Detecting Origin-Destination (OD) Mobility Flows From Geotagged Tweets in Greater Los Angeles Area

Song Gao\textsuperscript{1}, Jiue-An (Jay) Yang\textsuperscript{1,2}, Bo Yan\textsuperscript{1}, Yingjie Hu\textsuperscript{1}, Krzysztof Janowicz\textsuperscript{1}, Grant McKenzie\textsuperscript{1}

\textsuperscript{1}STKO Lab, University of California, Santa Barbara, CA, USA
\textsuperscript{2}San Diego State University, CA, USA

http://stko.geog.ucsb.edu/
Big Geo-Data

- Mobile Phone Data
- GPS-Enabled Taxis
- LBSN
- Social Media


FOR IMMEDIATE RELEASE: TUESDAY, MARCH 5, 2013

Census Bureau Reports 471,000 Workers Commute into Los Angeles County, Calif., Each Day

Los Angeles County, Calif., has among the highest number of commuters coming from another county in the nation, the U.S. Census Bureau reported today in new estimates released from the American Community Survey. Nationally, 27.4 percent of workers commute outside the county where they live.

Among workers in Los Angeles County, 471,345 live outside the county, according to 2006-2010 estimates from the American Community Survey. For example, 178,681 workers commute in from Orange County, 126,642 from San Bernardino County and 66,832 from Ventura County.

Meanwhile, 335,676 residents of Los Angeles County leave the county for work, with 181,744 going to Orange County (which was not significantly different from the number of workers coming in from the county), 57,390 to San Bernardino County and 36,602 to Ventura County.

"It is well known that Los Angeles County draws a lot of commuters to work. The detailed information in the American Community Survey tells us where Los Angeles County workers are coming from, where its residents work, and how its commuting patterns compare to those of other large counties," said Brian McKenzie, a Census Bureau statistician who studies commuting. "This information shapes our understanding of the boundaries of local and regional economies, as people and goods move across the nation's transportation networks."
$170 Million, 1% of households
Motivation

http://cdb.io/1vbu9Kl
Data

- Twitter Streaming API
- Covering Los Angeles County and 4 nearby counties.
- 6.8 million geotagged tweets from 110,868 users
- December 7, 2013 to January 7, 2014
About 11000 (10%) users tweet more than 5 geotagged tweets per day.
The distribution fits a truncated power function with the exponent value 1.94 and R-square 0.93.
Average inter-tweeting time

- The **mean** of individual average inter-tweeting time interval per day for all users is **126 minutes**, and the **median** is **79 minutes**.
- The majority (about **80%**) of users is within **190 minutes** per day.
Retrieving OD Flow

• **Step 1: individual-based trajectory detection**
  - Let $L_i = (l_1, l_2, l_3, \ldots)$ denote the temporal sequence of geotagged-tweet locations (latitude/longitude) of the user $u$.
  - $Z_i = (z_1, z_2, z_3, \ldots)$ represents the user’s location records at the traffic analysis zones (TAZ) scale.
  - Spatially clustering consecutive points if they were located inside the same TAZ polygon within the time threshold.
  - $C_i = (c_1, c_2, c_3, \ldots)$ represents a user’s ST clusters.

• **Step 2: place-based trip aggregation.**
  - Aggregate trips $(u, o, d, t)$ with the same origin $o$ and destination $d$ regions for all users at different temporal windows $t$ such as hourly, daily, or weekly.
  - OD matrices
Results

- On average, **24,000** daily trips were detected.
- The Pearson correlation coefficient between the survey data and the detected trips in **weekdays** is **0.91** (p-value 0.0017).
- A little lower for **weekends** **0.69** (p-value 0.05).
- Substantially lower for **Christmas** **0.59** (p-value 0.1233).

Table 1. The comparison of average morning peak-hour trips between the survey and the results detected from geotagged tweets.

<table>
<thead>
<tr>
<th>Time Window</th>
<th>Survey</th>
<th>Weekdays</th>
<th>Weekends</th>
<th>Christmas</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:00am – 5:29am</td>
<td>6.74%</td>
<td>2.31%</td>
<td>3.92%</td>
<td>5.68%</td>
</tr>
<tr>
<td>5:30am – 5:59am</td>
<td>7.12%</td>
<td>4.09%</td>
<td>5.22%</td>
<td>9.61%</td>
</tr>
<tr>
<td>6:00am – 6:29am</td>
<td>13.18%</td>
<td>8.53%</td>
<td>7.65%</td>
<td>9.61%</td>
</tr>
<tr>
<td>6:30am – 6:59am</td>
<td>12.36%</td>
<td>15.29%</td>
<td>10.63%</td>
<td>11.35%</td>
</tr>
<tr>
<td>7:00am – 7:29am</td>
<td>20.40%</td>
<td>24.80%</td>
<td>16.98%</td>
<td>12.66%</td>
</tr>
<tr>
<td>7:30am – 7:59am</td>
<td>14.92%</td>
<td>20.89%</td>
<td>23.69%</td>
<td>14.41%</td>
</tr>
<tr>
<td>8:00am – 8:29am</td>
<td>16.99%</td>
<td>15.73%</td>
<td>17.72%</td>
<td>23.58%</td>
</tr>
<tr>
<td>8:30am – 8:59am</td>
<td>8.28%</td>
<td>8.36%</td>
<td>14.19%</td>
<td>13.10%</td>
</tr>
</tbody>
</table>
OD Trips at TAZ-scale: Morning Peak (5-9 AM)
OD Trips at TAZ-scale: Morning Peak (5-9 AM)

Mean: 56.28
Median: 36.18

32 Mins
Net Flows at TAZ-scale: Morning Peak (5-9 AM)
OD Flow at County-Scale in Los Angeles Areas
GeoPrivacy!
Wrap-up

- Explored the possibility to use large-scale social media data to estimate regional OD trips
- Provide reliable estimates of temporal mobility flows on weekdays compared with the ACS
- Discover spatiotemporal flow patterns and variations in different scales
- Mining activity types using the textual parts of geotagged tweets
Thanks for your attention!

Song Gao
www.geog.ucsb.edu/~sgao