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

Song Gao[1], Jiue-An Yang[1,2], Bo Yan[1], Yingjie Hu[1], Krzysztof Janowicz[1], Grant McKenzie[1]

[1]STKO Lab, Department of Geography, University of California, Santa Barbara, CA, USA
Email: {sgao, jiueanyang, boyan, yingjiehu, jano, grant.mckenzie}@geog.ucsb.edu
[2]Department of Geography, San Diego State University, CA, USA

1. Introduction

Trajectory-based mobility research plays an increasing role in GIScience and related domains. Typically, the research results heavily depend on the quality and resolution of data that can be collected, e.g., via surveys. In travel behaviour and transportation studies, time and cost constrains are the limiting factors for the collection of large-scale individual travel behaviour data using traditional trip-diary surveys (McNally 2000). With the fast development of information and communication technologies (ICT), new data sources including GPS logs, smart card records, mobile phone data, and location-based social media have become potential alternatives or complementary approaches to study large-scale human mobility patterns and travel behaviour (Calabrese et al. 2011, Liu et al. 2012a, Yue et al. 2014). Human movement origin-destination (OD) information is of major importance in urban transportation modelling and infrastructure planning in order to optimize the use of street networks. The increasing use of social media like Twitter offers unprecedented opportunities to study individual activities, to know where users are at which time, and what they are talking about. In this work we study the reliability of detecting regional OD trips from individual geotagged tweets in comparison with survey data in a quantitative manner, and explore the spatiotemporal flow patterns extracted from social media.

We will investigate the research question of whether **OD trips mined from social media yield comparable results to expensive and labour intensive large-scale studies.** To do so, we will derive OD trips from geotagged tweets, aggregate them, and compare the results by correlating them to the American Community Survey data.

2. Data and Methods

2.1 Datasets

We collected 6.8 million geotagged tweets from 110,868 users in the Greater Los Angeles Area from December 7, 2013 to January 7, 2014. This area sprawls over five counties in the southern part of California, namely Los Angeles, Orange, San Bernardino, Riverside, and Ventura counties. We only use geotagged tweets whose sources are smart phones, including iPhone, Android, Blackberry and Windows Phones. This ensures that a geotagged-tweet reflects a person's physical location instead of a social-bot IP address or a default (hometown) location. Some initial data processing reveals that on average a user generates two geotagged tweets per day within the collection period. However, about 11000 (i.e., 10%) users tweet more than 5 geotagged tweets per day. Figure 1 shows that the distribution of the daily average number of geotagged tweets per user actually fits a truncated power function with the exponent value 1.94 and R-square 0.93. We also found the mean of individual average inter-tweeting time interval per day for all users to be 126 minutes, and the median is 79 minutes. In addition, as shown in Figure 2, the distribution of average inter-tweeting time interval per user varies from minutes to hours, and the majority (about 80%) of users is within 190 minutes per day. These preliminary analysis results help us understand the characteristics of geo-tweeting behaviours in our study area and guided us in the setting temporal-bands in the OD trip estimation algorithm discussed below.

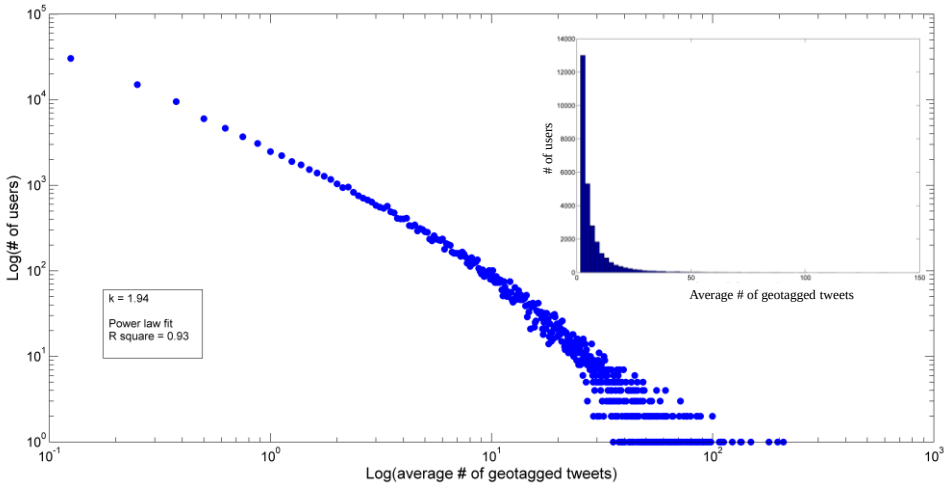

Figure 1: The log-log plot and histogram for the average number of geotagged tweets per user per day.

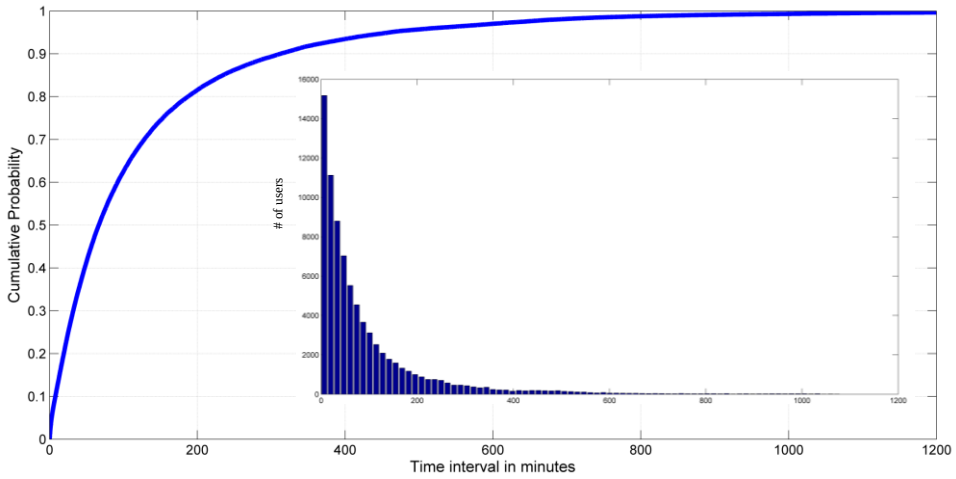

Figure 2: The histogram and cumulative probability distribution of individual average inter-tweeting time interval per day for all users

2.2 OD-Trip Estimation Approach

OD-trip estimation has been widely used for predicting travel demands in the conventional four-step model of transportation forecasting process. Our approach for estimating travel OD trips from geotagged tweets consists of two steps: *individual-based trajectory detection* and *place-based trip aggregation*.

In the first step, let $L_u = (l_{u,}^1 l_{u,}^2 ..., l_u^n)$ denote the temporal sequence of geotagged-tweet locations (latitude/longitude) of the user *u*. Then, we spatially joined all locations to the traffic analysis zones (TAZs) based on computing point-in-polygon relations which creates a second sequence $Z_u = (z_{u,}^1 z_{u,}^2 ..., z_u^n)$ of the user's location records at the TAZ scale. The spatial extent of a TAZ varies, ranging from large areas in the suburbs to as small as city blocks in central business districts. However, even for these small TAZs, the minimal extends of the bounding rectangles are about 600-1000 meters which is sufficient to filter the smartphone GPS uncertainty (typically up to 30 meters in our dataset). As s user might have multiple geotagged tweets within the same TAZ over a short period of time, these records do not contribute to the physical movements at the inter-TAZ level. Therefore, we spatially clustered those consecutive points if they were located inside the same TAZ polygon within the time threshold Δt which we set to 4 hours based on the knowledge from aforementioned inter-tweeting time analysis. The new sequence of TAZ clusters can be represented as

$C_u = (c_{u,}^1 c_{u,}^2 ..., c_u^n)$. Next, individual trips can be extracted as the paths between two consecutive clusters in different TAZs for any given user.

In the second step, we aggregate trips *(u, o, d, t)* with the same origin *o* and destination *d* TAZ regions for all users together at different temporal windows *t* such as hourly, daily, or weekly. The result is an asymmetric OD matrix whose element T_{ij} represents the total number of detected trips from the origin *i* to the destination *j* regions starting within a time period.

3. Results and Evaluation

3.1 Extracting Peak-Hour OD Trips at the TAZ-scale

The proposed OD-trip estimation approach has the flexibility to detect dynamic inter-TAZ mobility flows at different temporal windows. In order to compare the detecting results with 2008-2012 American Community Survey (ACS)[1] data for evaluation, we aggregate OD trips extracted from geotagged tweets in 30min time windows based on the leaving time of each trip in morning-peak hours 5am-9am as shown in Table 1. On average, we detected about 24000 daily trips and 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), and substantially lower for Christmas Day 0.59 (p-value 0.1233). The higher correlation between weekday trips and the survey at such a significance level than weekends and holidays meets our expectation since weekday trips have more regular patterns. Furthermore, we analyzed the trip-length distribution and found that it roughly follows a distance-decay distribution (Figure 3c and 3d) and the average length is about 56 km (35 miles). If we convert the trip distance into time using the local speed limit of 65 miles, the average time of all morning trips is about 32 minutes and very close to the survey data results of 29 minutes. All these results indicate that our OD-trip detection algorithm corresponds well with ACS data and can capture the overall characteristics of mobility flows in the study area using a big-data-driven approach.

In addition, the geovisualization of morning or evening peak-hour trips and netflow (inflow-outflow) patterns help us to identify the directed-flow changes in suburbs and downtown areas, as well as to better understand urban transportation dynamics (see Figure 3). Advanced spatiotemporal patterns and the linkages to land-use types can also be analyzed using the methods proposed by Guo et al. (2012) and Liu et al. (2012b) for further studies.

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

Time Window	Survey	Weekdays	Weekends	Christmas
5:00am – 5:29am	6.74%	2.31%	3.92%	5.68%
5:30am – 5:59am	7.12%	4.09%	5.22%	9.61%
6:00am – 6:29am	13.18%	8.53%	7.65%	9.61%
6:30am – 6:59am	12.36%	15.29%	10.63%	11.35%
7:00am – 7:29am	20.40%	24.80%	16.98%	12.66%
7:30am – 7:59am	14.92%	20.89%	23.69%	14.41%
8:00am – 8:29am	16.99%	15.73%	17.72%	23.58%
8:30am – 8:59am	8.28%	8.36%	14.19%	13.10%

[1]Search for Table S0801: *Commuting Characteristics by Sex* via http://www.census.gov/acs/www/

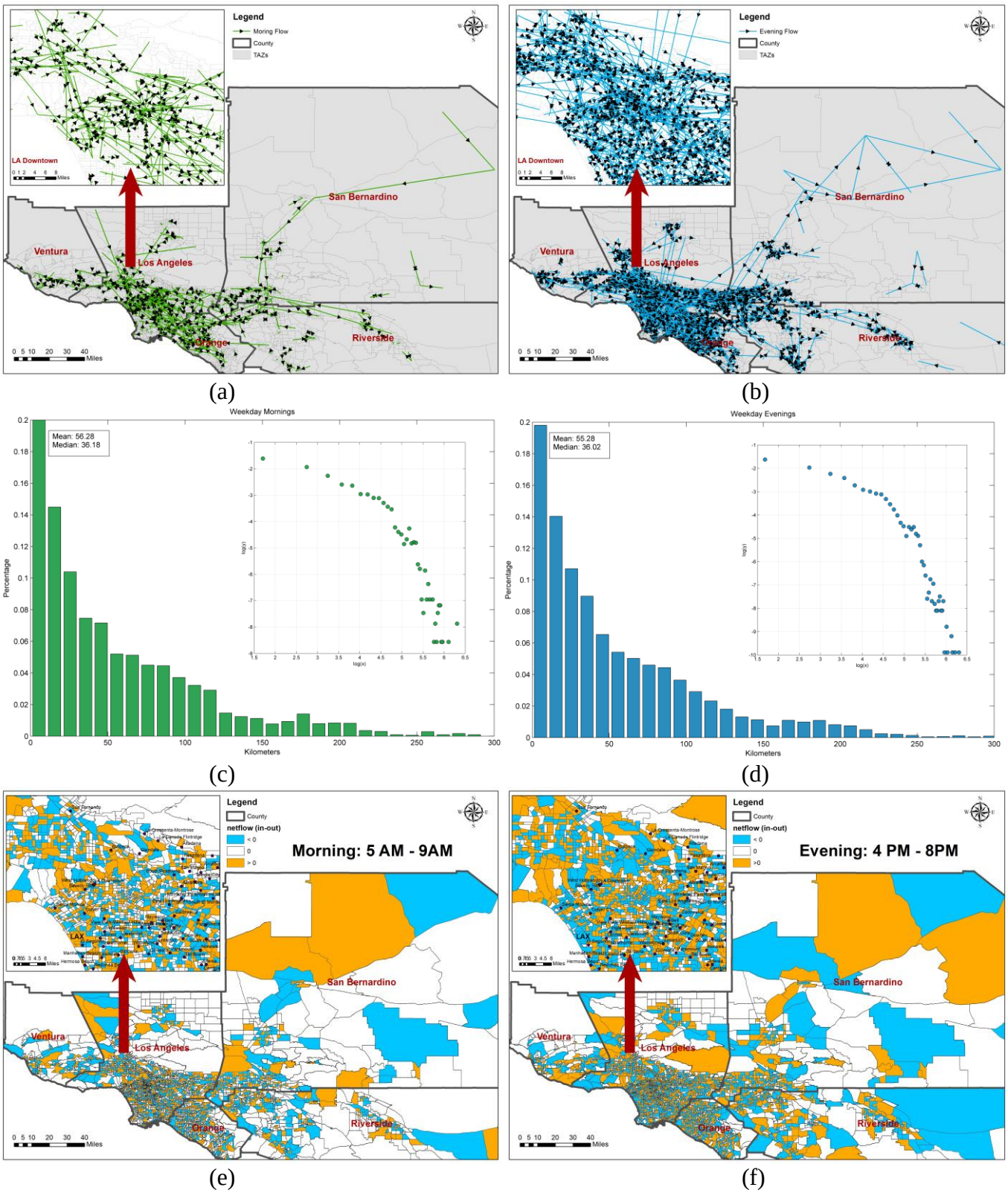

Figure 3: The spatial distributions of detected OD trips. (a) morning-peak directed pattern; (b) evening peak directed pattern; (c) morning-trip distance distribution; (d) evening-trip distance distribution; (e) morning netflow; (f) evening netflow at the TAZ scale.

3.2 OD Trips at the County-level

We spatially aggregated the OD trips from TAZ to the county level. Figure 4 shows the spatial distribution of detected daily OD trips for five counties. Surprisingly, even for such large-scale inter-county and intra-county flow patterns, our results show a perfect rank

matching with the ACS data, which demonstrates that Orange-Los Angeles has the largest inter-county mobility flows, San Bernardino-Los Angeles and Ventura-Los Angeles ranks the second and the third. But most trips occur inside the same county. Moreover, we also detected some regular trips between Riverside and four other counties, which weren't reported in the survey. Such analysis shapes well for the regional-flow patterns and spatial interaction structure, which is beneficial for regional transportation planning.

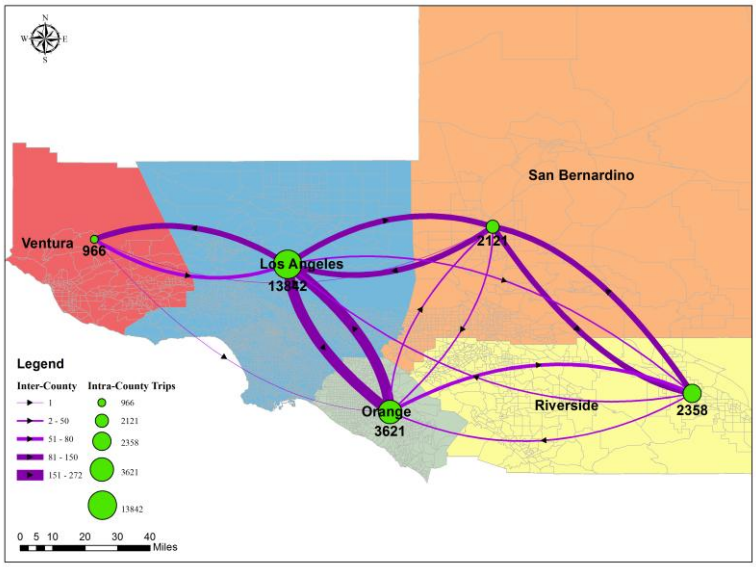

Figure 4: The spatial distribution of detected daily OD trips for five counties.

4. Conclusion and Future Work

In this research, we explored the possibility to use large-scale social media data to estimate regional OD trips. Our case study demonstrates that the proposed approach can provide reliable estimates of temporal mobility flows on weekdays compared with the community survey data and also help discover spatiotemporal flow patterns and variations at varying scales. As a next step we would like to investigate how to mine for activity types using the textual parts of tweets.

References

Calabrese F, Di Lorenzo G., Liu L and Ratti C. 2011. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4), 36-44.

Guo D, Zhu X, Jin H, Gao P, and Andris C. 2012. Discovering Spatial Patterns in Origin-Destination Mobility Data. *Transactions in GIS*, 16(3), 411-429.

McNally, MG, 2000. The activity-based approach. In: Hensher, D.A., Button, K.J. (Eds.), *Handbook of Transport Modelling*. Pergamon, Amsterdam, 53–70.

Liu Y, Kang C, Gao S, Xiao Y, and Tian Y. 2012a. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14(4), 463-483.

Liu Y, Wang F, Xiao Y, and Gao S. 2012b. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning*, 106(1), 73-87.

Yue Y, Lan T, Yeh AG and Li, QQ, 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society*, 1(2). 69–78.