Dynamic Origin-Destination Travel Demand Estimation using Location Based Social Networking Data

Fan Yang, Ph.D. Candidate (Corresponding Author)
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

Peter J. Jin, Ph.D.
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

Xia Wan, Ph.D.
Department of Civil and Environmental Engineering,
University of Wisconsin-Madison
1241 Engineering Hall 1415 Engineering Drive, Madison, WI 53706
Phone: 1-608-556-4289
Email: wan5@wisc.edu

Rui Li, Ph.D., Assistant Professor
College of Civil and Transportation Engineering,
Hohai University
Xi Kang Road 1#, Nanjing 210098, China
Phone: 86-15895985925
Email: liruihhu@163.com

Dr. Bin Ran, Ph.D., Professor
School of Transportation, Southeast University
No.2 Si Pai Lou, Nanjing 210096, China
and
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

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

4210 Words + 3 Tables + 8 Figures = 6960 Words

ABSTRACT

The Location-based Social Networking (LBSN) data have emerged as new data sources for studying travel demand. This paper investigates the feasibility of using LBSN data to estimate dynamic Origin-Destination (OD) travel demand for general trips. A combined non-parametric cluster and regression model is used to establish the relationship between LBSN data and the trip production and attraction. A modified gravity model based trip distribution method with three friction function variations is proposed to estimate the OD matrix. The proposed methods are calibrated and evaluated against the ground truth OD data from CMAP (Chicago Metropolitan Agency for Planning). The results demonstrate the promising potential of using LBSN data for dynamic OD estimation.

KEY WORDS: Dynamic Origin-Destination Estimation, Location-based Social Networking, Gravity Model

1 Introduction

Dynamic Origin-Destination (OD) travel demand information is an essential input for urban transportation planning and traffic operational applications. Traditional OD estimation methods typically rely on household surveys and roadside surveys, which tend to be costly, labor-intensive and time-consuming to undertake. Moreover, the survey data cannot provide up-to-date information to reflect the rapid changes in travel demand pattern since the traffic demand can vary significantly by time of day and day of week over different locations.

Researchers have been exploring various alternative data sources to derive dynamic OD matrices, including traffic counts data, cellular data, Bluetooth data, and GPS data, etc. Traffic-count based OD estimation methods *(1-4)*rely on an existing metering infrastructure, which may be expensive to install or maintain. The cellular data based OD estimation methods*(5-7)* have to solve the problem in matching the cellular cell and Traffic Analysis Zone (TAZ), which reduces the accuracy of trip estimation. For the GPS based methods*(8)*, collecting GPS data requires user consent on disclosing his/her trajectory data, therefore incentives are needed to promote participation in the GPS data collection, which can be expensive when sample size requirement is high. Bluetooth based OD estimation methods*(9)* suffer from sample size issue, the typical penetration rate is about 1% to 5% since the Bluetooth functionality may be turned off in many mobile devices to conserve battery.

With the increasing popularity of smartphones and tablets with Location-based Service (LBS) features, Location-based Social Networking (LBSN) services have attracted users of different income levels to broadcast their locations and activities through their LBSN applications. When it comes to travel demand analysis, especially the OD estimation, the social networking data have some unique advantages over the GPS, cell phone and Bluetooth data. The LBSN applications use the built-in GPS to obtain an accurate position of the user's current location, and the users will need to confirm the name of the place when they "check-in" at a location via LBSN, which ensures the data quality for trip origins and destinations for travel demand analysis. In addition, the penetration rate of social networking service is growing at a rapid pace. The leading LBSN service provider Foursquare has attracted 20 million registered users with an average 3 million check-ins per day by April 2012*(10)*. More importantly, the data

are updated in real-time, and the cost is low since it requires no auxiliary data collecting devices. Yang*(11)* proposed a combined clustering, regression, and gravity model to estimate daily OD matrix for non-commuting trips based on Foursquare check-in data in the Chicago urban area. Jin*(12)* used Foursquare data to estimate the daily home-based work trips, the home-based retail trips, and the general trips in Austin, Texas.

This paper utilizes the Foursquare check-in data to derive dynamic OD travel demand information. The rest of the paper is organized as follows. Section 2 introduces the data required to conduct LBSN-base OD estimation. Section 3 proposes a gravity model based method to derive OD matrix from the check-in data. The optimal setting of the model regarding friction function and venue classification is also identified. Section 4 evaluates the proposed model using ground truth OD matrix. The last section concludes this paper.

2 Data Collection

This section introduces the required data for dynamic OD travel demand estimation using the Foursquare data, which include the GIS data of the research area, the official OD data provided by the local MPO (Metropolitan Planning Organization), and the hourly and daily "check-in" statistics for Foursquare. The data collection method and data quality improvement methods are also introduced in this section.

2.1 Research area and the GIS data

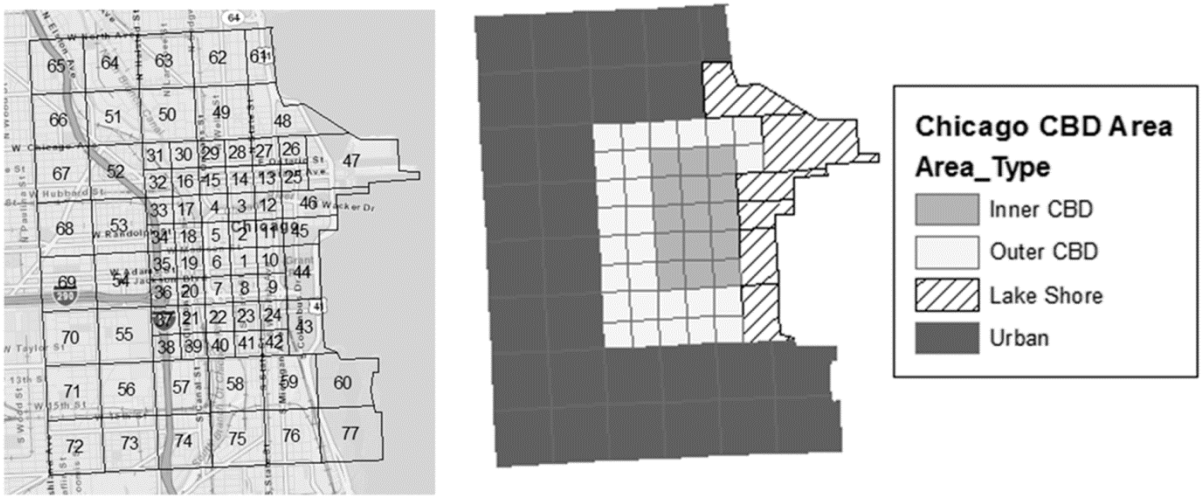


FIGURE 1 TAZ of the research area

This research is conducted in the Chicago central area, which is a representative urban area and has high coverage of Foursquare venues and users. The selected study area is bounded by North Avenue, Ashland Avenue, and Cermak Road, as shown in Figure 1. The official Traffic Analysis Zone (TAZ) system developed by the CMAP(Chicago Metropolitan Agency for Planning) is used for this study. Among the 77 TAZs, 47 zones are 1/4 mile by 1/4 mile located in the CBD area; while the other 30 zones are 0.5 mile by 0.5 mile located around the CBD. The

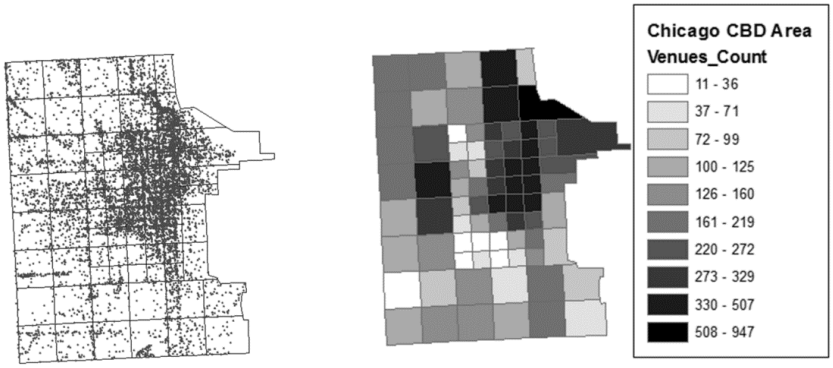
smaller zones reflect that the trips are condensed in the CBD area. Areas along the Lake Michigan (lake shore area) have many points of interests for tourists and have travel patterns from the other zones. Instead of using the CMAP ID system, the zone ID is rearranged in a spiral manner for the convenience of displaying OD flow pattern. The numbering starts from the center where the trips are most frequent. We define zone 1~15 as the inner CBD area, zone 16~42 the outer CBD area, zone 43~48 the lake shore area, and zone 49~ 77 the urban (Non-CBD) area.

2.2 Ground Truth OD Matrix

The ground truth OD matrix data were derived from CMAP's most recent analysis completed in 2010 *(13)*. The CMAP procedure uses an intervening opportunity distribution model to obtain the OD matrices, which uses the trip ends from the trip generation model as a measure of the number of satisfying opportunities, and a trip impedance measure to reflect the difficulty to travel between analysis areas. The CMAP matrices contain the modeled daily trip tables for three trip purposes, which includes home-based work auto person trips, home-based other auto person trips, and non-home based auto person trips. The CMAP also provides time-of-day factors to be applied to daily trips to obtain trips for the following time period: (1) the ten hour late evening-early morning off-peak period(8:00M~6:00AM); (2) the shoulder hour preceding the AM peak hour(6:00AM~7:00AM); (3) the AM peak two hours(7:00AM~9:00AM); (4) the shoulder hour following the AM peak hour(9:00AM~10:00AM); (5) a five hour midday period(10:00AM~2:00PM); (6) the two hour shoulder period preceding the PM peak hour(2:00PM~4:00PM); (7) the PM peak two hours(4:00PM~6:00PM), and; (8) the two hour shoulder period following the PM peak hour(6:00PM~8:00PM).

2.3 LBSN Data Collection

Foursquare provides API endpoints*(14)* for accessing a resource such as a venue, tip, or user, at a canonical URL. Third party programs can get the venue list and obtain the total number of check-ins at a venue via the API. 17,235 venues are collected in the research area, which are represented using scattered dots in Figure 2a. The number of venues in each TAZ is color coded in Figure 2b. The venues are more densely distributed in the middle-east part of the research area with those smaller zones than the other areas.



(a) Venue locations (b) Venue Density Distribution

FIGURE 2 Foursquare Venues Locations and their Categorical Distribution among TAZs

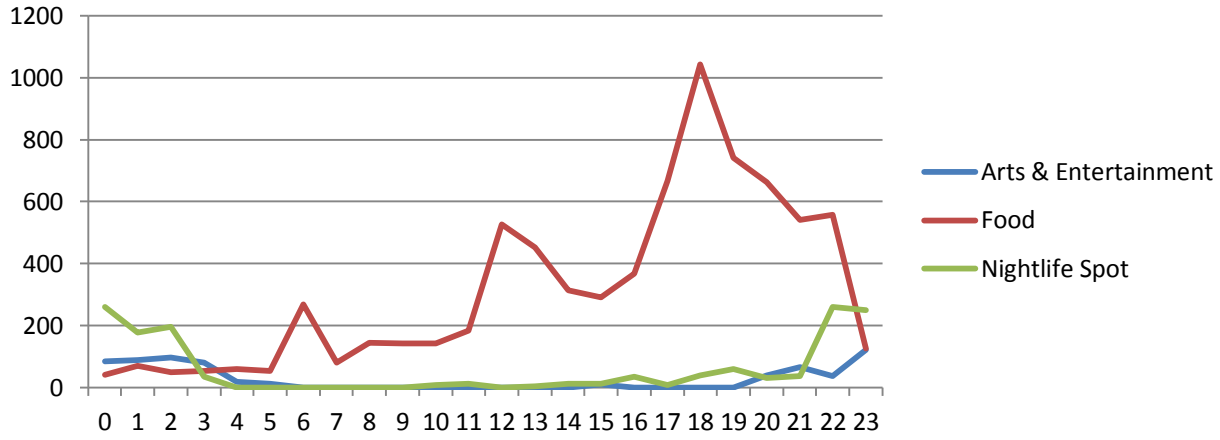
TABLE 1. Foursquare Venue Statistics by Categories

Category	# of Venues	Percentage	# of Check-ins	Percentage	Avg. Check-ins
Arts & Entertainment	1052	6.1%	564456	5.8%	537
College & University	751	4.4%	391506	4.0%	521
Food	2361	13.7%	2893552	29.7%	1226
Nightlife Spot	1094	6.3%	944981	9.7%	864
Outdoors & Recreation	942	5.5%	518660	5.3%	551
Professional & Other Places	4685	27.2%	1533994	15.7%	327
Residence	873	5.1%	189669	1.9%	217
Shop & Service	3056	17.7%	1608089	16.5%	526
Travel & Transport	1245	7.2%	1040593	10.7%	836
Unclassified	1176	6.8%	65125	0.7%	55
Sum	17235	100.0%	9750625	100.0%	566

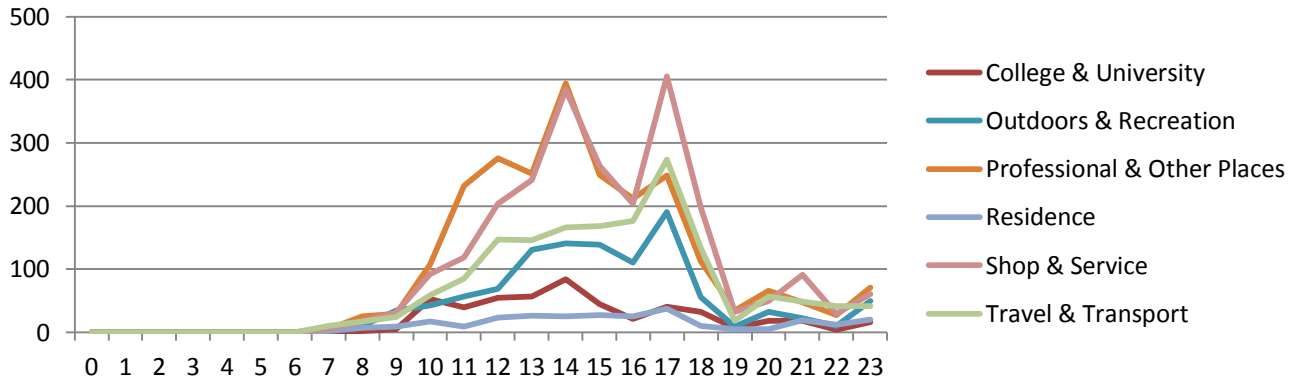
Fourquare classified the venues into 9 venue categories. Table 1 shows several statistical results of the venues aggregated by different categories. The "Professional & Other Places" category has the largest number of venues while the "Food" category attracted the most total check-ins. 1176 venues are not labeled with any categories, but the total check-ins in those venues are small enough to be eliminated in our travel demand analysis.

2.4 Preliminary Analysis of the Check-in Data

The data collection program runs continuously to collect the hourly snapshots of the total check-in records for each venue. The program ran for two months between May 21, 2013 and July 20, 2013. Figure 3a shows the hourly check-in pattern within a day for those venues with venue types "Food", "Arts & Entertainment" and "Nightlife Spot". Food venues peak around lunch and dinner time, with a greater peak during lunch. The "Arts & Entertainment" and "Nightlife Spot" venues receive more check-ins in the night than in the day. Figure 3b shows the hourly check-in pattern of the other categories. Most check-in activities happen in the day time (6:00am – 11:00pm), which indicates people are more active on Foursquare during the day than the night. All these patterns indicate reasonable spatial and temporal coverage of major activities of people in Chicago central area.



a.



b.

FIGURE 3 Hourly Check-in Pattern

3 Methodology

3.1 Methodology Framework

The methodology of converting hourly Foursquare check-in data to hourly general purpose OD matrices includes two main models: a non-parametric cluster and regression model that convert the hourly check-in data to trip productions and attractions, and a modified gravity model to obtain OD matrices using the productions and attractions estimated by the hourly check-in data, which is shown in Figure 4.

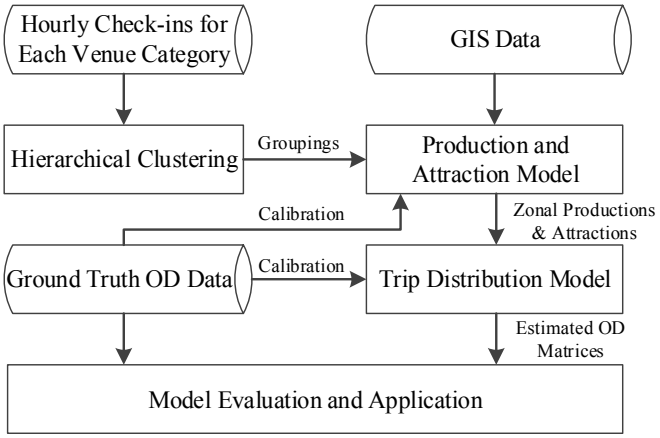


FIGURE 4 Proposed Framework for LBSN-Based OD Estimation

The methodology involves two calibration steps, in which the CMAP OD data is used as ground truth data to calibrate the model. In the first step, the hierarchical clustering method is used to group venue types with similar hourly check-in pattern in order to reduce the number of parameters in the model. Then the hourly check-ins in each groups are used as input to the production and attraction model to estimate the hourly productions and attractions in each TAZ. In the second step, the OD matrix is estimated by the trip distribution model using the productions and attractions obtained from the previous step as input. The trip distribution model is based on the gravity model and three friction functions are applied and compared. A singly-constraint balancing procedure is used to satisfy the relationship between productions and attractions. In the final procedure, the models built in the previous two steps are evaluated and compared, the final model is selected using Deviation to Diagonal and Coincidence Ratio as the standard.

As mentioned in Section 2.2, CMAP divided the OD data into 8 time periods. We used the data for the first four time periods (8:00PM~10:00AM) to calibrate the model parameters. Therefore the predicted OD data for the last four time periods can be used to verify the proposed methodology.

3.2 Hourly Origin-Destination Trip Model

The hourly OD trip model used in this study is based on the daily OD trip model formulation proposed in *(11)*. This research assumes that the hourly trip productions and attractions in a TAZ are directly related to the number of check-ins in each venue categories. The ratio of total check-ins at certain category of venues to the actual number of trips made from/to those venues need to be adjusted, in order to balance the bias of the different probabilities for users to check-in at different venue categories. Shops and restaurants may receive more check-ins motivated by promotions and discounts. Users may be more likely to check-in at recreational places than the Home and work locations. The functions are listed as:

$$P_i = \sum_{k=1}^{K} p_k x_{ik} + p_0 , \ i = 1,2, \dots, N$$
$$A_j = \sum_{k=1}^{K} a_k x_{jk} + a_0 , \ j = 1,2, \dots, N$$
$$T_{ij} = P_i \frac{A_j F_{ij}}{\sum_{j=1}^{N} A_j F_{ij}}$$
$$\textbf{\textit{s.t.}} \sum_{i=1}^{N} P_i = \sum_{j=1}^{N} A_j \quad (1)$$

Where

P_i: Trip production at origin zone *i*

A_j: Trip attraction at destination zone *j*

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

x_{jk}: Check-ins for venue type *k* in destination zone *j*

p_k, a_k: Coefficients for estimating the trip production/attraction contribution according to total check-ins for venue type *k*

N: The total number of TAZs

K: The total number of venue types

p_0, a_0: The constant terms

T_{ij} : Number of trips between origin zone *i* and destination zone *j*

F_{ij}: Friction factor for trips between origin zone *i* and destination zone *j*

The friction factor term (F_{ij}) represents the impedance of travelers to make trips of various distances. Three types of friction functions are to be experimented including the linear function, the Negative Exponential function*(15)* and the gamma function. Those friction functions are selected because 1) they have been used and validated in the existing studies; 2) the number of parameters in the friction functions needs to be small to avoid adding too much complexity to the proposed model. The forms of these friction functions are listed as follows.

1. Linear Function. $F_{ij} = \alpha + \beta d_{ij}$ (2)

2. Negative Exponential Function. $F_{ij} = \alpha e^{-\beta d_{ij}}$ (3)

3.Gamma. $F_{ij} = \alpha {d_{ij}}^{\beta} e^{\gamma d_{ij}}$ (4)

where

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

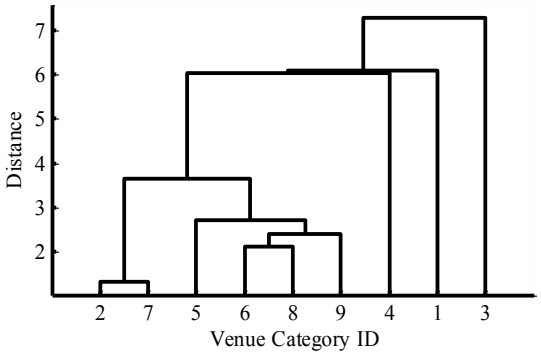
β : A positive or negative constant value which affects the distribution of shorter trips

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} : Distance between the centroids of origin zone i and destination zone j

The large number of model parameters in Equation (1) can lead to significant overfitting issue. Therefore, agglomerative hierarchical clustering method is used to group similar venue types. Agglomerative hierarchical clustering*(16)* is a bottom-up clustering method which starts with every single object in a single cluster, the closest pair of clusters is merged together if certain similarity criteria is satisfied in each successive iteration, until all of the objects are agglomerated as one cluster. The dissimilarity matrix is obtained by computing the Euclidean distance of the hourly number of check-ins in each venue type, which is illustrated in Figure 3.

The single linkage method is used in merging the venue types. The dendrogram is shown in Figure 5.



ID	Venue Category Name
1	Arts & Entertainment
2	College & University
3	Food
4	Nightlife Spot
5	Outdoors & Recreation
6	Professional & Other Places
7	Residence
8	Shop & Service
9	Travel & Transport

FIGURE 5 Dendrogram of the Venue Types

Another issue in Equation (1) concerns the trip balancing problem. The number of trips from zone i to other zones must equal to number of trips produced in zone i, and the number of trips from other zones to zone j must equal to number of trips attracted to zone j. However, Equation(1) only satisfies the first constraint. Therefore, iterative attraction trip-end balancing procedure is applied as illustrated in Figure 6.

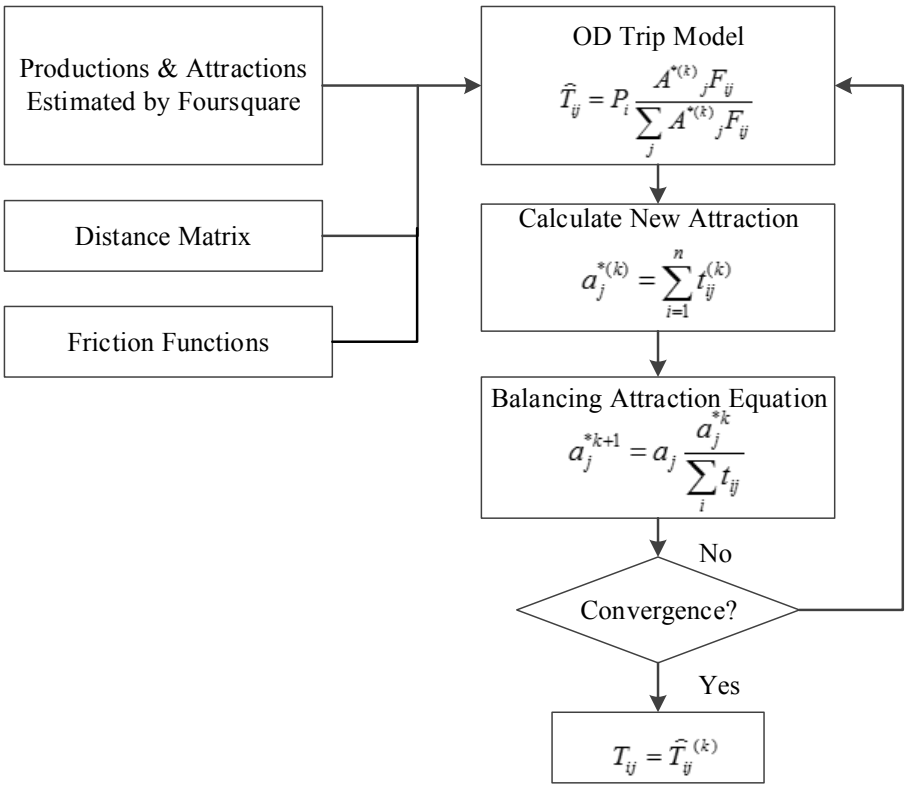


FIGURE 6 Iterative Attraction Trip-end Balancing Flow Chart

3.3 Model Calibration

The model calibration follows a two-phase procedure using the genetic algorithm to obtain the parameters which optimize the objective functions in each phase. The objective functions in the two phases are to minimize the Deviation to Diagonal (DD)*(11)* between the predicted value and the ground truth value defined as the following:

$$DD = \frac{\sum_{i=1}^{n}\left|\frac{\arg(x_i,y_i)-\pi}{\pi}\times 180-45\right|}{n} \quad (5)$$

Where

x_i, y_i: Elements in N-dimensional vector X,Y

n: Number of elements in vector X, Y

$\arg(x, y)$: The Azimuth of (x, y) when being projected into polar coordinate system.

DD is essentially the deviation between the Azimuth of (x, y) and the $y = x$ line in degree. DD is selected instead of the classic error measures such as Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE) because of the high fluctuations among trip production and attraction for different zones and trip frequency between different OD pairs. DD can provide a relatively unbiased count that absolute or relative error based measures. Smaller DD value indicates the model predicted values are closer to the ground truth values.

Firstly, trip production and attraction function is calibrated to obtain the coefficients p_n, a_n, and p_0 . Nine models with respect to nine venue classification methods are calibrated in total. Then the parameters associated with the trip distribution function are calibrated. A total of 27 different models are to be calibrated and compared with the proposed eight venue classification methods and three friction functions.

In addition to the two Deviation to Diagonal indexes calculated in the two model calibration phases, which include the DD of the production and attraction function and the DD of the trip distribution function, we also used the Coincidence Ratio (CR) to evaluate the performance of the 27 models.

The Coincidence Ratio measures the percent of the area that "coincides" for the two curves of distributions to compare*(17)*. In evaluating the fitness of the model, we compare the percentage of trips in each trip length interval for CMAP'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:

$$CR = \frac{\sum_{i=1}^{n}\min(p_i^M, p_i^O)}{\sum_{i=1}^{n}\max(p_i^M, p_i^O)} \quad (6)$$

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 CMAP.

n : Number of intervals.

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 CMAP results indicates a better model.

All of the three measures for the 27 models tested are listed in Table 4.

TABLE 2. Deviation to Diagonal (DD) and Coincidence Ratio (CR) Calibration Results

		n=9	n=8	n=7	n=6	n=5	n=4	n=3	n=2	n=1
ProAttrDD		43.4	44.5	44.4	46.6	46.8	43.2	43.3	45.3	43.4
Trip DD	Linear	29.2	29.3	29.8	29.8	29.9	29.3	29.3	29.9	29.3
	NegExp	28.9	29.0	29.0	29.6	29.7	28.7	28.7	32.8	32.7
	Gamma	32.5	31.3	32.5	33.0	29.9	29.2	28.4	30.8	29.0
CR	Linear	0.77	0.77	0.77	0.75	0.75	0.74	0.74	0.76	0.74
	NegExp	0.82	0.83	0.83	0.82	0.83	0.81	0.83	0.83	0.55
	Gamma	0.49	0.55	0.55	0.46	0.73	0.73	0.56	0.64	0.74

As indicated in Table 2, the models using negative exponential friction function generally have lower DD values and higher CR values. We selected the three venue classifications model with negative exponential friction function through comprehensive consideration. The calibrated parameters for Equation (1) for Chicago central area are listed in Table 3, which will be used for further study of dynamic travel patterns.

TABLE 3. The Calibrated Parameters

Notation	Value	Explanation
N	77	Number of Traffic Analysis Zones
n	3	Number of Venue Groups
G_1	Venue Group 1	College & Universities, Nightlife Spots, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport
G_2	Venue Group 2	Art & Entertainment,
G_3	Venue Group 3	Food
a_0	8.76	The constant term in the attraction function
a_1	0.021	The scaling factor for Venue Group 1 in the attraction function
a_2	0.295	The scaling factor for Venue Group 2 in the attraction function
a_3	0.003	The scaling factor for Venue Group 3 in the attraction function
p_0	0.92	The constant term in the production function
p_1	0.001	The scaling factor for Venue Group 1 in the production function
p_2	0.111	The scaling factor for Venue Group 2 in the production function
p_3	0.064	The scaling factor for Venue Group 3 in the production function
α	0.540	The parameter in the negative exponential function
β	0.001	The parameter in the negative exponential function

4 Model Evaluation and Application

This section compares the estimated Foursquare OD matrix with the CMAP ground truth matrices in each of the eight time periods defined in Section 2.2. As a supplement to the Deviation to Diagonal and the Coincidence Ratio error indexes, this section also uses visual measures including the trip length distribution curves and the OD heat map to evaluate the proposed model.

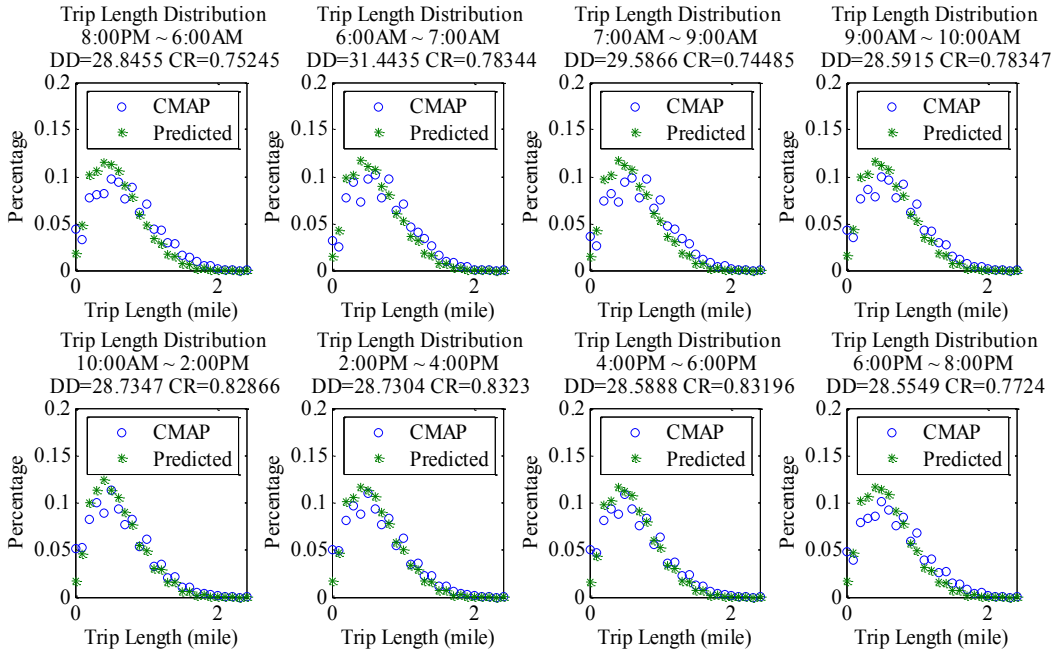


FIGURE 7 Trip Length Frequency Comparisons

Figure 7 compares the trip length distribution curves of the ground truth OD data and the estimated OD using the Foursquare data in each time period. The Deviation to Diagonal and the Coincidence Ratio error indexes are also shown in Figure 7. The trip length distribution curves visualize the statistics used to calculate the coincidence ratio. Generally the estimated distribution curves are consistent with the CMAP data but slightly overestimate the frequency of short trips in each time period. Paired-Sample t-Tests are conducted in each time period to test the null hypothesis that the pairwise difference between the two trip length distribution curves has a mean equal to zero at the 5% significance level. The returned results of h = 0 and p=1 in every time period indicate that t-test does not reject the null hypothesis, which proves that the Foursquare OD matrix and CMAP ground truth OD matrix have similar trip length distributions. As mentioned in Section 3.1, the parameters of the model are calibrated using the first four time periods (8:00PM~10:00AM) in the ground truth data. The similarity in the CMAP data and the estimated results for the last four time periods (10:00AM ~8:00PM) indicates the hourly OD trip model can be applied to estimate OD result on an hourly basis.

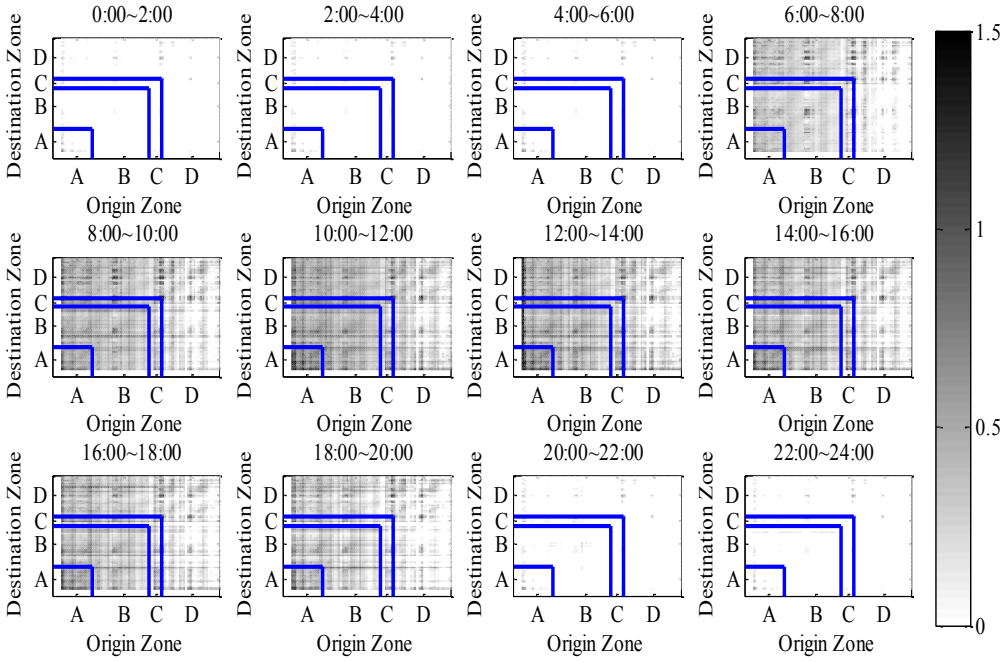


FIGURE 8 Bihourly Zonal OD Flow Pattern (Log10 values)
(A – Inner CBD, B – Outer CBD, C – Lake Shore, D – Urban)

Figure 8 shows the zonal OD flow pattern estimated using Foursquare data in every two hours interval. The horizontal axis represents the origin zone ID, and the vertical axis is the ID of the destination zone. The zone numbers are the same as those shown in Figure 1. Both axes are divided by zones defined in Section 2.1. Zone A is the inner CBD area, zone B is the outer CBD area, zone C is the lake shore area, and zone D is the urban area. Each grid (i, j) in the diagram displays the Logarithm Trip Intensity LTI_{ij} from zone i to zone j, which can be calculated as:

$$LTI_{ij} = \log_{10}(T_{ij}) \tag{7}$$

As indicated in Figure 8, the travel demand are fairly low from 8:00PM to 6:00AM, high from 8:00 to 16:00, which reflect the regular working and resting schedule. Typically, rush hour should be from 6:00 to 10:00 and 16:00 to 19:00. Generally, the bihourly flow pattern is also consistent with the empirical experience.

5 Conclusion

This paper introduces a new dynamic Origin-Destination travel demand estimation methodology using the Location-based Social Networking (LBSN) data. A gravity model based method is proposed to estimate OD matrix based on the Foursquare "check-in" data. The CMAP OD data are divided into eight time periods by applying the time-of-day factors to the daily OD matrix, where the first four time periods are used to calibrate the model parameters and the last four time periods are used to validate the method. The agglomerative hierarchical clustering

method is used to group similar venue types to avoid the overfitting issue in model calibration. Three friction functions are proposed to estimate the effect of trip distances. A total of 27 model variations are calibrated and evaluated using the Deviation to Diagonal (DD) and Coincidence Ratio (CR) measures. The model with the negative exponential friction function and three fraction factors for all venues categories is chosen. In addition, the feasibility of using LBSN data for dynamic OD estimation is validated through the trip length frequency distribution comparison. The model is also applied to estimate bihourly zonal OD pattern.

The comparison in the previous section suggests that the proposed LBSN based dynamic OD matrix estimation method can produce reasonable result. However, some discrepancies still exist between the Foursquare and CMAP OD matrix. The following error sources may contribute to the differences: 1) The ground truth data may not reflect the real dynamic OD pattern. 2) Foursquare venues may not cover all the locations in a TAZ, therefore trips made in the uncovered locations cannot be estimated. 3) The trip distribution method strongly relies on the gravity model to estimate the OD matrix from the production and attraction data. The zonal differences in geographical characteristics are neglected, and the resolution of the gravity model may yield estimation errors.

It should be pointed out that the proposed method still relies on the input of a ground truth OD matrix from MPO. However, different from traditional methods which use the MPO OD matrix as a baseline matrix, the MPO OD matrix is used only to determine the correlation ratio between the check-ins reported in one type of venues and the implied trip production or attraction. With such ratios determined, LBSN data can be used to directly generate OD matrices rather than modifying an existing MPO OD matrix.

However, the study still has several limitations to be addressed in future work. The proposed methodology is based on the gravity model. Future studies could explore other trip distribution models to improve the accuracy of OD estimation. In addition, it is necessary to apply the model to other geographical areas to evaluate the transferability of the proposed method.

ACKNOWLEDGEMENT

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

REFERENCES

1. 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.
2. 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.
3. Cascetta, E. and S. Nguyen. A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Research Part B: Methodological*, Vol. 22, No. 6, 1988, pp. 437-455.
4. Bell, M. G. H. The Estimation of an Origin-Destination Matrix from Traffic Counts. *trsc*, Vol. 17, No. 2, 1983, pp. 198-217.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. Yang, F., J. Jin, Y. Cheng and B. Ran. Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data. *International Journal of Sustainable Transportation*, Accepted, 2013
12. Jin, J., F. Yang, M. Cebelak, B. Ran and C. M. Walton. Urban Travel Demand Analysis for Austin Tx USA Using Location-Based Social Networking Data. *The 92nd Annual Meeting of Transportation Research Board*, 2013
13. CMAP. *Travel Model Documentation*. Chicago, 2010.
14. Foursquare. The Foursquare Platform https://developer.foursquare.com/overview/ Accessed Mar 20,2012.
15. Bossard, E. G. Retail Trade Spatial Interaction. In *Spreadsheet Models for Urban and Regional Analysis*, Rutgers University, New Brunswick, 1993.
16. Friedman, J., T. Hastie and R. Tibshirani *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.
17. Martin, W. A. Report 365 Travel Estimation Techniques for Urban Planning. 1998