon, S.E., 2010. *The Internet and social inequalities*. London, UK: Routledge.

Wong, D.W.S. and Lee, J., 2005. *Statistical analysis and modeling of geographic information*. New York, NY: Wiley and Sons.

Wong, D. and Shaw, S.-L., 2011. Measuring segregation: an activity space approach. *Journal of Geographical Systems*, 13, 127–145. doi:10.1007/s10109-010-0112-x

Wyly, E.K., 1996. Race, gender, and spatial segmentation in the twin cities. *The Professional Geographer*, 48 (4), 431–444. doi:10.1111/j.0033-0124.1996.00431.x

Xu, C., Wong, D.W., and Yang, C., 2013. Evaluating the "Geographical Awareness" of individuals: an exploratory analysis of Twitter data. *Cartography and Geographic Information Science*, 40 (2), 103–115. doi:10.1080/15230406.2013.776212

Yu, H., 2007. Visualizing and analyzing activities in an integrated space-time environment: temporal geographic information system design and implementation. *Transportation Research Record: Journal of the Transportation Research Board*, 2024, 54–62.

Zhou, C., *et al*., 2004. Discovering personal gazetteers: an interactive clustering approach. *In*: *Proceedings of the 12th annual ACM international workshop on Geographic information systems*. New York, NY: ACM, 266–273.