Urban Travel Demand Analysis for Austin TX USA using Location-based Social Networking Data

Peter J. Jin, Ph.D. (Corresponding Author)
Department of Civil, Architectural, and Environmental Engineering,
The University of Texas at Austin,
1616 Guadalupe St., Ste. 4.202, Austin, TX 78701
Phone: 1-512-232-3124
Email: jjin@austin.utexas.edu

Fan Yang, Ph.D. Candidate
Department of Civil and Environmental Engineering,
University of Wisconsin-Madison
1241 Engineering Hall 1415 Engineering Drive, Madison, WI 53706
Phone: 1-608-609-3386
Email: fyang29@wisc.edu

Meredith Cebelak, Ph.D. Candidate
Department of Civil, Architectural, and Environmental Engineering,
The University of Texas at Austin,
1616 Guadalupe St., Ste. 4.202, Austin, TX 78701
Phone: 1-512-232-3124
Email: mdegaspari@utexas.edu

Dr. Bin Ran, Ph.D., Professor
Department of Civil & Environmental Engineering, University of Wisconsin-Madison,
1415 Engineering Drive, Madison, WI 53706, USA
Phone: 1-608-262-0052 Fax: 1-608-262-5199
Email: bran@wisc.edu
and
School of Transportation, Southeast University
No.2 Si Pai Lou, Nanjing 210096, China

C Michael Walton, Ph. D.
Department of Civil, Architectural, and Environmental Engineering,
The University of Texas at Austin
1 University Station C1761, Austin, TX 78712-0278
Phone: 1-512-471-1414
Email: cmwalton@mail.utexas.edu

Submitted for Presentation and Publication
to the 92nd Transportation Research Board Meeting
Submission Date: Aug. 1st, 2012

4870 Words + 2 Table + 8 Figures = 7370 Words

ABSTRACT

The location-based social networking (LBSN) is a location-sensitive service interactively carried out by users with mobile devices, such as smart phones, to “checkin” with the “venues” reflecting their daily activities. With its increase popularity and sophistication, the location-based social networking (LBSN) data have emerged as a new data source for studying urban travel demand. Comparing with traditional Origin-Destination (OD) estimation method such as survey based or traffic count based methods, LBSN data has the potential to provide OD estimation with much higher temporal resolution at much lower cost. In this paper, the Foursquare LBSN data was used to analyze the OD demand for the urban area near Austin, Texas, USA. A gravity model with two-regime friction factor functions is proposed to estimate the O-D matrix. The proposed methods are calibrated and evaluated against the ground truth O-D data from CAMPO (Capital Area Metropolitan Planning Organization). The results illustrate the promising potential of using LBSN data for urban travel demand analysis and monitoring.

KEY WORDS: Origin-Destination estimation, Location-based social networking, gravity model

INTRODUCTION

Origin-Destination (O-D) matrices are key inputs for urban transportation planning and traffic operational applications, and can be classified into three main categories: the traditional survey-based methods, the traffic-counts based methods*(13)*, and the positioning technology based methods. Traditional survey-based O-D estimation methods, such as the household survey and the roadside survey method, can be time-consuming, expensive to undertake, are usually confined to a limited number of households, are not conducted on a continuous basis, and can cause distraction to traffic (roadside survey). Additionally, the sample size issue and sample bias can render the reliability of this type of method. Traffic-count based O-D estimation methods*(4, 12)*, use the link traffic counts from traffic detectors to update an existing O-D matrix. While there are advantages of significant time and labor when compared to the survey-based methods, these methods rely on an existing metering infrastructure, which may be expensive to install or maintain, and may have coverage issues impacting the reliability of certain parts of the O-D matrix to not be updated (e.g. arterial roads and local streets without detector coverage). Traffic-count based methods also depend on an existing O-D matrix that may not accommodate changes in land use or transportation network.

In recent years, positioning technologies such as GPS, cell phone, and Bluetooth technology have become important data sources for traffic flow monitoring, traveler information provision, and advanced traffic and demand management. Similar to the survey based method, the GPS-based methods*(15)* also suffers from sample size issues and sampling bias. Other limitations include privacy concerns require user consent on disclosing collected GPS trajectory data and time and labor costs associated with membership, contracts, services, and potential incentives need to promote participation in the GPS data collection when large sample sizes are required. Wireless location technologies (WLT) available from wireless carriers to derive the O-D matrix have been recently studied*(2, 3, 11)*. WLT uses cellphone positions that are collected by tracking wireless signal transition events when cellphones cross the boundaries of virtual regions in the cellular network. The effectiveness of this method is limited by high spatial resolutions that can only be triggered when cellphone is on a call and by other events, which can only be triggered during off-call periods. These limitations do not provide high enough spatial resolution for meaningful O-D estimation resulting in the collection of only partial trajectories of cellphone users' traveling activities. Additionally, matching issues between the data collection areas (e.g. cellular cells) using cellphones and Traffic Analysis Zone (TAZ)s can further reduce the accuracy of trip estimation. Bluetooth technology is another emerging technology to collect travel demand data. Barceló et al.*(1)* investigated the use of Bluetooth device for travel time and O-D data collection. With a dense network of Bluetooth reading devices, the vehicle trajectories within a traffic network can be collected via the Bluetooth IDs (a number assigned during the device manufacturing, eliminating any privacy issues). One key limitation is the sampling rate of typically 1% to 5% penetration*(1)*, which is attributed to the ability to turn off Bluetooth functions in many mobile devices.

Technological advances have allowed smartphones and tablets with LBS (Location based service) features to become affordable and accessible to people of various income levels. Concurrently, the fast development in social networks led by Facebook and Twitter also attracts a tremendous amount of users that actively updated their personal activities online including their locations. The combination of social networking and LBS results in the location-based social networking (LBSN) services. Recently, researchers began to conduct data mining of social networking sites to study the spatial pattern of cellphone user behavior. Cheng*(9)* proposed a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets. Backstrom*(5)* utilized the network of associations between Facebook users to predict the location of an individual based on the location of his/her friends. Cheng et al.*(8)* studied the human mobility patterns by analyzing the social networking data.

With respect to travel demand analysis, specifically O-D estimation, social networking data has some unique advantages over the GPS, cell phone, and Bluetooth data. Social networking services allow users to share their locations with their friends by using mobile applications on a smartphone or tablet. The applications use built-in GPS to generate an accurate trajectory of the user's traveling activities. Announcing arrival at a location through such application is called "check-in," which is associated with particular venues (e.g., a restaurant). For a successful "check-in," the user need to confirm the name of a venue ensuring the data quality for trip origins and destinations for a travel demand analysis. The penetration rate of social networking service is growing at a rapid pace providing a sample size much larger than other methods. More importantly, the data is updated in real-time and is a low cost option due to the lack of auxiliary data collecting devices.

This paper proposes a novel O-D estimation method that utilizes the LBSN check-in data provided by Foursquare. This rest of the paper is organized as follows. Section 2 introduces the background of LBSN services. The data processing procedures and methods are introduced in Section 3. In section 4, a preliminary analysis is conducted for the characteristics of check-ins collected. Section 5 proposes a gravity model based method to derive O-D matrix from the check-in data. Section 6 calibrates and evaluates the proposed model using the CAMPO O-D matrix. The optimal setting of the model regarding friction function and venue classification is also identified. Then, in Section 7, the calibrated model is applied to study the bihourly and daily O-D patterns. The final section concludes this paper.

BACKGROUND

Overview of location based social network (LBSN)

Social networking is an online service building and reflecting social relations among people who share interests and/or activities. Facebook, launched in February 2004, is currently the largest social networking site in the world with more than 600 million active users. LBSN, also called LBS, refers to a special social networking service that uses GPS features to locate users and allows members of the communities to broadcast their locations and activities through their mobile devices. Due to its creative ways of interacting with customers, venues actively participate in LBSN services, and the estimated number of LBSN websites worldwide has reached more than a hundred by 2011*(17)*. The first commercially available LBSN system was Dodgeball *(7)* founded in 2000 and later acquired by Google in 2005. The current leading LBSN service is Foursquare. Foursquare users can "check-in" at public places through Foursquare website, text messages, or mobile applications. They are then awarded points and sometimes "badges". The company reported it had attracted 10 million registered users by June 2011, with about 3 million "check-ins" per day*(16)*. Among the users, 50 percent came from US and 50 percent were male.

Characteristics of Foursquare data

Due to its high popularity and comprehensive functionality, Foursquare is selected as the main data source for this study. Foursquare provide developers the access to their data through OAuth2 (Open Authorization) API (Application Programming Interface). Data that can be exported from the interface include the user and check-in statistics from each venue and also the detailed information each venue such as location, coordinates, and type. The penetration rate of Foursquare is currently the highest among all existing LBSN systems, which provides the largest LBSN sample size for travel demand analysis.

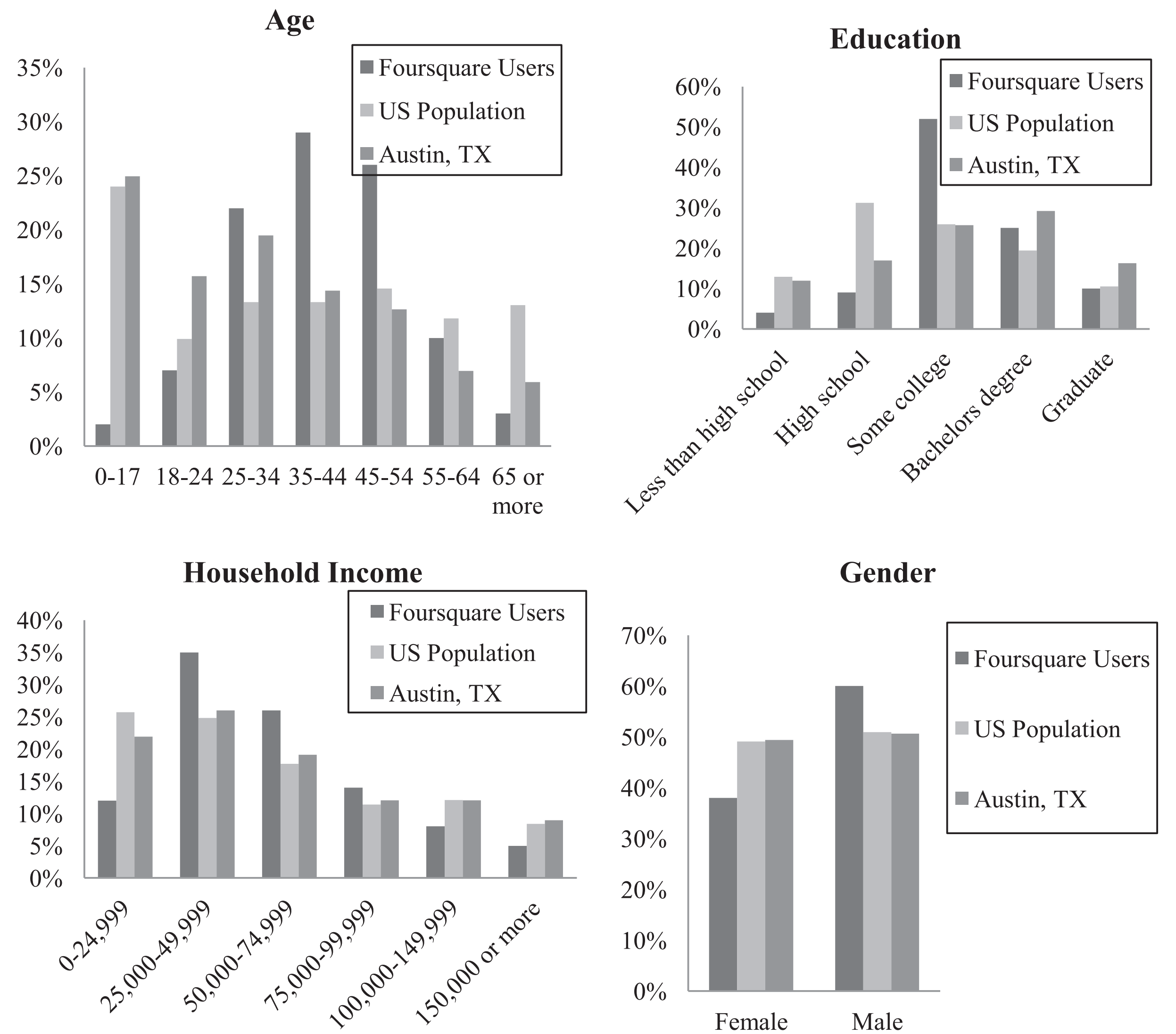


FIGURE 1 Demographic distribution of Foursquare users v.s. US drivers/population.

Figure 1 shows the demographic characteristics of foursquare users*(14)*. The gender information comes from the worldwide statistics of Foursquare users in May, 2011. The other demographic information such as the age, education, and household income is estimated by Google's DoubleClick Ad Planner in February, 2012. Ad Planner uses a hybrid methodology combining sample user data from various Google products and services and direct-measured site-centric data. The results should be very close to the real Foursquare's user data*(14)*. In Figure 1, the corresponding demographic distribution for U.S. population*(19)*, and Austin city population*(20, 21)* is also provided. It can be observed that the demographic distribution of Foursquare users generally matches that of U.S. drivers or general population. Inconsistencies include the limited coverage over population greater than 55 years old, less coverage over female travelers, population with educational levels less than high school, and low-income households. These population groups may have less number of trips made and miles travelled than the other population groups although the lack of coverage over population group with lower educational level may be caused by the IT knowledge requirement for using LBSN services.

FOURSQUARE DATA

The data used in this paper include the GIS data of the Austin central area and O-D data from CAMPO, and the hourly "check-in" statistics Foursquare. The GIS data are used to define the boundary of the traffic analysis zone. The official O-D matrix is used as the ground truth data for model calibration. The Foursquare data are collected through a Java program which runs every hour to obtain the numbers of check-in increases of all the venues within the research area. The Austin central area is a representative urban area and has high coverage of Foursquare venues and users. The selected study area is bounded by Pecan Street, Lime Creek Road, SH 71, SH 45, and Caldwell Lane, as shown in Figure 2. The official Traffic Analysis Zone (TAZ) system developed by CAMPO is used for this study.

CAMPO Origin-Destination (O-D) matrix data

The ground truth O-D matrix data were derived from CAMPO's most recent analysis completed in 2010*(22)*. The CAMPO procedure uses an Atomistic Trip distribution model to obtain the O-D matrices, which is a spatially disaggregated gravity model that allocates intrazonal trips using radius data for each TAZ and the trip length frequency distributions determined from the Travel Surveys. The CAMPO matrices contain the modeled daily trip tables for 17 trip purposes, which are listed as follows:

- Home Based Work Person Trips Direct
- Home Based Work Person Trips Strategic
- Home Based Work Person Trips Complex
- Home Based Non-work Retail Person Trips
- Home Based Non-work Other Person Trips
- Home Based Non-work Primary Education Person Trips
- Home Based Non-work University/College Person Trips
- Home Based Non-work UT-Austin Education Person Trips
- HBNW/NHB (Non-work) Airport Person Trips
- Non-home Based Work-related Person Trips
- Non-home Based Other Person Trips
- Non-home Based External Commuter/Visitor Vehicle Trips
- Commercial Truck/Taxi Vehicle Trips
- External Local Auto Vehicle Trips
- External Local Truck Vehicle Trips
- External Through Auto Vehicle Trips
- External Through Truck Vehicle Trips

For this study, the first 13 categories were combined to attain the total trips taken. All Home Based Work trips were combined into one category. Home Based Non-work Retail and Home Based Non-work Other will be the other categories included for this study.

Since these CAMPO O-D matrices were obtained only two years before the Foursquare data are collected, the difference between the O-D matrix derived from CAMPO and Foursquare should be minimal if the latter one can provide accurate results.

The Foursquare data collection

The data collection program utilized the latest version (Version 2) of Foursquare API*(18)*. The API endpoints provide methods for accessing the venue real-time status through a URL. A computer program is written to obtain the latest status of all venues in the Austin area every hour. The program started running since June 11, 2012. Data from the first three weeks till July 2, 2012 are used for model calibration and analysis. An additional one-day data from July 18th are used to study the bi-hourly travel pattern

PRELIMINARY ANALYSIS OF THE CHECK-IN DATA

In this section, a preliminary analysis is conducted on the characteristics of the check-in data by investigating how the check-ins distribute in space and when people use Foursquare LBSN service. The locations of the 19,170 venues are represented using a dot in Figure 2a. The venues are more densely distributed in central area.

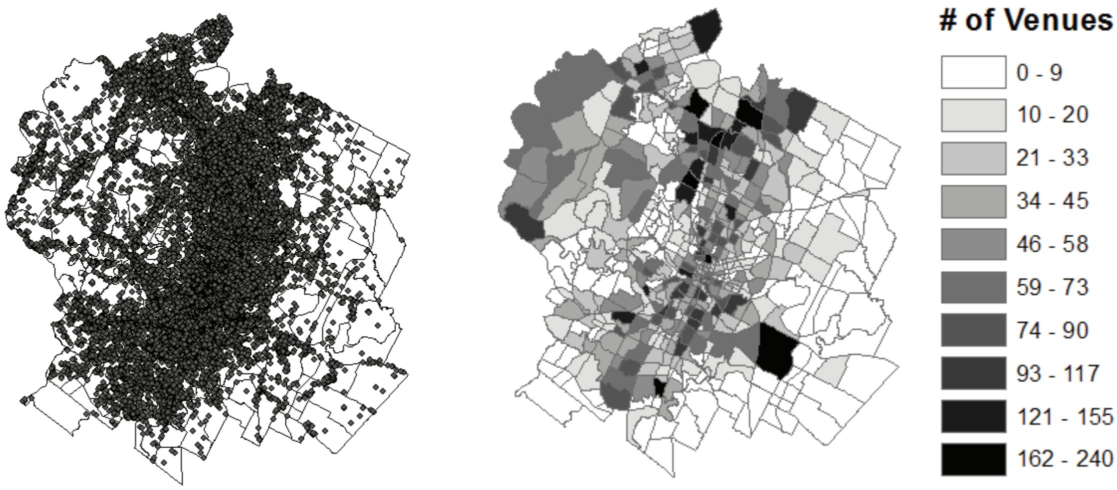


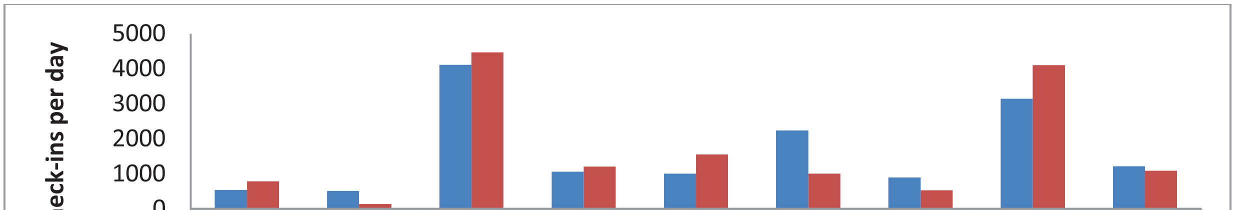
FIGURE 2 Foursquare venues locations and their categorical distribution among TAZs.

TABLE 1 Austin Area Foursquare Venue Characteristics

Category	# of Venues	Percentage	# of Check-ins	Percentage	Avg. Check-ins
Colleges & Universities	719	3.8%	367866	5.5%	512
Shops & Services	5187	27.1%	1389636	20.9%	268
Food	2809	14.7%	2021897	30.4%	720
Nightlife Spot	547	2.9%	669712	10.1%	1224
Arts & Entertainment	592	3.1%	324249	4.9%	548
Travel & Transport	792	4.1%	479305	7.2%	605
Professional & Other Places	4679	24.4%	832999	12.5%	178
Great Outdoors	1596	8.3%	278065	4.2%	174
Residences	711	3.7%	182825	2.7%	257
Unclassified	1538	8.0%	102692	1.5%	67

Table 2 shows several statistical results of the venues aggregated by different categories. The “Shops & Services” and “Professional & Other Places” categories have the largest number of venues. However, “Food” attracted the most total check-ins. The “Nightlife Spot” category had the highest average check-ins. The “Unclassified” venues totaled 1538, but only had 1.5% of the total check-ins which is a small enough percentage to be eliminated in our travel demand analysis.

To investigate the temporal characteristics of how people use the Foursquare LBSN services, the number of check-ins are aggregated by each venue category for weekdays and weekends and for different hours of the day.



	Arts & Entertainment	Colleges & Universities	Food	Great Outdoors	Nightlife Spot	Professional & Other Places	Residences	Shops & Services	Travel & Transport
Weekday	544	517	4110	1062	1011	2242	896	3145	1216
Weekend	791	138	4464	1208	1553	1012	534	4099	1089

FIGURE 3 Weekday versus weekend check-in pattern.

Figure 3 shows the check-in patterns in weekdays versus weekends. More check-ins can be found at “Colleges and Universities,” “Professional & Other Places,” “Residences,” and “Travel & Transport” locations during the weekdays than during the weekends; while the check-in frequency for the “Arts & Entertainment,” “Shops & Services,” and “Nightlife Spots” locations is much lower during the weekdays than the weekend. “Great Outdoors” and “Food” attracts slightly more check-ins on the weekend.

5. METHODOLOGY

5.1. Trip Distribution Modelling

The base model used in this study for trip distribution is a gravity model based formulation proposed in *(23)*.

$$P_i = \sum_n \sigma_p x_{in} \text{ , } i = 1, 2, \ldots, 77$$

$$A_j = \sum_n \sigma_a x_{jn} \text{ , } j = 1, 2, \ldots, 77$$

$$\widehat{T}_{ij} = P_i \frac{A_j F_{ij}}{\sum_j A_j F_{ij}} \tag{1}$$

Where

x_{in}: Check-ins for venue type *n* in origin zone *i*

x_{jn}: Check-ins for venue type *n* in destination zone *j*

σ_p: The ratio of trip production to Foursquare check-ins.

σ_a: The ratio of trip attraction to Foursquare check-ins.

$\hat{T}_{ij}$: Trips made between origin zone *i* and destination zone *j*.

P_i: Production from zone *i*

A_j: Attraction of zone *j*

F_{ij}: Friction function

The two adjustment ratios, σ_p and σ_a, are used to adjust the estimated trip production and attractions to cope factors such as Foursquare penetration rate over total travelers and venues in a TAZ and Foursquare user's willingness to check in. The friction factor function (F_{ij}) uses the following three formulations including the linear function, the Negative Exponential function*(10)* and the gamma function as the following.

$$\text{Linear: } F_{ij} = \alpha + \beta d_{ij} \tag{2}$$

$$\text{Negative exponential: } F_{ij} = \alpha e^{-\beta d_{ij}} \tag{3}$$

$$\text{Gamma: } F_{ij} = \alpha \cdot d_{ij}{}^{\beta} e^{\gamma \cdot d_{ij}} \tag{4}$$

Where

α: a positive scaling factor controlling the overall range of function values

β : a positive or negative constant value which affects the distribution of shorter trips

γ : A parameter of transport friction related to the efficiency of the transport system between two locations. γ is always negative and can affect the distribution of longer trips.

d_{ij}: The Manhattan distance between the centroids of origin zone i and destination zone j in miles.

Furthermore, in our preliminary analysis, it is found that the impedance as indicated by CAMPO OD data has different trends for short-distance trips and long-distance trips. Therefore, the friction factor function is revised to become a two-regime function.

$$F_{ij}(d_{ij}) = F_{ij}^{(s)}(d_{ij}) I_{d_{ij} \le T_d} + F_{ij}^{(l)}(d_{ij}) I_{d_{ij} > T_d} \tag{5}$$

where $I_{[clause]}$ is an indicator function for a logic clause ($I_{[clause]} = 1$ if the clause is true; otherwise, $I_{[clause]} = 0$), the superscript *s* and *l* indicates short-distance trip regime and the long-distance trip regime, respectively, and T_d is the threshold to determine the regime. Model parameters in both regimes and T_d needs to be calibrated using ground truth OD matrix. Selecting from Equation 2 to 4, 9 different combinations of friction function types are explored.

5.2. Model Calibration

A total of 15 different models are to be calibrated and compared with the proposed five venue classification methods and three friction functions. The singly-constrained method is implemented for trip balancing. The genetic algorithm is used to obtain the parameters for each model, and the objective function is to minimize the MAE (Mean Absolute Error) between the modeled O-D matrix and the ground truth O-D matrix. We use the coincidence ratio to evaluate the performance of the 15 models. The coincidence ratio measures the percent of the area that "coincides" for the two curves or distributions to compare*(6)*. The daily average Foursquare check-in data for the entire study period are used as the inputs to the gravity model to match the characteristics of the CAMPO O-D matrix. In evaluating the fitness of the model, we compare the percentage of trips in each trip length interval for CAMPO's survey trips and the predicted trips. The trip length interval is defined as 0.25 miles in our study, and the maximum trip length is 6 miles long which results in 25 intervals. The coincidence ratio is defined as the following

$$CR = \frac{\sum_i \min(p_i^M, p_i^O)}{\sum_i \max(p_i^M, p_i^O)} \tag{6}$$

Where p_i^M : the percentage of trips in interval *i* in the predicted trips from Foursquare data.

p_i^O : the percentage of trips in interval *i* in the survey trips from CAMPO.

CR takes the value in [0, 1]. When CR = 0, the two distributions are completely different; while when CR = 1, the two distributions are identical. In this study, higher coincidence ratio between Foursquare results and CAMPO results indicates a better model. The coincidence ratios for the 15 models tested are listed in Table 4.

TABLE 2 Coincidence Ratios for the Different Combination Friction Factor Models

		Long Trip		
		Linear	Neg. Exp.	Gamma
Short Trip	Linear	0.31	0.85	0.69
	Neg. Exp.	0.32	0.32	0.70
	Gamma	0.31	0.70	0.70

Table 4 listed the calibration results for two-regime friction factor function. It is found that linear model has issues for describing the trip friction characteristics of long trip; while negative exponential function may not yield satisfactory results for short trips. The best overall results is achieved by using linear model for short trips and negative exponential for long trips. The best friction factor model with parameters is as the following.

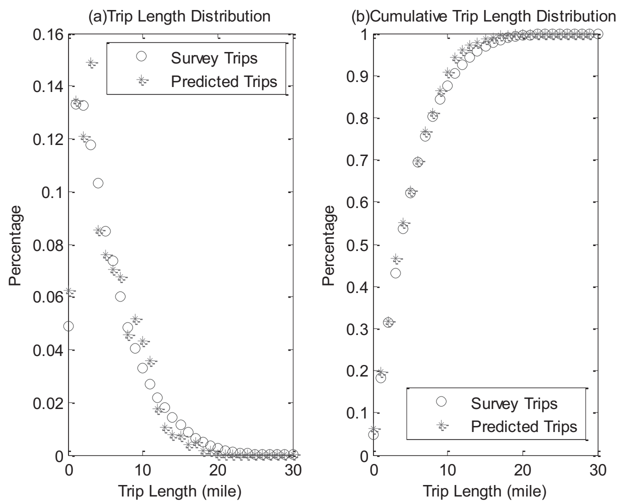
$$F_{ij}\left(d_{ij}\right) = \begin{cases} 0.0029 + 0.0030 \cdot d_{ij}, & d_{ij} \le 8.26 \\ 0.7710 \cdot e^{-0.0035 d_{ij}} & d_{ij} > 8.26 \end{cases} \tag{7}$$

6. MODEL EVALUATION AND APPLICATION

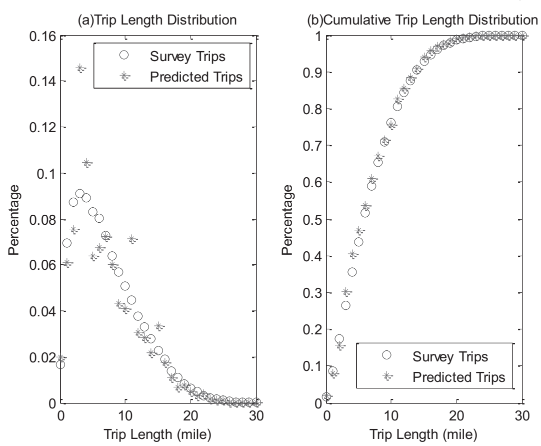
6.1. Model Evaluation

In order to compare the calibrated Foursquare O-D matrix with CAMPO ground truth matrix three different approaches are used including the comparison of the trip length distributions, the zonal O-D flow patterns and the zonal trip generation/attraction heat maps.

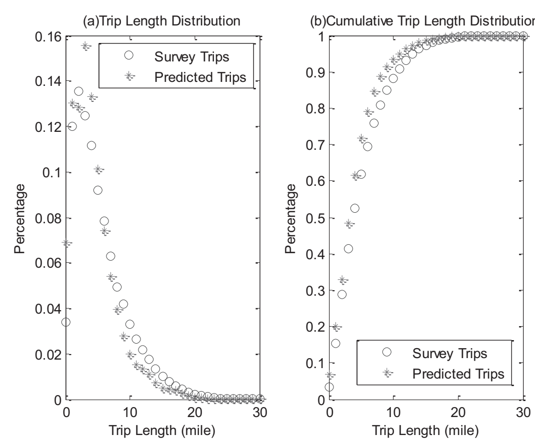
The trip length distribution curves are the same as those defined for calculating the coincidence ratio, whose curves can illustrate how well the model output matches the ground truth data. Figure 4a demonstrates the comparison results between the survey and predicted trips. Relatively consistent matching can be observed, although, Foursquare data slightly underestimates the number of long trips. Figure 4b illustrates the cumulative trip length distributions. It can be observed that the trips predicted by Foursquare data accumulate faster for shorter trips than the CAMPO trips. In general, the two curves follow the same paths and data points are located within close proximity to one another in both plots, demonstrating the feasibility of the proposed method.



(A) Total Trips



(B) Home-based Work Trips



(C) Home-based Retail Trips

FIGURE 4 Trip length distributions.

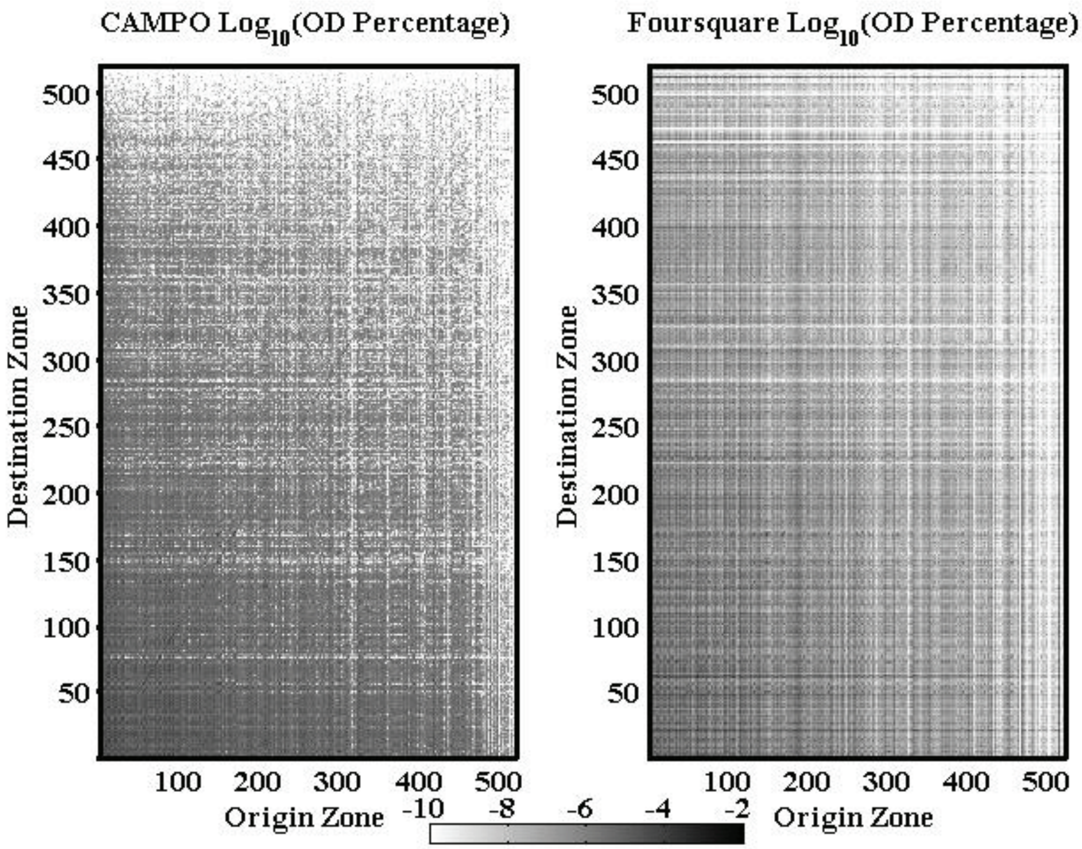


FIGURE 5 Zonal O-D pattern comparison.

By sorting all TAZs based on its CAMPO zonal attractions, Figure 5 compares the O-D flow pattern between the CAMPO O-D matrix and the Foursquare matrix. The zonal flow pattern can be regarded as the visualization of the O-D matrices. The horizontal axis represents the origin zone, and the vertical axis is the attraction rates in numerical order. Each grid (*i*, *j*) in the diagram displays the adjusted O-D flow intensity I_{ij} from zone *i* to zone *j* defined as the following.

$$I_{ij} = \log_{10}\left(\frac{\hat{T}_{ij}}{\sum_i \sum_j \hat{T}_{ij}} \right) \tag{8}$$

Dark color represents high O-D flow, and light color suggests low O-D flow. As shown in Figure 5, the Foursquare flow pattern and the CAMPO flow pattern exhibit similar characteristics in terms of the frequency distribution. Slight inconsistency can be found for OD pairs between the lower 200 origin zones and higher 200 destination zones where Foursquare slightly underestimated the OD flow. The zonal trip productions and attractions obtained from the CAMPO O-D matrix and the modeled matrix are color coded in Figure 6. Larger values were represented by darker colors and smaller values by lighter colors. O-D pairs with zero trip counts in CAMPO data are identified as blank areas in the O-D flow figure. Analysis between the ground truth data and the modeled data suggest a consistence between the CAMPO O-D and modeled matrices. However, some inconsistencies can be found within the production heat maps. These inconsistencies can be attributed to the relatively lower number of check-in of

Foursquare data at residences, which may also explains the slight inconsistencies in Figure 6. Despite the inconsistencies, in general, the above comparisons indicate significant similarity between the O-D matrix generated from the model and the CAMPO O-D matrix.

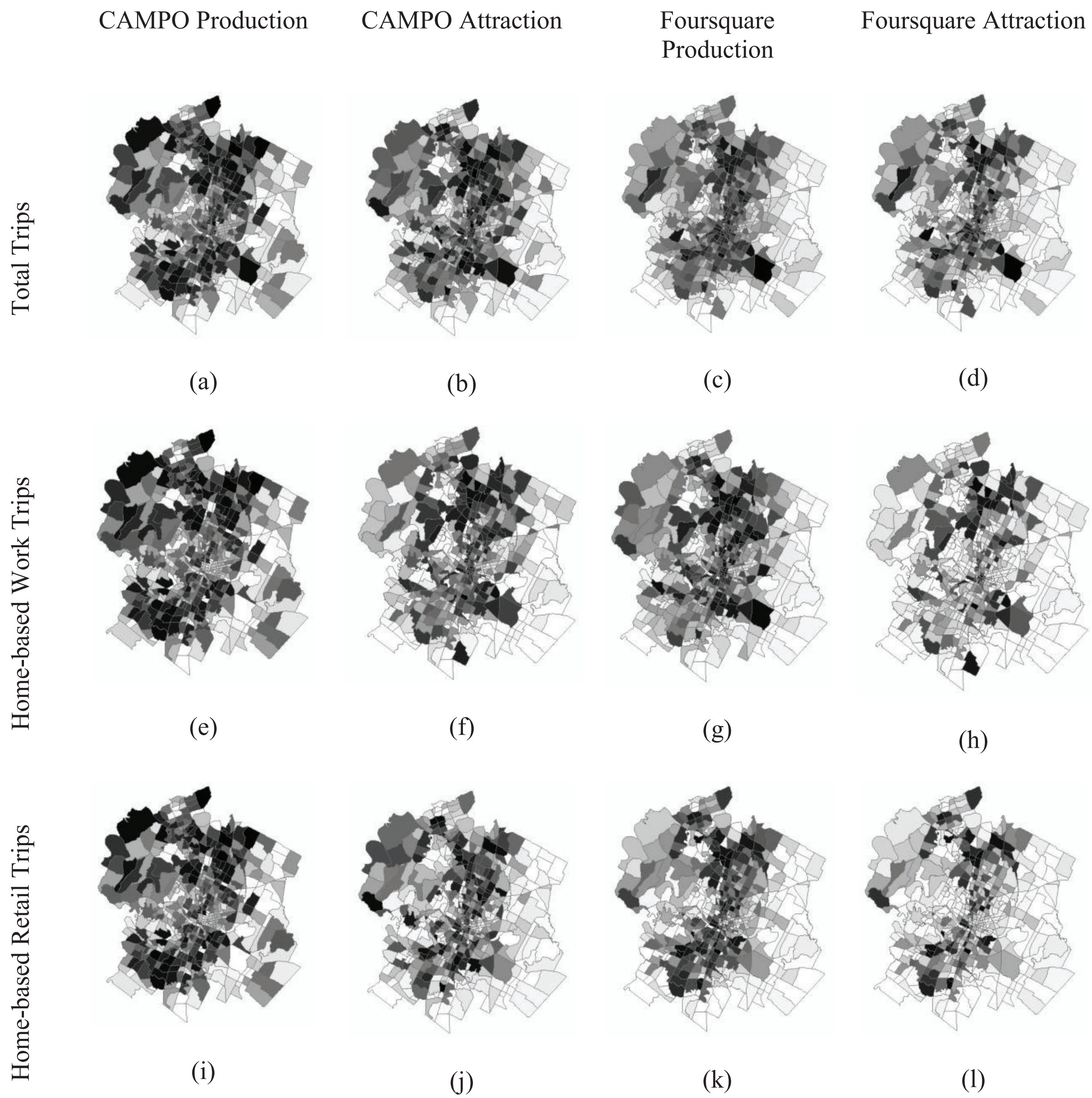


FIGURE 6 Zonal production/attraction heat maps.

6.2. Model Applications

In this section, the calibrated gravity model is applied to the daily Foursquare check-ins statistics to analyze the bi-hourly travel demand pattern in the Austin area. Admittedly, since the CAMPO data represents static average daily travel patterns for workdays, the calibrated model is only accurate for the average workday O-D matrix. But to explore its potential as a dynamic OD monitoring tool, the model is extended to a bi-hourly temporal resolution to investigate whether the resulting patterns are consistent with the empirical experience. We use the OD demand data from Foursquare for Wednesday, July 18th, 2012, which is not within the three-week range for the model calibration. Both the zonal attraction heat map (Figure 7) and the OD pattern diagrams (Figure 8) similar to Figure 5 are provided for each hour.

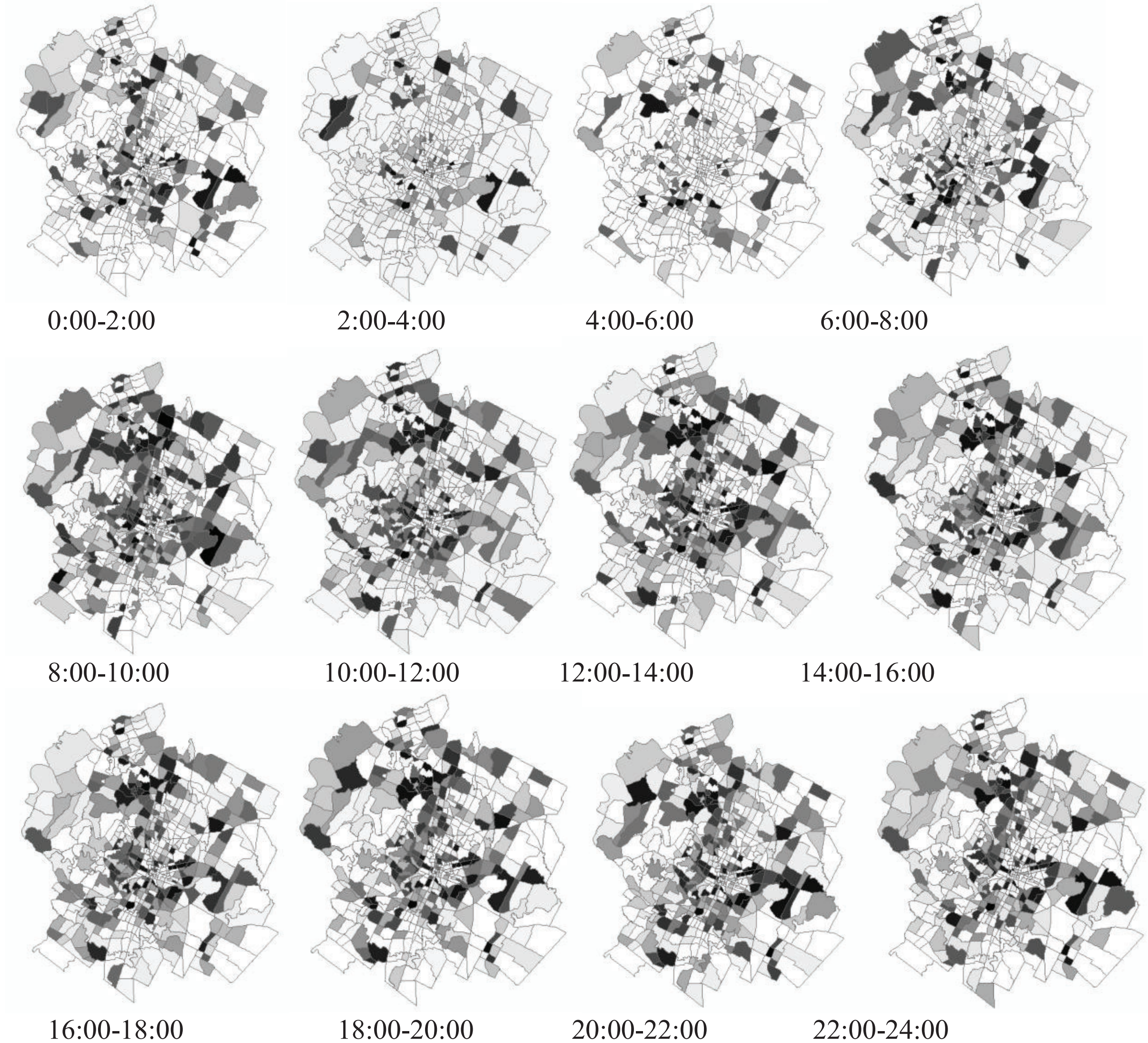


FIGURE 7 Bihourly zonal trip attraction heat map for all trip purposes.

The bi-hourly zonal attractions pattern fits well with the expected daily activities in the Austin area. The trip intensity during the night time is less than that during the day time. The activity level reaches the minimal between 2 to 6 am. The morning activity peak is around 8-10 am. Around 12 to 2 pm, a noon activity peak is observed which is consistent with the lunch time. Another activity peak can be found around 6 to 8 pm, when most dinning, shopping, and entertainment activities may occur. Foursquare data exhibits good coverage over the night time activities in Austin and some residual travels can also be observed between 0 to 2 am indicating people returning home from their late night activities. Figure 8 illustrates the distribution of trips among different zones. TAZs are still in the descend order CAMPO zonal attraction. OD flows distribute more evenly during the morning than the afternoon. During activity peak periods, trips may distribute more intensively between high-attraction zones, which can create pressure on the surrounding transportation infrastructures.

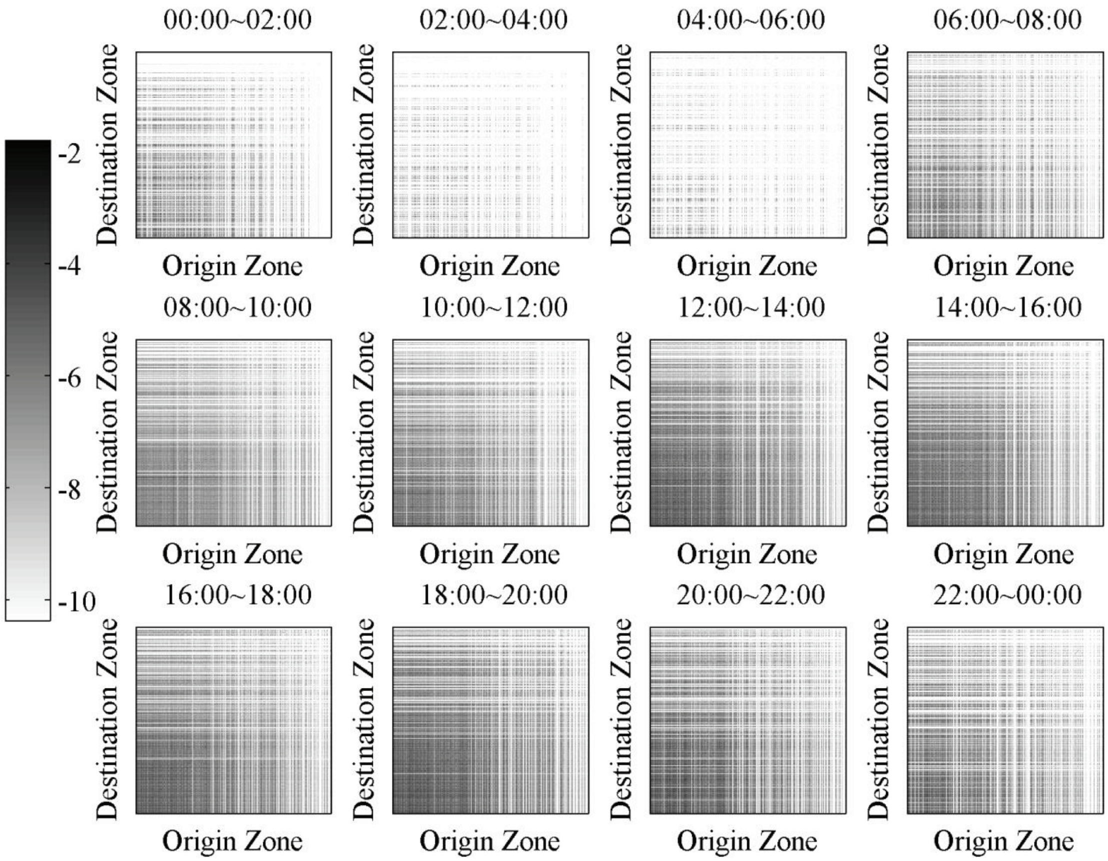


FIGURE 8 OD trip distribution for all trip purposes (zones ordered by zonal attractions).

7. CONCLUSION

This paper investigates the feasibility of using the location-based social networking (LBSN) data to analyze the urban travel demand pattern in Austin TX USA. The study uses the checkin data from the leading LBSN provider, Foursquare, and the ground truth OD matrix from CAMPO. Compared with the traditional O-D estimation methods, LBSN data have better spatial and temporal coverage, built-in user verification, real-time updating capability, and much lower data collection cost. A gravity model based method is proposed to estimate O-D matrix based on the Foursquare "check-in" data. To fit the Foursquare data and the travel demand characteristics of Austin area, a two-regime friction factor model is proposed. Three different types of friction factor functions are evaluated for both the short-distance and long-distance trips respectively using using the coincidence ratio between the model and the ground truth O-D matrix. The model with the linear friction function for short-distance trips and negative exponential friction function for long-distance trips achieves the best results. Using the calibrated model, we further investigate the static and dynamic geographical zonal production and attraction pattern and OD flow pattern. The results are found to be consistent with the travel and activity routines in the Austin area.

In addition, we extended the model to obtain bihourly O-D matrix, the result O-D pattern are consistent with the empirical knowledge, which implies promising potentials of using LBSN data to monitor travel demand.

ACKNOWLEDGEMENT

The authors would like to thank Foursquare® for allowing the research team to obtain data through their developer API and Capital Area Metropolitan Planning Organization (CAMPO) for providing the 2010 OD data for this study. This study is partially supported by the National Key Basic Research Development Program of China (No.2012CB725405).

REFERENCES

1. Barcelo, J., L. Montero, L. Marques and C. Carmona. Travel Time Forecasting and Dynamic Origin-Destination Estimation for Freeways Based on Bluetooth Traffic Monitoring. *Transportation Research Record*, No. 2175, 2010, pp. 19-27.
2. Friedrich, M., K. Immisch, P. Jehlicka, T. Otterstatter and J. Schlaich. Generating Origin-Destination Matrices from Mobile Phone Trajectories. *Transportation Research Record*, No. 2196, 2010, pp. 93-101.
3. Caceres, N., J. P. Wideberg and F. G. Benitez. Deriving Origin-Destination Data from a Mobile Phone Network. *Intelligent Transport Systems*, Vol. 1, No. 1, 2007, pp. 15-26.
4. Watson, J. R. and P. D. Prevedouros. Derivation of Origin-Destination Distributions from Traffic Counts - Implications for Freeway Simulation. *Network Modeling 2006*, No. 1964, 2006, pp. 260-269.
5. Backstrom, L. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. *WWW 2010*, 2010.
6. Martin, W. A. Report 365 Travel Estimation Techniques for Urban Planning. 1998.
7. Nan, L. and C. Guanling. Analysis of a Location-Based Social Network. *Computational Science and Engineering, CSE '09.* , 2009, pp. 263-270.
8. Cheng, Z., J. Caverlee, K. Lee and D. Sui. Exploring Millions of Footprints in Location Sharing Services. *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 2011.
9. Cheng, Z., J. Caverlee and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users. *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 759-768.
10. Bossard, E. G. Retail Trade Spatial Interaction. In *Spreadsheet Models for Urban and Regional Analysis*, Rutgers University, New Brunswick, 1993.
11. Pan, C. X., J. G. Lu, S. Di and B. Ran. Cellular-Based Data-Extracting Method for Trip Distribution. *Traffic and Urban Data*, No. 1945, 2006, pp. 33-39.
12. Doblas, J. and F. G. Benitez. An Approach to Estimating and Updating Origin–Destination Matrices Based Upon Traffic Counts Preserving the Prior Structure of a Survey Matrix. *Transportation Research Part B: Methodological*, Vol. 39, No. 7, 2005, pp. 565-591.
13. De Dios Ortuzar, J., A. M. Ivelic, H. Malbran and A. Thomas. The 1991 Great Santiago Origin-Destination Survey: Methodological Design and Main Results. *Traffic Engineering and Control*, Vol. 34, No. 8, 1993.
14. Albert Sun, J. V.-D. a. Z. S. *A Week on Foursquare*. The Wall Street Journal. New York, 2011.
15. Wolf, J., R. Guensler and W. Bachman. Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1768, No. 1, 2001, pp. 125-134.

16. Tsotsis, A. Foursquare Now Officially at 10 Million Users http://techcrunch.com/2011/06/20/foursquare-now-officially-at-10-million-users/ Accessed March 20,2012.
17. Schapsis, C. Location Based Social Networks, Location Based Social Apps and Games - Links http://bdnooz.com/lbsn-location-based-social-networking-links/#axzz1totECiLc
18. Foursquare. The Foursquare Platform https://developer.foursquare.com/overview/ Accessed Mar 20,2012.
19. Howden, L. M. and J. A. Meyer. *Age and Sex Composition: 2010*. USBS, 2011.
20. USCB. State & County Quickfacts http://quickfacts.census.gov/qfd/news.html Accessed Nov. 10, 2012.
21. CLRSearch. Austin Demographics Summary http://www.clrsearch.com/Austin_Demographics/TX/ Accessed Nov. 10, 2012.
22. CAMPO. *2005 Base Year Travel Demand Model Calibration Summary for Updating the 2035 Long Range Plan*. Capital Area Metropolitan Planning Organization, T. D. o. T., and the Texas Transportation Institute, Austin, 2010.
23. Yang, F., P. Jin, Y. Cheng and B. Ran. Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data. *International Journal of Sustainable Transportation*, 2012,Under Review.