Activity Space Estimation with
Longitudinal Observations of Social Media Data

Jae Hyun Lee1
Graduate Student Researcher, University of California, Santa Barbara
Email: lee@geog.ucsb.edu (Corresponding author)

Adam Davis1
Graduate Student Researcher, University of California, Santa Barbara
Email: awdavis@geog.ucsb.edu

Seo Youn Yoon2
Associate research fellow, Korea Research Institute for Human Settlements
E-mail: syyoon@krihs.re.kr

Konstadinos G. Goulias1
Professor, University of California, Santa Barbara
E-mail: goulias@geog.ucsb.edu

1 GeoTrans Lab. Department of Geography,
University of California Santa Barbara Santa Barbara, CA 93106, USA.
Phone: 805-284-7835

2 Korea Research Institute for Human Settlements 254 Simin-daero, Dongan-Gu, Anyang-Si,
Gyeonggi-do 431-712, South Korea Tel.: +82 31 380 0362.

Number of Words (including refs) =
Number of References =
Number of Tables and Figures = 10*250 = 2,500
Total equivalent =

GEOTRANS REPORT 2015-07-01

A shorter version of this report was submitted for presentation at the 95th Annual
Transportation Research Board Meeting, Washington, D.C.
January 10-14, 2016
Activity Space Estimation with
Longitudinal Observations of Social Media Data

Jae Hyun Lee*
Adam Davis
Seo Youn Yoon
Konstadinos G. Goulias

*corresponding author (lee@geog.ucsb.edu)
Department of Geography and GeoTrans
University of California Santa Barbara
Santa Barbara, CA 93106-4060

Abstract

In this paper, we demonstrate the use of an inexpensive and easy-to-collect long-term dataset to address the problems caused by basing activity space studies off short-term data. In total, we use 63,114 geotagged tweets from 116 unique users to create individuals’ activity spaces based on minimum bounding geometry (convex hull). By using polygon density maps of activity space, we found clear differences between weekday and weekend activity spaces, and were able to observe the growth trajectory of activity space over 17 weeks. In order to reflect the heterogeneous nature of spatial behavior and tweeting habits, we used Latent Class Analysis. As a result, three of the unique growth trajectories were found and compared both statistically and spatially. By comparing those latent growth trajectories with the total number of tweets and land use characteristics underneath the activity space, we estimate the optimal duration to capture individuals’ activity space. Additionally we extract the users’ major activity locations using a multilevel latent class model, and compare it with the growth patterns of their activity spaces. As a result, we found the growth of activity space were not depending on their major tweeting locations. In light of this, we can conclude that data sourced from Twitter can be a valuable source for long-term activity space research.

Paper submitted for 95th Annual Transportation Research Board Meeting, Washington, D.C.
January 10-14, 2016
1. INTRODUCTION

1.1. Objective

An activity space is the area through which a person travels during their day. Individuals’ activity spaces tend to center around fixed points in their life such as home, work, and school, and expand to cover the range of other locations they visit to pursue their various needs and wants. Activity spaces are particularly relevant to transportation research because they frame the travel decisions made by individuals and can help us understand what features of an area draw people to it. Travel diaries have generally been the primary data source for activity space estimation, but they are costly to collect over extended time periods. To address this shortcoming, we propose the use of harvested social media data to estimate long-term longitudinal activity spaces.

This study seeks: 1) to illustrate the value of harvested tweets as an inexpensive source of longitudinal data that can be used to measure activity spaces; 2) to estimate an optimal duration for activity data collection; and 3) to investigate the ways in which individuals’ activity spaces grow and change over time. Twitter users are clearly not a representative sample of the residents of this region or any region and even the heaviest users do not post from every location they visit during the day. So, Twitter is not the ideal source for studying all aspects of activity space. However, because it is possible to collect a much longer sample than any conventional survey design could, Twitter can give a very clear idea of how people’s experienced activity spaces expand over time.

1.2. Literature Review

Activity spaces emerged from Hägerstrand’s concept of time-space prisms and early work in the fields of time geography, behavioral geography, and travel behavior (Miller, 1991; Golledge & Stimson, 1997; Kitamura, Yoshii, & Yamamoto, 2009). Action/activity spaces represent the spatial component of an individual’s time-space prism (Kitamura et al., 2009). Golledge and Stimson define an action space as the totality of a person’s interaction with their environment, including movement through it and communication with people in it; activity spaces are the subset of action spaces that people actually move through with some regularity. Action spaces are highly individualized, but they follow predictable patterns – people center their activities on anchor points like home and work and interact less frequently with areas that are farther afield (Golledge & Stimson, 1997, Chapter 8; Neutens, Delafontaine, Scott, & De Maeyer, 2012). Travel behavior researchers’ applications of action/activity spaces more closely match Golledge’s definition of an activity space, so to avoid confusion we will use that term throughout this paper.

Though activity spaces are a fairly straightforward concept, there is substantial disagreement about how they should be measured or characterized spatially (Miller, 1991). The activity space a person experiences through travel is a subset of what Dijst describes as a “potential action space,” which contains all the possible locations at which that person believes they can pursue their desired activities (Dijst, 1999; Susilo & Kitamura, 2005) and that they can reach accounting for the spatial and temporal constraints provided by their schedule of mandatory activities (Miller, 1991; Neutens et al., 2012). Representations of activity spaces are usually generated from a set of activity locations (measured as points); two-dimensional representations attempt to fit one or more shapes around these points or around the routes between these points, and density-based representations map the space onto a smooth grid of values, with more activity-rich areas taking a higher value. Different representations produce different results (Schönfelder & Axhausen, 2003; Susilo & Kitamura, 2005), but there is no clear right choice for all research pertaining to activity spaces. Methods used depend on the fineness of detail in the source data (e.g. an explicit record of trips is required in order to build action spaces based on trips) and area of
Activity surfaces are intended to address the fact that people do not use all parts of their respective activity spaces equally. Kernel density measures assign a value to each cell of a raster grid cells based on the frequency of activity locations nearby. Instead of providing a discrete boundary around the places a person is likely to visit, activity surfaces report the probability that each area will be the site of an activity. People choose activity locations partly based on their proximity to other planned activities, allowing for trip chaining (Golledge & Stimson, 1997, p. 286). Schönfelder and Axhausen find that the quality of the resulting surface depends substantially on the bandwidth and shape of the function used to create the surface, though they suggest that a bandwidth of 1000m accurately accounts for people walking to locations near their activity sites (Schönfelder & Axhausen, 2003).

Discrete two-dimensional representations of action spaces lack the fineness of detail that activity surfaces provide, but they are much simpler to analyze because they clearly demarcate areas that are inside or outside the activity space. Minimum convex hulls (polygons that contain all activity points and have no internal angles greater than 180 degrees) are one of the simplest ways to represent an activity space. The method was first used by ecological researchers to estimate the home range of a population (analogous to that animal’s activity space) (Worton, 1987) and has been applied to human activity spaces in research examining the effects of urban form on activity spaces at the scale of the individual (Fan & Khattak, 2008) and household (Buliung & Kanaroglou, 2006). Convex hulls are very straightforward to compute, but they are clearly imperfect in the event that the person’s activity space is irregularly shaped, perhaps because the underlying geography makes some areas undesirable or inaccessible. Despite the shortcomings, research in ecology and other fields has found convex hulls to be useful and less biased than many other measures (Jennrich & Turner, 1969; Worton, 1987). Ellipse-based representations also originate from efforts to define home ranges in ecology (Jennrich & Turner, 1969). Since ellipses are generally designed to exclude some points, they are less affected by outliers than area convex polygons.

Most work using elliptical representations has sought to characterize the whole of a person’s activity space with one ellipse, but Schönfelder and Axhausen create separate ellipses for destinations of home- and work-based trips to investigate separate activity patterns (Schönfelder & Axhausen, 2003). Though home and work locations can be estimated for users who tweet very frequently, these methods have not yet been formalized. The work of Rai et al. tested various curved geometries to see which produced realistic activity spaces most efficiently (Rai et al., 2007), but conclude that none of the geometry-creation algorithms is clearly best.

Two-dimensional representations of activity spaces have also been constructed by adding a buffer around the shortest path between successive activity locations, (Schönfelder & Axhausen, 2003; Oliver, Schuurman, & Hall, 2007; Kim & Ulfarsson, 2015). Buffer widths depend on the purpose of the study and often approximate some estimate of “walking distance” (Kim & Ulfarsson, 2015). Shortest-path network buffers could be made from harvested tweets from particularly frequent users, but this would require assuming that the user either posts at each destination (which is unlikely), or intersperses each set of away-from-home tweets with at least one from home (which some users seem to).

Because activity/travel diaries provide a complete record of an individual’s movement, they are the obvious data source for studying space-time prisms and activity spaces (Miller, 1991), but these are difficult to collect. Because of the cost and difficulty of gathering long-term data, most studies use diaries covering three or fewer days of data (Dijst, 1999; Fan & Khattak, 2008; Kim & Ulfarsson, 2015). Short term datasets are most useful as cross-sectional studies which allows them to compare different people’s activity spaces, but cannot provide a clear sense of the variability of each person’s schedule from one day.
diaries and other information to estimate the change in person-based accessibility (measured in number of
people in one’s activity space) over time (Neutens et al., 2012).

No conventional data source provides a longer continuous stretch of longitudinal travel data than
Mobidrive, which collected six weeks of data from the residents of two German cities (Axhausen,
Zimmermann, Schöpfel, Rindsführer, & Haupt, 2002). This data source has served as the basis for
multiple studies of activity spaces because its longitudinal nature allows researchers to study the ways
individuals’ activity spaces grow and change over time, as they visit new places (Schöpfel &
Axhausen, 2003; Susilo & Kitamura, 2005). Susilo and Kitamura investigate day-to-day variability in
individuals’ activity space size using a particularly unique measure — the second moment of activity
location (essentially the variance of activity locations) — that provides a numerical value for the size of a
person’s activity space without asserting the specific area that it covers. This measure is particularly well-
suited for studying variability in activity space (Susilo & Kitamura, 2005).

Six weeks is a very long time to run a survey, but it is unclear whether this is long enough to fully
encompass the variability in individuals’ activity patterns. Social media data can provide a much longer
(though much less detailed) record of individuals’ travel for little or no cost. Twitter is a particularly
promising data source because of its widespread popularity, the frequency with which some users tweet,
and the ease of data collection using Twitter’s APIs. Though some startup effort is required to develop a
stable Twitter data collection process, once collection has started it allows for indefinite continuous
longitudinal study.

Though it is much more popular among people under the age of 30, Twitter was used by roughly 23% of
the American internet-using population in 2014, and usage is increasing in every age group, though usage
is still far heavier among people under the age of 30 (Duggan et al., 2015). It is relevant to note the
privacy and quality concerns inherent in harvested social media data. Though Volunteered Geographic
Information sources (such as Open Street Map) are increasingly valuable resources in transportation and
geography research, user-generated data often contains errors (Goodchild, 2007); in contrast the
geographic information in tweets is provided automatically by the mobile device, diminishing this error
considerably. Tweets are public by default, but users may not know how much information they are
sharing, making Twitter data a partial example of “coerced” geographic information (McKenzie &
Janowicz, 2014). Any use of harvested social media data should seek to protect user privacy as much as
possible, which this study does by displaying only aggregate information. Most uses of Twitter data in
transportation research have sought to exploit it as a source of real-time information about network
disruptions (Chan & Schofer, 2014; Mai & Hranac, 2013; Pender, Currie, Delbosc, & Shiawakoti, 2014) or
major disasters (Ukkusuri, Zhan, Sadri, & Ye, 2014), but the location information provided in geotagged
tweets could be more broadly applicable. One recent study used geotagged tweets to estimate an origin-
destination matrix for the Los Angeles area (Lee, Gao, & Goulias, 2015). In addition, this type of data
may also contain inherent biases because a study of activity spaces in the Puget Sound region found a
positive (though not quite statistically significant) relationship between internet use and activity space
size, which suggests that Twitter users may travel somewhat more than the general population (Kim &
Ulfarssson, 2015).

It is unlikely that any user posts frequently enough that the location information in their tweets would be
quite as detailed as a continuous GPS record, but many people use the system so frequently over a long
time period. The fact that some users tweet much more than others would be a major source of internal
bias for many applications, but it reinforces Twitter’s value as a source of longitudinal data about a core
set of users. Though it is uncertain just how much information a single tweet can contain, a geotagged
tweet provides a high degree of certainty that the user was at the reported location at the reported time. If we gather enough tweets from a user, it is relatively straightforward to visually identify anchor points like home and work locations, but we have not yet formalized a method with which to do this for large numbers of users. A dense long-term record of tweets should also make it possible to estimate detailed schedules for users that could provide comparable detail to travel diaries. Though we consider these applications of social media data to be very compelling, they are beyond the scope of this paper, which focuses on taking initial steps in using harvested tweets as the basis for activity space estimation.
2. DATA USED

2.1. Twitter Data
The main data source for this study was a nearly continuous collection of tweets in southern Santa Barbara County, CA. The collection used a bounding box of 119.5-120 degrees west longitude and 34.3 to 34.5 degrees north latitude and ran from November 23, 2014 to April 6, 2015, with multiple brief interruptions totaling no more than 4 days over the course of the collection period. Though the harvesting process yielded millions of tweets, many of these were not actually from within the study region (due to Twitter’s use of imprecise “place” tags in addition to gps-based geotags). Tweets were filtered by location and source (in order to exclude tweets from social-robots (Lee et al., 2015)), leaving roughly 150,000 tweets, each of which contained a time-stamped location for a user in the Santa Barbara region. For the purposes of this study, only those users who posted at least two tweets per day during the study period were considered, this left a final total of 63,114 tweets from 116 unique users. Figure 1 shows the geo-tagged tweets in southern Santa Barbara area for 17 weeks, which cover most of residential and commercial areas of southern Santa Barbara.

![Figure 1. Geo-tagged Tweets from November 23, 2014 to April 6, 2015](image)

2.2. Business Establishment Data
Since people’s activity spaces reflect their travel in pursuit of opportunities, opportunity-based land use metrics could provide a useful way to track the expansion and change of activity spaces (Chen et al., 2011; Davis, Lee, Deutsch-Burgner, McBride, & Goulia, 2015; Lee et al., 2015). Since opportunities are difficult to measure, we calculate proxies for opportunity density and diversity based on business establishments found in the Dun & Bradstreet National Establishment Time-Series (NETS) Database (2010), which contains an inventory of business establishments in the US. As a proxy for opportunity diversity we use the total number of employees of all businesses within a user’s activity space; as a proxy for diversity, we use Shannon Entropy based on the two-digit Standard Industrial Code (SIC) for each business establishment. This measure quantifies the uncertainty in the categorization of randomly selected entity, providing a simple measure of diversity (DeJong, 1975). To calculate the Shannon index, we determine the class proportion for each SIC category present in a given activity space \( \left( p_i; \text{where } i \in R \right) \), we then multiply each class probability by its logarithm and sum across all \( R \) SIC categories present (Equation 1).
\[ Shannon Entropy = - \sum_{i=1}^{n} p_i \ln p_i \]  

(1)

3. METHODS

3.1. Defining Activity Space

As discussed in the literature review, researchers have developed many different representations of activity spaces, including buffers, ellipses, and convex hulls. Because Twitter data provides a particular advantage for long-term activity space generation, we require a method that can capture the growth patterns of activity spaces from week to week (we choose one week as the primary unit for creating activity spaces because cyclical patterns are more stable and consistent). The two methods that seem most promising for this purpose are circular buffers around each of a user’s tweet locations and a minimum convex hull around all of a user’s tweets. Because it is unlikely that any user tweeted from each of their activity locations over any substantial time period, circular buffers are particularly likely to exclude important activity locations, while convex hulls should miss less of a person’s true activity space.

For this project, we calculated 17 weekly activity spaces using a four-step approach, described in Figure 2, using ArcGIS 10.2.2. The first step is to group all of an individual’s tweets from a given week and convert the geotags into a form usable in a GIS. In the second step, we create buffers around the location of each tweet. These buffers were calculated to account for inaccuracy in mobile device geocodes and to incorporate short-distance pedestrian travel around activity locations. The third step is to dissolve the boundaries among overlapping circular buffers and combine all of a user’s separate buffers into a single multipart object. Lastly, a minimum bounding geometry (convex hull) was created around the combined buffer areas.

Step 1: Geotagged tweets

Step 2: Create buffer surrounding Tweets’ location

Step 3: Merge buffers

Step 4: Create action space using convex hull

Figure 2. Defining Activity Space with Geotagged Tweets

This process was used to produce 17 weekly activity spaces for each heavy Twitter user in the study area. In order to investigate the longitudinal growth patterns of activity space, we accumulate these activity spaces by week. For example, the accumulated second week’s activity space is created by merging
(dissolving) the first week’s and second week’s tweets and generating a concave hull activity space. Figure 3 shows the expansion of activity space over time from one user in our sample. Although this person has very small activity space in the first week, it expands over time as they visit a wider range of their preferred destinations.

Figure 3. Expansion of Activity Space over Time

3.2. Latent Class Analysis

In order to classify the growth pattern of the users’ activity spaces over the 17-week study, we use a Latent Class Cluster Model because it can capture the heterogeneity of individuals’ activity space growth patterns by classifying them into groups (Vermunt & Magidson, 2013). Another advantage of this model is that it allows each latent class to have its own growth pattern instead of using fixed growth coefficients for the entire sample. In other words, our model produces a data-driven growth trajectory for each latent class; using a different trajectory for each group is particularly useful to identify the point at which activity space growth plateaus. Equation 2 illustrates a latent class analysis model without covariates.
(Vermunt & Magidson, 2013): \(x\) denotes a single nominal latent variable, \(y_{it}\) is the response variable. \(T\) is the number of individuals. In this paper, the class membership is a function of the spatial area and land use characteristics of individuals’ activity space over the 17 weeks.

\[
f(y_i) = \sum_{x=1}^{K} P(x) \prod_{t=1}^{T} \{y_{it} | x\} \quad (2)
\]

Some twitter users use the social service by far more than others and this needs to be accounted for when developing latent classes and user memberships. Since tweets are nested within users, we use a Multilevel Latent Class model for descriptive analysis of geotagged tweets and the users. Then, the results of this analysis are used in the host-hoc analysis. We use all of the geo-tagged tweets’ locations (longitude and latitude) to define low-level classes, and Twitter users are classified as high-level classes simultaneously. In this way, we classify members of our sample based on the geographical locations of their tweets as well as the intensity of tweets in certain locations. We can then compare these results with the growth patterns of their activity spaces. The following two equations are used to estimate Multilevel Latent Class model in Equations 3 and 4 (Lukočienė, Varriale, & Vermunt, 2010, p. 252).

\[
f(y_k) = \sum_{x=1}^{H} P(w_k = h) \prod_{j=1}^{n_k} f(y_{kj} | w_k = h) \quad (3)
\]

\[
f(y_{kj} | w_k = h) = \sum_{l=1}^{L} P(x_{kj} = l | w_k = h) \prod_{i=1}^{I} (y_{ki} | x_{kj} = l, w_k = h) \quad (4)
\]

where,

- \(i\): response variable
- \(j\): lower level unit (cluster of tweets)
- \(k\): higher level unit (class of users)
- \(x_{kj}\): lower-level class membership (class membership of geotagged tweet);
- \(y_{ki}\): responses of individual \(i\) from group \(k\) (longitude and latitude of tweets for \(l\) from group \(k\))
- \(w_k\): higher-level class membership (class membership of Twitter users)
- \(n_k\): the number of individuals within group \(k\) (the number of twitter users within group \(k\))
- \(y_k\) the full set of responses of group \(k\) (the full set of geotagged tweets of group \(k\))
- \(L\): number of lower classes (class of geotagged tweets)
- \(H\): number of higher classes (class of Twitter users)

In the first equation (higher level; Equation 3), \(P(w_k = h)\) corresponds to the probability that user \(k\) belongs to Latent Class \(h\) and \(f(y_{kj} | w_k = h)\) is the conditional probability density for response vector of individual \(j\) belong to group \(k\), which is conditional on latent class membership \(h\). In the second equation, \(P(x_{kj} = l | w_k = h)\) is probability that individual \(j\) of group \(k\) in latent class \(l\) given condition of higher latent class \(h\). \((y_{ki} | x_{kj} = l, w_k = h)\) is conditional probability density for response variable \(i\) of individual \(j\) in group \(k\) given condition of the membership to lower-level class \(l\) and higher level class \(h\).

In order to estimate this model, we use the Expectation-Maximization (EM) algorithm along with Newton-Raphson Maximum Likelihood Estimation (Vermunt & Magidson, 2002). For a given set of starting values, maximum likelihood iterations is performed with the EM algorithm until the change in likelihood value is smaller than a criterion. Then Newton-Raphson runs until a predefined limit of
convergence is reached. The results of estimation with the combination of two algorithms are quite stable, but this type of model is still very sensitive to local maxima of the likelihood function. This issue can be resolved by testing multiple models with different sets of parameter start values (Goulias, 1999).

An operational issue that arises from this type of model is that degrees of freedom are rapidly exhausted as we increase the number of parameters (usually by increasing the number of classes). This can make it difficult to achieve both model identification (ability to estimate parameters) and converge (subsequent estimation step parameters are not close enough). To address this issue, we use a hierarchical approach when estimating this model,

a) First we estimate the model with an assumption that there is only one class.
b) We proceeded by increasing the number of latent classes until model identification becomes impossible and the resulting classes become too small.
c) We select the most suitable number of classes based on multiple goodness of fit criteria, including Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and the Consistent Akaike Information Criterion (CAIC), based on (McCutcheon, 2002; Nylund, Asparouhov, & Muthén, 2007).

However, we need an additional model-fitting strategy to decide number of classes for the multi-level latent class model. We use a three-step approach to fit the model because it is less computationally expensive than a two-step approach, and it helps to separate both levels (Lukočienė et al., 2010). The first step is to determine the number of lower-level classes regardless of the multilevel structure. Then we fix the number of lower-level classes and determine the number of higher-level classes. Lastly, we re-determine the number of lower-level classes with a fixed number of higher-level classes. This third step aims to test whether the number of lower-level classes is affected by the multilevel data structure, and we repeat the process until the number of classes stabilizes.

Although twitter data does not provide users’ home locations, we can identify their major tweeting locations with long term locational information from users. In order to do that, we first use the multilevel latent class model to classify individuals using all of their tweeting locations. Then, we compare it with growth pattern of activity space. In this way, we classify individuals based on the intensity of tweets in space as well as total number of tweets. As a result, five classes of activity locations (lower classes) and five groups of twitter users were identified (Table 1). This indicates there are five main spatial centers of geotagged tweets, and five types of users tweeted differently in terms of spatial distributions of geotagged tweets.

Figure 4a shows the major class centers of geotagged tweets in blue and activity centers in red of users depending on higher-level latent classes, and their size of symbols denotes the size of latent classes. All of the tweets’ class-membership can be found in Figures 4b and 4c. The major activity centers are located in Isla Vista for classes 1, 2, and 4 (68.6% of tweets), downtown Santa Barbara for class 5 (8.3%), and upper area across southern Santa Barbara for class 3 (23%). User classes were more centered in the middle of the region because users travel and tweet throughout the region, which can be seen in Figure 4c and Table 2. Table 2 shows the cross tabulation of multilevel latent classes and their relationship with the place names users reported. Each user’s class has a corresponding major activity location, the strongest corresponding relationship can be found in the activity location class 5 – user class 5 (downtown Santa Barbara), whereas the least one was found in class 4 for both class memberships (west Isla Vista). In fact, this group is one of the most mobile group within Isla Vista area. However, the most mobile group (spatially well-distributed tweets) across entire southern Santa Barbara area is group class 3, mostly engaged in upper Santa Barbara area. Although the first and second user classes are mostly engaged with
the Isla Vista area, the second group tends to have slightly dispersed spatial behavior. These results are compared to the tweet’s reported place locations (Table 2, bottom). Users in classes 1, 2, 4 and 5 tweet mainly from Isla Vista, and Santa Barbara, but the class 3 tweets from both Santa Barbara and Goleta. In summary, Isla Vista and downtown Santa Barbara contain the heaviest concentration of major activity locations for members of our sample, but these users travel (and tweet) all over the region.

Table 1. Model fit indices for multilevel latent class model for activity locations

<table>
<thead>
<tr>
<th>Step</th>
<th>Model</th>
<th>Model Types</th>
<th>LL</th>
<th>BIC(LL)</th>
<th>AIC(LL)</th>
<th>AIC3(LL)</th>
<th>CAIC(LLL)</th>
<th>Npar</th>
<th>Class. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step1</td>
<td>Model1</td>
<td>1-Cluster</td>
<td>285281.4</td>
<td>-570519</td>
<td>-570555</td>
<td>-570551</td>
<td>-570515</td>
<td>4</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Model2</td>
<td>2-Cluster</td>
<td>372846.0</td>
<td>-745593</td>
<td>-745674</td>
<td>-745665</td>
<td>-745584</td>
<td>9</td>
<td>0.0073</td>
</tr>
<tr>
<td></td>
<td>Model3</td>
<td>3-Cluster</td>
<td>400481.2</td>
<td>-800808</td>
<td>-800934</td>
<td>-800920</td>
<td>-800794</td>
<td>14</td>
<td>0.0109</td>
</tr>
<tr>
<td></td>
<td>Model4</td>
<td>4-Cluster</td>
<td>429640.1</td>
<td>-859070</td>
<td>-859242</td>
<td>-859223</td>
<td>-859051</td>
<td>19</td>
<td>0.0097</td>
</tr>
<tr>
<td></td>
<td>Model5</td>
<td>5-Cluster</td>
<td>440162.5</td>
<td>-880060</td>
<td>-880277</td>
<td>-880253</td>
<td>-880036</td>
<td>24</td>
<td>0.0274</td>
</tr>
<tr>
<td></td>
<td>Model6</td>
<td>6-Cluster</td>
<td>445976.5</td>
<td>-891632</td>
<td>-891895</td>
<td>-891866</td>
<td>-891603</td>
<td>29</td>
<td>0.0303</td>
</tr>
<tr>
<td></td>
<td>Model7</td>
<td>7-Cluster</td>
<td>449591.6</td>
<td>-898807</td>
<td>-899115</td>
<td>-899081</td>
<td>-898773</td>
<td>34</td>
<td>0.0578</td>
</tr>
<tr>
<td>Step2</td>
<td>Model8</td>
<td>5-Cluster 1-GClass</td>
<td>415026.5</td>
<td>-888060</td>
<td>-888277</td>
<td>-888253</td>
<td>-888036</td>
<td>24</td>
<td>0.0274</td>
</tr>
<tr>
<td></td>
<td>Model9</td>
<td>5-Cluster 2-GClass</td>
<td>465657.7</td>
<td>-930995</td>
<td>-931257</td>
<td>-931228</td>
<td>-930966</td>
<td>29</td>
<td>0.0283</td>
</tr>
<tr>
<td></td>
<td>Model10</td>
<td>5-Cluster 3-GClass</td>
<td>473275.9</td>
<td>-946176</td>
<td>-946484</td>
<td>-946450</td>
<td>-946142</td>
<td>34</td>
<td>0.0272</td>
</tr>
<tr>
<td></td>
<td>Model11</td>
<td>5-Cluster 4-GClass</td>
<td>470321.5</td>
<td>-940212</td>
<td>-940565</td>
<td>-940526</td>
<td>-940173</td>
<td>39</td>
<td>0.0168</td>
</tr>
<tr>
<td></td>
<td>Model12</td>
<td>5-Cluster 5-GClass</td>
<td>497547.2</td>
<td>-994608</td>
<td>-995006</td>
<td>-994962</td>
<td>-994564</td>
<td>44</td>
<td>0.0151</td>
</tr>
<tr>
<td></td>
<td>Model13</td>
<td>5-Cluster 6-GClass</td>
<td>498652.7</td>
<td>-996764</td>
<td>-997208</td>
<td>-997159</td>
<td>-996715</td>
<td>49</td>
<td>0.0153</td>
</tr>
<tr>
<td></td>
<td>Model14</td>
<td>5-Cluster 7-GClass</td>
<td>499665.0</td>
<td>-998733</td>
<td>-999222</td>
<td>-999168</td>
<td>-998679</td>
<td>54</td>
<td>0.0152</td>
</tr>
<tr>
<td>Step3</td>
<td>Model15</td>
<td>5-Cluster 1-GClass</td>
<td>285281.4</td>
<td>-570474</td>
<td>-570547</td>
<td>-570539</td>
<td>-570466</td>
<td>8</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Model16</td>
<td>5-Cluster 2-GClass</td>
<td>400834.3</td>
<td>-801481</td>
<td>-801635</td>
<td>-801618</td>
<td>-801464</td>
<td>17</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td>Model17</td>
<td>5-Cluster 3-GClass</td>
<td>436883.3</td>
<td>-873479</td>
<td>-873715</td>
<td>-873689</td>
<td>-873453</td>
<td>26</td>
<td>0.0072</td>
</tr>
<tr>
<td></td>
<td>Model18</td>
<td>5-Cluster 4-GClass</td>
<td>465664.1</td>
<td>-930941</td>
<td>-931258</td>
<td>-931223</td>
<td>-930906</td>
<td>35</td>
<td>0.0176</td>
</tr>
<tr>
<td></td>
<td>Model19</td>
<td>5-Cluster 5-GClass</td>
<td>497547.2</td>
<td>-994608</td>
<td>-995006</td>
<td>-994962</td>
<td>-994564</td>
<td>44</td>
<td>0.0151</td>
</tr>
<tr>
<td></td>
<td>Model20</td>
<td>5-Cluster 6-GClass</td>
<td>505017.5</td>
<td>-1009449</td>
<td>-1009297</td>
<td>-100976</td>
<td>-1009396</td>
<td>53</td>
<td>0.0276</td>
</tr>
<tr>
<td></td>
<td>Model21</td>
<td>5-Cluster 7-GClass</td>
<td>506226.9</td>
<td>-1011769</td>
<td>-1012330</td>
<td>-101268</td>
<td>-1011707</td>
<td>62</td>
<td>0.0240</td>
</tr>
</tbody>
</table>

Table 2 Cross Tabulation of User Classes and Activity Location Classes

<table>
<thead>
<tr>
<th>Activity Location Clusters</th>
<th>User Classes</th>
<th>Location/Profile Clusters size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusters</td>
<td>GClass1</td>
<td>GClass2</td>
</tr>
<tr>
<td>Cluster1</td>
<td>0.1434</td>
<td>0.8535</td>
</tr>
<tr>
<td>Cluster2</td>
<td>0.7686</td>
<td>0.0000</td>
</tr>
<tr>
<td>Cluster3</td>
<td>0.0000</td>
<td>0.1195</td>
</tr>
<tr>
<td>Cluster4</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Cluster5</td>
<td>0.0248</td>
<td>0.0452</td>
</tr>
<tr>
<td>Isla Vista</td>
<td>0.9200</td>
<td>0.8460</td>
</tr>
<tr>
<td>Goleta</td>
<td>0.0210</td>
<td>0.0870</td>
</tr>
<tr>
<td>Santa Barbara</td>
<td>0.0450</td>
<td>0.0620</td>
</tr>
<tr>
<td>All Others</td>
<td>0.0130</td>
<td>0.0050</td>
</tr>
</tbody>
</table>

User Clusters’ size: 41 (0.3521) 32 (0.2752) 21 (0.1812) 16 (0.1385) 6 (0.0530)
Figure 4. Activity Locations with Multilevel Latent Classes
5. RESULTS AND FINDINGS

As described in the methods section, we create cumulative weekly activity spaces for heavy Twitter users using minimum convex hulls; for each of these activity spaces, we enumerate the land use characteristics of the areas they cover as a proxy for the opportunities available in the activity space. We then estimate a latent class model for activity space growth patterns. These processes were accomplished using ArcGIS 10.2.2, SPSS 22.0, and Latent GOLD 5.0.

5.1. Aggregated Density of Activity Space

In total, 116 individuals’ end-of-period (cumulative weeks 1-17) activity spaces were created (Figure 5a), each with a unique shape and coverage. In order to see the aggregated pattern of activity space, we created polygon density maps (Figure 5b). The zone with an intense red color that spans from the center of the Santa Barbara region to Isla Vista (a town populated mainly by university students) indicates the area included in the most users’ activity spaces (93-100 Twitter users in this case). The particularly high density of activity spaces in Isla Vista indicates that younger adults make up a large share of our sample, which generally matches the known demographics of Twitter users discussed in the literature review.

Although some people’s activity spaces extend north into a mountainous area or east to Summerland, the majority of our sample stayed in the area of downtown Santa Barbara (the area just east of the middle of the map), old town Goleta (west of the center), Isla Vista (southwest), and Camino Real shopping center (west). Although old town Goleta is included in many users’ activity spaces, it contains fairly few actual tweets, indicating that few people in our sample pursued activities in this area (as shown in Figure 1). Though this area may not be visited by many of our users, its central location means that many people pass through it when traveling between Isla Vista and Santa Barbara, meaning its inclusion in the activity spaces is plausible.

We also explore the differences between total weekday and weekend activity spaces (Figure 5c and 5d). Since they are based on activities from fewer days, weekday activity spaces are generally larger than weekend ones. The major difference between the two maps can be seen in downtown Santa Barbara and old town Goleta, where many business establishments are located; both areas are heavily visited on weekdays, but are included in the weekend activity spaces of far fewer of the users we studied; in contrast, the relative dominance of Isla Vista is much more pronounced on weekends (Figure 5d). This suggests activity locations that people need to visit regularly (such as home, work and school) plays important role in creating activity spaces. Isla Vista serves as a major anchor point for its residents (84.8% of whom are aged between 18-24 years, according to Census 2010) and students attending UC Santa Barbara. This is also a particularly attractive site for weekend socializing. Though downtown Santa Barbara contains the densest concentration of business establishments in the region, our respondents visited it less than they did Isla Vista on weekends. This likely reflects the popularity of Isla Vista (famous for its informal entertainment).
Figure 5. Activity Space and Density Maps of Activity Space
5.2. Longitudinal Growth Pattern of Activity Space

Figure 6 illustrates the growth of activity space over the 17 weeks of our study, using the same color scheme as the other maps. In the first week, the highest activity space density was found in the Isla Vista area. This area of high concentration expands over time mainly towards downtown Santa Barbara and Goleta area between 3rd week and 5th week as more users leave Isla Vista. This can be also found in the chart at the bottom of Figure 6, which shows the average size of users’ activity spaces in squared kilometers; there is a slight plateau of growth of activity space between 3rd and 7th week. After that, individuals’ activity spaces resume growing, on average, but the aggregated sample activity space has already reached its maximum size within about 3-4 weeks. The rate of increase of average activity space size slows around 14th week, which matches what is shown in the density maps, since the maps of 15th and 17th week are very similar. This suggests that the ability to gain new information from flat activity spaces produced from social media data decreases substantially after about 14 weeks, because in this time period most users will have visited most of the areas they ever visit. The plateau from week four to week seven corresponds to the holiday season and universities’ winter breaks; this suggests many people’s local activity spaces stopped expanding because they were out of the area.

5.3. Heterogeneity of Longitudinal Growth Pattern

The average growth pattern of activity spaces for Twitter users in Santa Barbara cannot reflect the heterogeneous nature of individuals’ spatial behaviors or the differences in their tweeting habits. In order to address this shortcoming, we used Latent Class Analysis. As we discussed in the methods section, this model allowed us to identify possible sub-groups in our sample and estimated average growth trajectories for each group. The development process for this model entails running it repeatedly, increasing the number of classes until it fails to converge or identifies a meaningless class (one with too few members). Table 3 shows the goodness of fit indices of the series of the latent class analysis model. According to the changes in BIC, AIC3 and CAIC the three class model describes the heterogeneity of growth patterns well because models with more classes do not represent substantial improvements over it. This indicates that adding more parameters by increasing the number of classes does not improve the latent class model.

Figure 7 shows the proportion of users in each latent class (left) and the longitudinal growth profiles of activity space by the different class (right). The first class is largest, but the three are relatively balanced. In terms of growth trajectories, the three classes are very different. The first class shows the moderate growth pattern and its growth trajectory (red) is almost same with the average growth trajectory (gray dashed line). However, this class shows a more severe plateau between 3rd and 7th week. Members of the second class (green) have the largest activity spaces on average. Their activity spaces grow rapidly until they begin to reach a plateau around weeks 11-13. On average, their activity spaces are roughly twice as large as the third class, and about 30km² bigger than the first class. Members of the third class have the smallest activity spaces, and their average activity spaces remain very small until 9th week, when they begin to expand more rapidly.

The density maps of activity space growth can be found in Figure 7 (bottom). The first class shows a nearly identical growth pattern to the sample as a whole. The activity spaces of this class’s members start from the Isla Vista area and expanded towards downtown Santa Barbara and Goleta. As discussed, there was almost no change for this class between weeks 4 and 9, so the density maps of week 3, 5, and 7 are similar. Class 1’s activity spaces increase again from the 9th week, and most class members appear to have visited most of the same areas, since the middle of the activity space density map appears fairly homogeneous. The second group has the largest activity space overall at all points in time, and most of the region was covered by most members of second group from week 9. In addition, their activity spaces are not concentrated on Isla Vista, but instead appear to be spread throughout the region even in the
earlier weeks. The third class has very unique growth pattern. Although it includes several people who only tweeted from downtown Santa Barbara, this group is especially heavily concentrated in Isla Vista, and most users in this group did not tweet from outside that small area until 11\textsuperscript{th} - 13\textsuperscript{th} weeks.

Figure 6 Growth Patterns of Activity Space

Table 3. Goodness of Model-fit Indices for Latent Class Analysis

<table>
<thead>
<tr>
<th></th>
<th>LL</th>
<th>BIC(LL)</th>
<th>AIC(LL)</th>
<th>AIC3(LL)</th>
<th>CAIC(LL)</th>
<th>Npar</th>
<th>Class.Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Cluster</td>
<td>-8516.11</td>
<td>17193.84</td>
<td>17100.22</td>
<td>17134.22</td>
<td>17227.84</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>2-Cluster</td>
<td>-7870.21</td>
<td>16068.42</td>
<td>15878.42</td>
<td>15947.42</td>
<td>16137.42</td>
<td>69</td>
<td>0.0104</td>
</tr>
<tr>
<td>3-Cluster</td>
<td>-7550.53</td>
<td>15595.42</td>
<td>15309.05</td>
<td>15413.05</td>
<td>15699.42</td>
<td>104</td>
<td>0.0061</td>
</tr>
<tr>
<td>4-Cluster</td>
<td>-7369.03</td>
<td>15398.80</td>
<td>15016.05</td>
<td>15155.05</td>
<td>15537.80</td>
<td>139</td>
<td>0.0074</td>
</tr>
<tr>
<td>5-Cluster</td>
<td>-7186.80</td>
<td>15200.73</td>
<td>14721.61</td>
<td>14895.61</td>
<td>15374.73</td>
<td>174</td>
<td>0.0110</td>
</tr>
<tr>
<td>6-Cluster</td>
<td>-7022.98</td>
<td>15039.47</td>
<td>14463.97</td>
<td>14672.97</td>
<td>15248.47</td>
<td>209</td>
<td>0.0010</td>
</tr>
</tbody>
</table>
Figure 7. Proportions of Clusters and Growth patterns of activity space by latent classes
5.4. Post-hoc Analysis

Although this inexpensive data source does not provide many concrete details about its users or their posts, we can explain some of the differences between latent user classes by examining the frequency with which users tweeted, and the land use of the areas their activity spaces covered.

First, we compute the accumulated number tweets per a user to compare it with growth pattern of activity space, and aggregated that by groups and weeks (figure 8a). As expected, the first group has very similar pattern with the mean of entire sample and the second group posted the largest number of tweets. Interestingly, though the third group tweeted from a much smaller range of areas, they posted nearly as frequently as the first group. This suggests that the size of an activity space is not simply a function of the number of tweets a user makes from mobile devices, and may reflect actual differences in behavior. Additionally, although the activity spaces stop growing, none of the groups shows much decrease in total number of tweets. This suggests that the number of weeks required to determine the spatial extent of activity spaces can be identified by noting the point at which an increase in number of geotagged tweets stops contributing to an expansion of activity space.

Land use characteristics can also be used to examine the growth of activity spaces. Business density and diversity (as described in the second part of the Data Used section) were calculated for each activity space and then aggregated by latent class and by week (Figure 8b, 8c). Growth in number of employees covered by activity spaces appears to follow a very similar to the growth pattern to the area of the activity spaces (Figure 8b). The first class and the mean of entire sample are very similar in this chart, and the second class has exposed to by far larger amount of opportunities than the others. The third group is associated with the smallest number of opportunities. The growth pattern of business establishment diversity shows a unique pattern. For the first two classes, diversity increases until it approaches the region-wide diversity of around 3.5. This plateau is reached faster than were plateaus in area or density. This is presumably because downtown Santa Barbara contains the areas peak diversity as well as a large share of its businesses, so activity space diversity will not continue to increase after the activity space covers this part of the region, which the classes do at various times as shown in Figure 7. Unlike the activity space growth pattern, there are some slight decreases in these longitudinal trajectories of means. This is because of missing cases. In other words, some people started use geo-tagged tweets while we collected data therefore, smaller size of activity space will be added in the middle of weeks. This will lead to decreases in the number of employees as well as diversity. The reason why it did not happened in the activity space, we used values that are estimated by model to create group-wise activity space growth trajectories.

Lastly, individuals’ the growth patterns of activity space are compared with their spatial distribution of tweets (the results of explanatory analysis in section 4). Table 4 illustrates the cross-classified table of activity space growth patterns and segments of users classified based on their major locations of tweets. Overall, it does not seem to have a strong corresponding relationship between two results of classifications. Although the moderate activity space growth pattern (class 1) has the largest proportions in most activity location based user classes except the fifth class, those are slightly larger when it compares to marginal distributions. Although the class 4 and 5 seem to have unique patterns (6.3 % at location class 4 – growth class 2, 0% at location class 5 – growth class 1), it is hard to conclude that those groups’ spatial tweeting behavior affect the growth of activity space growth due to the small sample size.
a) Accumulated Number of Tweets for 17 weeks

b) Average of Business Employees in Activity Space

c) Average of Business Diversity in Activity Space
Figure 8 Post-hoc Analysis (Number of tweets, average number of employees, and Business diversity)
Table 4. Cross-Classification Table of Activity Space Growth Clusters and Activity Location-based Clusters

<table>
<thead>
<tr>
<th>Activity location-based User Classification</th>
<th>Activity Space Growth Cluster</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Count</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>% within Activity Space Cluster</td>
<td>39.0%</td>
<td>36.6%</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>% within Activity Space Cluster</td>
<td>49.9%</td>
<td>34.4%</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>% within Activity Space Cluster</td>
<td>57.1%</td>
<td>23.8%</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>% within Activity Space Cluster</td>
<td>50.0%</td>
<td>6.3%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>% within Activity Space Cluster</td>
<td>0.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>51</td>
<td>35</td>
</tr>
<tr>
<td>% within Activity Space Cluster</td>
<td>44.0%</td>
<td>30.2%</td>
</tr>
</tbody>
</table>

6. SUMMARY AND CONCLUSION

In this paper, we demonstrate the use of an inexpensive and easy-to-collect long-term dataset to address the problems caused by basing activity space studies off short-term data. In total, we use 63,114 geotagged tweets from 116 unique users to create individuals’ activity spaces based on minimum bounding geometry (convex hull). By using polygon density maps of activity space, we found clear differences between weekday and weekend activity spaces, and were able to observe the growth trajectory of activity space over 17 weeks. In order to reflect the heterogeneous nature of spatial behavior and tweeting habits, we used Latent Class Analysis. As a result, three of the unique growth trajectories were found and compared both statistically and spatially. By comparing those latent growth trajectories with the total number of tweets and land use characteristics underneath the activity space, we estimate the optimal duration to capture individuals’ activity space. Additionally we extract the users’ major activity locations using a multilevel latent class model, and compare it with the growth patterns of their activity spaces. As a result, we found the growth of activity space were not depending on their major tweeting locations. In light of this, we can conclude that data sourced from Twitter can be a valuable source for long-term activity space research.

Although we successfully estimated activity spaces, there are several drawbacks to our methods. The construction of convex hulls can result in including the unvisited areas and ignores spatial variability in the activity patterns of individual users. Long-distance travel (to areas outside of the Santa Barbara area) was excluded in part because convex hulls would likely be an inappropriate means of studying it, but many users likely left the region during the holiday season, making it part of their true activity space. Therefore, our immediate step is to develop a method to address issues from convex hull, and collect social media data over a larger spatial area.
References


