IV WPSC
IV WORLD PLANNING SCHOOLS CONGRESS JULY 3-8TH 2016 RIO, BRAZIL

Who uses geolocated social media in the United States?

Samuel Maurer, PhD Student
Urban Analytics Lab
Department of City & Regional Planning
University of California, Berkeley
maurer@berkeley.edu

Keywords: participation, activity spaces, big data, social media

There has been increasing interest from both city planning scholars and practitioners in using geolocated social media posts as a way to learn about urban activity patterns: which public spaces are most popular at different times, what people's travel patterns look like, and so forth. Information about residential demographics, employment, and motor vehicle traffic is often already available from census agencies or roadway sensors, but widespread smartphone adoption now provides novel information about people's actual uses of urban space.

However, not all residents of cities have access to smartphones, and those who do may not use the type of apps like Twitter, Instagram, Flickr, or Foursquare that create public traces of location and activity context. From large-scale surveys, we know that users of these services in the United States tend to be younger and wealthier than average residents, and are less likely to be from disadvantaged racial or ethnic groups. This leads to concern that city planners utilizing this data may inadvertently privilege some members of the community over others.

This paper demonstrates a methodology to assess the demographic characteristics of geolocated social media data for metropolitan areas in the United States, using data form the U.S. Census. We will report findings for the entire U.S. if possible, and otherwise will focus on the San Francisco Bay Area, a region of about 9 million people in the state of California. Our objective is to provide city planning researchers and practitioners with information and tools for validating whether

particular datasets are likely to represent a fair sample of urban residents or one that is biased in particular directions.

Using Twitter's developer API, we collect public social media posts in real time as they are either posted directly to Twitter or cross-posted from other serves. A large portion of users create posts from their smartphones, and some have enabled location tagging. In this case the posts include latitude/longitude coordinates from the GPS chip in the smartphone. If possible, we will augment the Twitter data with similar data from other social media platforms.

With a large enough data sample, we can infer the approximate home location of individual users from their patterns of evening and nighttime activity. Based on a proof-of-concept for the San Francisco Bay Area, we find that just 10 days of social media activity is enough to infer a plausible, approximate home location for 10% of users in the sample. (We validate the location inferences against users' public profiles, in cases where they have provided relevant information.)

Using these locations, we match the users to home census tracts, which are small, relatively homogeneous districts of about 4,000 inhabitants which are defined in a consistent manner by the U.S. Census Bureau. We cannot make direct demographic claims about the social media users, because of possible biased selection within the tracts, but we are able to assess the characteristics of tracts which are over- or under-represented in the dataset: whether they have systematic differences in income, age, race/ethnicity, or geographic location within the region, compared to tracts with average representation.

Our initial findings for the San Francisco Bay Area indicate that contrary to expectations, people who create large volumes of geolocated social media posts actually live in census tracts with typical income and race/ethnicity distributions for the region. However, they are likely to come from tracts with much younger age profiles than average, including areas that are known for housing university students and people in their 20s.

In California alone, geolocated social media posts have already been used to study which locations are most popular for tourists vs. locals, which public parks are popular at different times, and how people's activity patterns agree with or diverge from

zoned land uses. These new findings suggest that such studies likely speak for a broad socioeconomic cross-section of the population, but will under-represent the elderly.

The tools developed in our paper will be available for others to use as open-source software libraries. Along with our findings (based on geolocated social media posts from Summer and Fall 2015), this will aide city planning researchers and practitioners in using social media activity traces more knowledgeably and responsibly — by validating whether particular datasets are likely to represent a fair sample of urban residents or a sample that is biased in particular directions.

References:

- Duggan, M., et al. (2015), "Social Media Update," Pew Research Center.
- Kling, F. & A. Pozdnoukhov (2012), "When a City Tells a Story: Urban Topic Analysis,"
  ACM GIS '12.
- Ratti, C., et al. (2006), "Mobile landscapes: using location data from cell phones for urban analysis," *Environment and Planning B*, vol. 33, no. 5.
- Tasse, D. & J. Hong (2014), "Using Social Media Data to Understand Cities," *Workshops on Big Data and Urban Informatics '14*.