Location-Based Social Networking Data: An Exploration into the Use of a Doubly-Constrained Gravity Model for Origin-Destination Estimation

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

Meredith Cebelak, P.E., (Corresponding Author)
Graduate Research Assistant
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: mcebelak@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

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., P.E., Professor
Ernest H. Cockrell Centennial Chair in Engineering,
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

Corresponding Author: Meredith Cebelak
Words: 5693 + (2 Tables + 5 Figures)*250 = 7443

ABSTRACT

Trip distribution is an invaluable portion of the transportation planning process leading to the creation of origin-destination (OD) matrices. Location-based social networking (LBSN) has increased in popularity and sophistication, emerging as a new travel demand data source. Users of LBSN provide location-sensitive data interactively via mobile devices, including smartphones and tablets. This data has the potential to provide origin-destination estimates with significantly higher temporal resolution at a much lower cost in comparison with traditional methods. This paper proposes a LBSN OD estimation model based on the doubly-constrained gravity model to improve a previously-proposed model based on the singly-constrained gravity models. The proposed methodology is calibrated and comparatively evaluated against the OD matrix generated by the singly-constrained gravity model based method as well as a reference matrix from the local metropolitan planning organization. The results of this method illustrate significant improvement in reducing the OD estimation errors caused by the sampling bias from the singly-constrained gravity model based method.

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

INTRODUCTION

Trip distribution is a significant step in the transportation planning process which generates the static or dynamic origin-destination (OD) trip patterns to be used by traffic assignment models. The existing data collection methods for trip distribution can be classified into three main categories: survey-based, traffic counts, and positioning technology based. Survey-based methods such as telephone, in-person interview, mail or email survey can collect complete the social-demographic information of travelers and households and trip information. These methods are time-consuming and labor-intensive and can only generate static travel demand information at low frequency (e.g. every 5-10 years) due to the funding and resource limitations. Traffic count based methods calibrate an OD matrix based on traffic detector data *(1-5)*. These methods have relatively low cost and can provide dynamic OD information, if calibrated properly. However, they require an existing OD matrix and rely on traffic assignment models to generate accurate traffic flow data to be compared with field detector data in model calibration. The use of positioning technologies for OD data collection has the potential of producing OD data with much higher spatial and temporal resolution, larger sample size, and less cost than survey-based methods *(6-8)*. The complication lies in 1) the penetration rate of a specific position technology in the mobile devices of travelers, 2) privacy protection, and 3) the uncertainty in determining trip purposes and destinations due to positioning errors.

Within the US, the affordability and accessibility attributed to recent technological advances has allowed smartphones and tablets with location based service features to be available to individuals of diverse income levels. This in conjunction with the fast development of social networks attracts a substantial amount of users active in relaying their personal activities online often including their locations. Location-based social networking (LBSN) combines the aspects of social networking with the location based services features, which provides some advantages over other positioning technologies *(9)*. User activities produce trip purposes and destination information through applications with built-in GPS by "checking-in" at particular venues. The sample provided from this methodology has the potential to be larger than other methods due to the penetration rate of social networking services growing at a rapid pace. Furthermore, the lack of auxiliary data collection devices and availability of real-time updated data make this method of data collection an attractive low cost option. In a previous study, a singly-constrained gravity model based method was proposed and evaluated *(10)*. Although the study revealed promising potentials of LBSN data for OD estimation, the proposed model still has some limitations, especially the significant bias in OD patterns related to short-distance trips and residential areas.

This paper proposes a doubly-constrained gravity model based method whose improved learning capability, when compared to the singly-constrained method, during model calibration can reduce the sampling bias of LBSN check-in data. Section 2 of this paper introduces the state of practice for data collection. The methodology and procedures are introduced in Section 3. Next, Section 4 provides details on the experimental design as well as results from the proposed algorithm. Finally, Section 5 concludes the paper and provides some areas for the continuation of this research effort.

BACKGROUND

State of Practice Review on OD Estimation

Conventionally, OD matrices are derived by expanding sample OD matrices collected from traditional household travel behavior surveys based on social-demographic and economic data for a planning area. These survey methods include personal home interviews, telephone interviews, mail survey, and/or internet survey. Personal home interviews are one of the most complete data sources with the highest response rate, 60-70%, when compared to other household survey methods *(6)*. Home interviews are the most expensive and time consuming method, while telephone, mail, and internet based surveys have significantly less cost and time involved in collection. This reduction in time and cost comes with a decrease in response rate and introduces sampling biases. In recent years, Global positioning systems (GPS) assisted travel surveys have become popular both in the US and internationally *(11)*. However, significant incentives as well as logistical issues, including battery outages leading to incomplete data and loss of GPS units, were identified. Moreover, studies have shown that participants may be burdened by the extended length of GPS surveys, equipment complications, and privacy concerns *(11)*, and samples are often biased *(7)*.

Traffic-count data has been implemented as a data collection method for use in OD matrices. Studies have shown that OD matrix creation is possible given traffic volumes for each transportation link *(1-5, 12)*. However, many different matrices can be reproduced from observed traffic counts and deployment of a comprehensive detector infrastructure on all viable routes would be required. Additionally, concerns about the accuracy of estimated traffic conditions between fixed detectors has been discussed in recent research efforts *(13)*. A complementary survey method for the traffic count based method is the roadside intercept survey, which provides additional information regarding the OD composition of traffic flow at a road section.

In recent years, the emergence of secondary planning data sources, such as GPS, cellphone, and Bluetooth, has caught researchers' attention. Different from the aforementioned GPS based survey, recent research efforts have demonstrated the feasibility of replacing traditional survey methods with OD data directly derived from GPS trajectories generated by travelers' in-vehicle devices *(6, 8)*. Cellular phones have been explored for their data collection capabilities through their employment of wireless location technologies (WLT). Studies *(14, 15)* have shown that cellular phone technologies were both theoretically and experimentally feasible with reasonably precise estimation results. Penetration rates needed to achieve the spatiotemporal coverage of a network are between 2 and 3% *(16)*. There are limitations with the technology. The spatial resolution of the cellphone positions may be within a cellular cell or location area that may include multiple TAZs (traffic analysis zones). The LBS (location based service) data based method can significantly increase the spatial resolution, but users may not turn on LBS function or report their LBS data due to privacy concerns. Recently, Bluetooth has been noted to be a low cost and user-friendly method for data collection *(17-19)*. Employing a unique media access control (MAC) assigned by each devices manufacturer alleviates privacy concerns affiliated with other methods of data collection. However, the technology is limited by the short ping cycle that could lead to devices being over sampled, the potential for a single vehicle to have multiple Bluetooth capable devices, as well as the ability to turn off Bluetooth functions within a device. Additionally, the variability of Bluetooth samples could yield

objectionable expansion errors which negates the technologies ability to independently create an estimation for an OD matrix *(20)*.

Currently, research is being conducted to determine the ability for "Big", vehicle-to infrastructure (V2I), and smartphone data to be used for OD matrix creation. "Big" data includes transactional (i.e. credit card purchase and payment records, product/services logs), interactional, and observational data *(21)*. While this data sources has great potential, there are limitations to the incorporation into transportation planning, specifically with the ability for data to be shared. Additionally, data capture, management, and storage pose potential difficulty with utilization, and biases may exist. Similar to the credit card data mentioned within "Big" data, transactional data has been researched for potential use to improve transit planning *(22)*. The study noted that concerns with market penetration, sampling bias, privacy concerns, as well as errors with transaction/routing assignment exist with the method. V2I has recently been explored as a potential new data source, indicating that the use of the dedicated short-range communications connecting vehicles to infrastructure would have the potential to collect data on every vehicle within the system, effectively eliminating the need for an estimated OD matrix *(23)*. With the exception of V2I test-beds, this method of data collection is not viable at this time and privacy concerns would need to be address prior to acceptance.

Location-Based Social Network (LBSN) and Austin LBSN Data Characteristics

Location-based services (LBS) are services that use location and time data as a control feature. This feature has been encompassed within social networking to create location-based social networking (LBSN). With the increased popularity of sites like Facebook, Twitter, and Foursquare that include LBSN, this form of data collection has been explored recently for comprehension of spatial patterns of users. The first study exploring this area was by Li and Chen *(24)* and utilized the Markov-based location predictor to determine future locations of users with an accuracy of 49%. The relationships between geographic movements, the temporal dynamics of human movements, and social networking ties have been investigated in various studies *(9,25,26)*. Additionally, studies by Backstrom et al. *(27)* and Cheng et al. *(28)* demonstrated the ability to predict user locations via the user's friends and content, respectively.

Many social networking sites have added features that allow users to "check-in" to a place of interests which is called a "venue". This capability allows individuals to share and save places that have been visited with fellow users and friends. Foursquare is the most popular site that includes this feature, and as of January 2013 has over 30 million users worldwide with over three billion check-ins. Users of this particular site included businesses, which encourage check-ins through promotions and discounts. Due to the site's popularity, high penetration rate, and large sample size, researchers have used the LBSN data available to investigate mobility patterns across spatial, temporal, and social aspects *(29, 30)*.

The research team was among the first to used Foursquare data to specifically estimate an OD matrix in *(10)*. This study examined non-commuting trips within the Chicago urban area demonstrating the promising potential of the methodology. In *(31)*, we furthered this effort by examining the use of check-in data to analyze the OD demand for Austin, TX using a singly-constrained gravity model with a two regime friction factor, illustrating the potential of LBSN data for travel demand analysis and monitoring. The detailed LBSN OD estimation model based on singly-constrained gravity model is as the following.

$$P_i = \gamma * x_i \quad (1)$$

$$A_j = \varepsilon * x_j + \frac{1}{N}\sum_j(\gamma - \varepsilon)x_j \quad (2)$$

$$T_{ij} = P_i * \frac{A_j * F(d_{ij})}{\sum_k A_j * F(d_{ij})} \quad (3)$$

where

P_i: the productions for zone i

x_i: the total check-ins in zone i, $x_i = \sum_p x_{ip}$, where $p = 1, \dots, P$ indicates the pth venue type.

A_j: the attractions for zone j

γ: the adjustment factor to zonal trip production from Foursquare check-in counts

ε: the adjustment factor to zonal trip attractions for Foursquare check-in counts

$\frac{1}{N}\sum_i(\gamma - \varepsilon)x_i$: the residual term for zone i that ensures the total production equal to the total attraction.

T_{ij}: the number of trips between origin zone i and destination zone j.

$F(d_{ij})$: the friction function where d_{ij} is the travel cost between zone i and j.

To calibrate the model, the trip balancing process is applied to the following equations iteratively.

$$A_j^{(n)} = \sum_i T_{ij}^{(n-1)} \quad (4)$$

$$T_{ij}^{(n)} = P_i * \frac{A_j^{(n)} * F(d_{ij})}{\sum_k A_j^{(n)} * F(d_{ij})} \quad (5)$$

METHODOLOGY

The proposed model attempts to address several limitations from the previous model. First the zonal productions and attractions generated by the previous model usually results in symmetric patterns due to the uniform distribution of the residual term among all zones (See Equation 2). Second, the singly constraint model only tunes the zonal attractions. Third, the converging rate of the singly-constrained gravity model is relatively slow, which causes the model calibration to be slow and premature. To address those limitations, we propose a new model based on doubly-constrained gravity model and zone-specific residual assignment as follows.

$$P_i = \gamma * x_i \quad (6)$$

$$A_j = \varepsilon * x_j + x_j^{\rho} / \sum_j x_j^{\rho} \sum_j(\gamma - \varepsilon)x_j \quad (7)$$

$$T_{ij} = \beta_i * P_i * \alpha_j * A_j * F(d_{ij}) \quad (8)$$

where

ρ: the power of location factor

β_i: the balancing factor for the productions

α_j: the balancing factor for the attractions

$x_i^{\rho} / \sum_i x_i^{\rho}$: redistribute the residual based on the zonal check-in counts for zone i.

In this way, the residuals are assigned based on the check-in intensity rather than evenly distribute among all zones. The initial values of P_i and A_j are calculated directly from the Foursquare check-in counts based on equations 6 and 7. The T_{ij} is then calculated from P_i and A_j based on equation 8.

The doubly-constrained model can be calibrated by iteratively updating α_j and β_i. In this study, we set $\alpha_j^{(0)} = 1$ and $\beta_i^{(0)} = 1$. The values of α_j and β_i are updated using the following.

$$P_i^{(n)} = \sum_j T_{ij}^{(n-1)} \quad (9)$$

$$A_j^{(n)} = \sum_i T_{ij}^{(n-1)} \quad (10)$$

$$\beta_i^{(n)} = \frac{1}{\sum_j \alpha_j^{(n-1)} * A_j^{(n)} * F(d_{ij})} \quad (11)$$

$$\alpha_j^{(n)} = \frac{1}{\sum_i \beta_i^{(n-1)} * P_i^{(n)} * F(d_{ij})} \quad (12)$$

For consistence, the same friction function was engaged for the doubly-constrained model that provided the best coincidence ratio (CR) in the single-constrained model. The CR measures the percent of the area that "coincides" for the two curves/distributions that are being compared *(32)*. The friction function combined the linear model for short trips and the negative exponential model for long trips as shown in the following equation.

$$F(d_{ij}) = \begin{cases} \theta + \lambda * d_{ij}, & d_{ij} < \sigma \\ \mu * e^{-\rho * d_{ij}} & d_{ij} \geq \sigma \end{cases} \quad (13)$$

where θ, λ, μ, and ρ are factors that were optimized through the genetic optimization algorithm and the d_{ij} is the Manhattan distance between the centroids of origin zone *i* and destination zone *j* in miles. The dual-regime formulation is used to capture CAMPO's special treatment on OD pairs with short distance.

EXPERIMENTAL DESIGN

Study Area

The city of Austin, TX was selected as the study area for this paper. Austin is a diverse city that encompasses an area of 272 mi^2 and has an estimated population of almost one million people. The city of Austin *(33, 34)* was demographically compared to US Foursquare *(35)* users as well as the general US *(36)*, as shown in Figure 1. It should be noted that the Foursquare users have a higher proportion of individuals between the ages of 25 and 54, which constitutes 80% of the sites users. This age group also has a greater distribution than is seen in Austin, TX and the US. Additionally, there are significantly more female users of Foursquare (65% women compared to 35% men), which is also notably different than the distribution of gender within Austin and the US. Examining the educational and income trends of the Foursquare user, it is noted that within the income categories of $25,000 through $74,999 as well as within the "Some College" category there is an over representation when compared to the Austin and US data. Finally, it should be noted that Foursquare prohibits users under the age of 13, which is shown in the percentages of 17 and under and "Less than High School" users. The above potential sampling bias needs to be properly addressed when converting the number of Foursquare check-ins to trip counts.

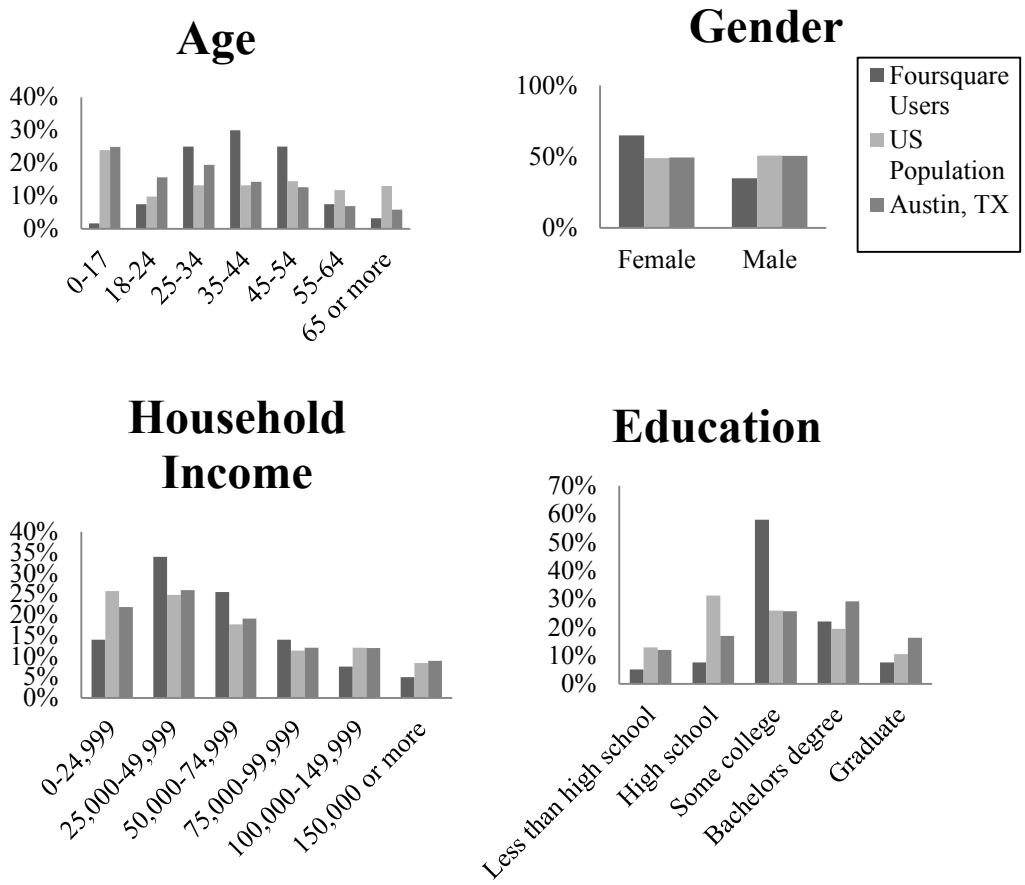


FIGURE 1 Comparative Demographics.

The Capital Area Metropolitan Planning Organization (CAMPO) has identified 520 TAZs within the city of Austin's jurisdiction, which will serve as the study area for this paper. CAMPO's 2005 Travel Demand Model (TDM) serves as the reference data used for the analysis. It should be noted that CAMPO data is not considered the ground truth data due to the limitation of current data collection methods. It serves as a reference data for identifying critical empirical OD patterns. The trip purposes identified within the CAMPO study were combined into eight categories:

1. Home-based Work (HBW)
2. Home-based Non-work Retail (HBR)
3. Home-based Non-work Other (HBO)
4. Home-based Non-work University of Texas (UT) (HBUT)
5. Non-work Airport (NWAir)
6. Non-home Based Work (NHBW)
7. Non-home Based Other (NHBO)
8. Non-home Based External (NHBE)

Data Collection

Foursquare data was collected by first identifying the venues within the study area. Figure 2 shows the 19,710 venues identified within the study area, demonstrating the special coverage of the data. It should be noted that all TAZs with the exception of three, highlighted in

the figure, had at least one venue with the majority of venues located within the denser urban areas.

FIGURE 2 Venue Locations within the Study Area by Individual Location and Density.

Once the venues were identified, a trolling algorithm was utilized to collect check-ins for the creation of an hourly rate for each venue during the analysis period, Tuesday, June 11 through Tuesday, July 2, 2012. The data collected included the venue ID, venue name, category, latitude, longitude, number of check-ins per hour, and the number of unique users. An initial analysis of the check-ins was performed to verify that categories were assigned to each venue. These categories, shown in Table 1, include Arts & Entertainment, College & University, Food, Professional & Other Places, Nightlife Spots, Residences, Great Outdoors, Shops & Services, and Travel & Transport. Since categories are assigned by venue creators and are optional, some venues did not have category assignment. For these venues, a key word search was performed to assign the appropriate primary category, when possible.

TABLE 1: Foursquare Category Venue and Check-in Statistics.

Category	# of Venues	Percentage	# of Check-ins	Percentage	Avg. # Check-ins per Venue
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 Spots	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 1 provides a categorical breakdown of the number of venues and check-ins collected. Of the ten venue categories, the Shops & Services category has the largest percentage of venues, while the least is associated with the Nightlife Spots category. Check-ins are most frequently associated with the Shops & Services and the Food categories, which account for 51.3% of all of the check-ins. The Residences category has the least number of check-ins at 2.7% and a moderately low number of venues within the sample size. Average number check-ins per venue was also calculated for each category, with the largest average number of check-ins coming from the Nightlife Spots category (1224 check-ins) and the least coming from the Unclassified category (67 check-ins). It should be noted that the top three average check-ins where in the previously mentioned Nightlife Spots, as well as the Food and Travel & Transport categories. While the Nightlife Spots and Food categories are to be expected as they are social activities, the large number of Travel & Transport check-ins is unexpected. Additionally, due to the low percentage (1.5%) of check-ins for the Unclassified venue category, it was determined that their removal from the study would be without negatively impacting the analysis.

MODEL CALIBRATION

For the calibration of the proposed model, a genetic algorithm was implemented. This algorithm within MATLAB optimizes through the mimicking of the principles of biological evolution via the repeated modification of a population of individual points using rules modeled on gene combinations in reproduction. This optimization strategy was selected for the improved chances of finding a global solution due to the algorithm's random nature. Within the algorithm's calculations, "individuals" are randomly selected from the current "population" and used as "parents" of the "children" for the next generation. This process is repeated and the population eventually "evolves" toward an optimal solution.

The genetic algorithm was used to obtain parameters for the friction function, and the production and the attraction calculations that would in turn minimize the mean absolute error (MAE) between the modeled OD matrix and the reference CAMPO OD matrix. To evaluate the performance of these parameters, a coincidence ratio (CR) was used and calculated from the following formula:

$$CR = \frac{\sum_i min(p_w^M, p_w^O)}{\sum_i max(p_w^M, p_w^O)} \quad (14)$$

where
p_w^M: the percentage of trips within the trip length interval w in the predicted trips from the check-in data, where the trip length interval is used to aggregate the trip counts with an aggregation interval of one mile (1.609 km).
p_w^O: the percentage of trips within the trip length interval w in the survey trips from CAMPO.

The value for the CR ranges from 0, when the distributions are completely different, to 1, when the distributions are exactly the same. For this study, the higher the CR between the check-in and the CAMPO results for each model, the better the model. Table 2 provides the results from the genetic optimization. In general, the calibrated parameters are similar except for the attraction scaling factor ε and the friction factor function parameters θ and σ. Significant improvement can be observed for the CR and MAE values from the proposed model.

TABLE 2: Genetic Optimization Parameters

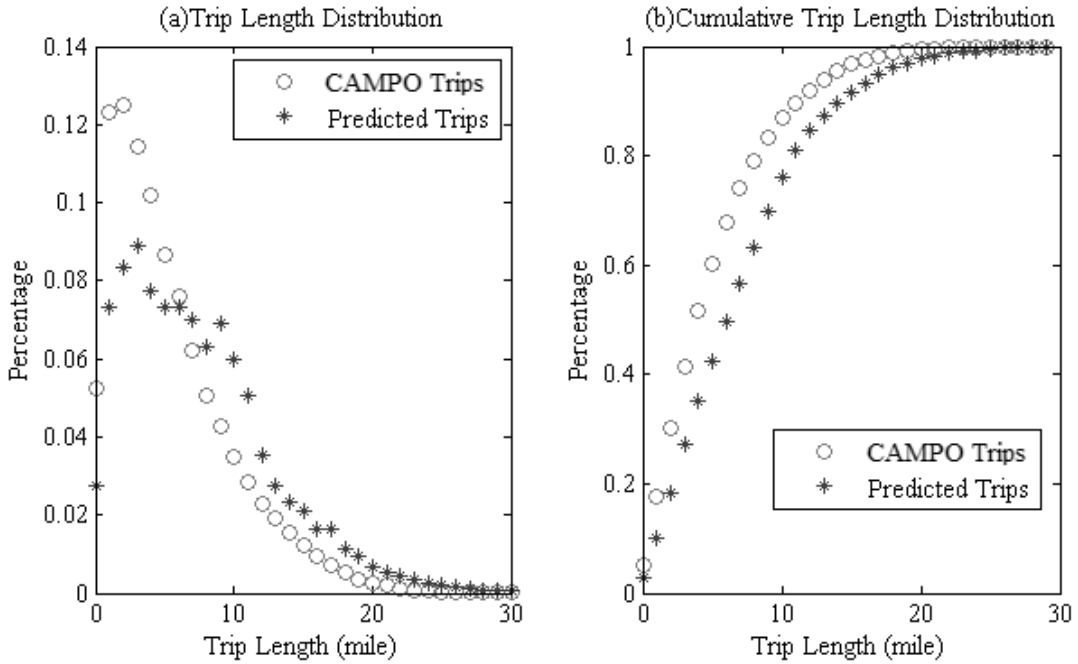
Parameter	Singly-Constrained	Doubly-Constrained
γ	1.02690	0.47334
ε	1.74412	0.66967
η	N/A	0.21198
θ	0.00100	0.16755
λ	0.01252	0.04407
μ	1.51817	2.05600
ρ	0.00283	0.00438
σ	11.18205	5.22909
CR	0.7456	0.9523
MAE	15.9348	10.2134

EXPERIMENTAL EVALUATION

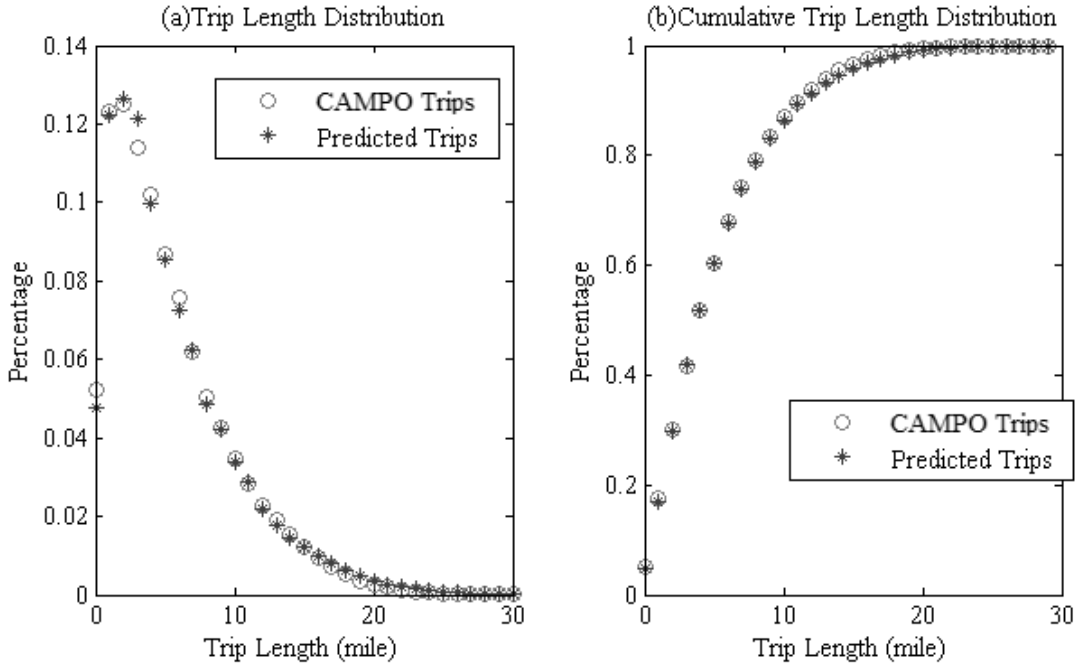
A comparison between the calibrated singly-constrained, doubly-constrained, and CAMPO OD matrices was done by examining the trip length distributions, the zonal trip production and attraction rates, and the zonal OD flow patterns.

Trip Length Distributions

Similar to the coincidence ratio, trip length distribution curves were examined to illustrate how closely the model output data matches the reference data. Figure 3 shows the trip length distributions (a) and the cumulative trip length distributions (b) for the singly- and doubly-constrained models compared with the reference CAMPO OD matrix. Examination of the Trip Length Distribution portion of the figure shows the doubly-constrained model (Figure 3b) is relatively constant with respect to the general curvature. However, for the singly-constrained model (Figure 3a), under estimation occurs for short trips and slight over estimation occurs for long trips. For the singly-constrained cumulative distribution figure, slight under estimation is consistently shown for the curve. While the curves do follow generally the same paths, the deviations indicated lend themselves to further fine tuning of this method.



(a) Singly-Constrained Model Trip Length Frequency Results



(b) Doubly-constrained Model Trip Length Frequency Results

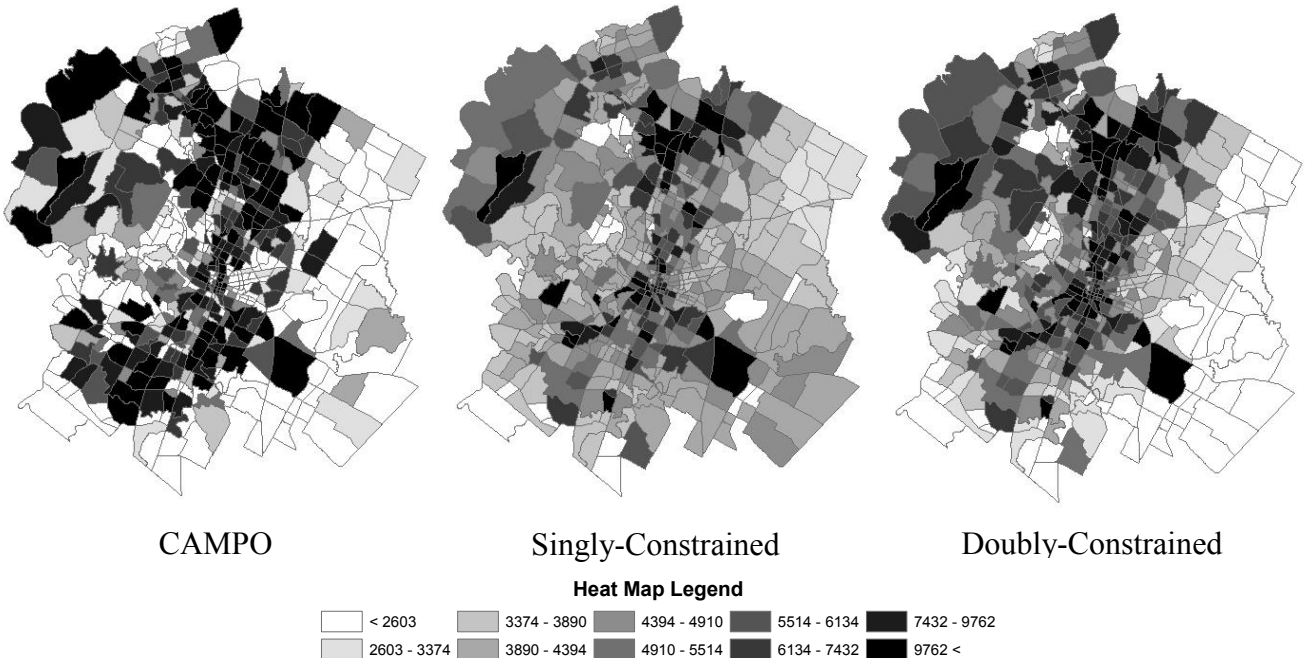
FIGURE 3 Trip Length Distributions for Doubly Constrain

Zonal Production and Attraction Rates

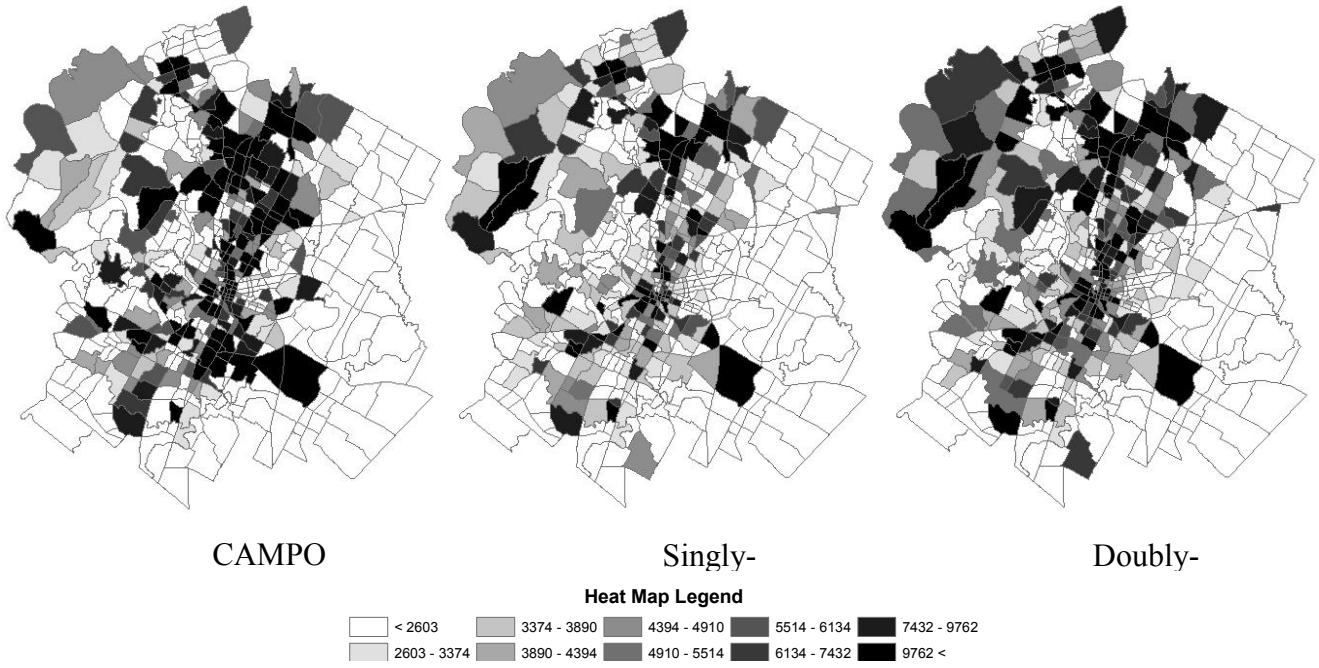
To determine the validity of the methods used to associate the check-ins to the various venues throughout the study area, heat maps were created showing the productions and attractions for each model. Figure 4a demonstrates where the methodology excels and where there are limitations for the production calculations. Using the CAMPO production map as a

reference map, the singly-constrained model shows high production areas that are significantly less in number. Additionally, the singly-constrained model shows mid-level production area through the study region while the CAMPO map is more polarized. Conversely, the doubly-constrained map shows production rates that are similar in magnitude to the CAMPO map through the study region where TAZs that include the central business district, airport, as well as areas dense with living, entertainment, retail, and food venues are consistently depicted as large production generators.

Figure 4b provides heat maps for the attractions for each of the models, highlighting where the methodology excels and where there are limitations for the attraction calculations. Once again using the CAMPO attraction map as the reference, the singly-constrained model predicts attraction rates similar to the CAMPO model for many areas, but suffers from the inability to associate high attraction rates to all of the TAZs identified within the CAMPO map. The doubly-constrained map demonstrates the models ability to better identify areas with high attraction rates. However, the map highlights areas where over estimation occurs, namely in the northwestern portion of the map. It should be noted that although CAMPO data are used as a reference, the data still has its limitations, for example the high variations in trip frequencies among zones potentially caused by under- or over- sampling in certain zones.



(a) Production Comparison Maps



(b) Attraction Comparison Maps

FIGURE 4 Production and Attraction Comparison Maps.

Model Matrix Comparison

The next step in the analysis of the methodology was to examine the zonal flow pattern for each model, which can be regarded as the visualization of the OD matrices. Destination zones are located along the horizontal axis, while origin zones are along the vertical axis. The OD flow intensity, I_{ij}, is calculated using the following equation:

$$I_{ij} = log_{10}\left(\frac{T_{ij}}{\sum_m \sum_n T_{mn}}\right) \quad (15)$$

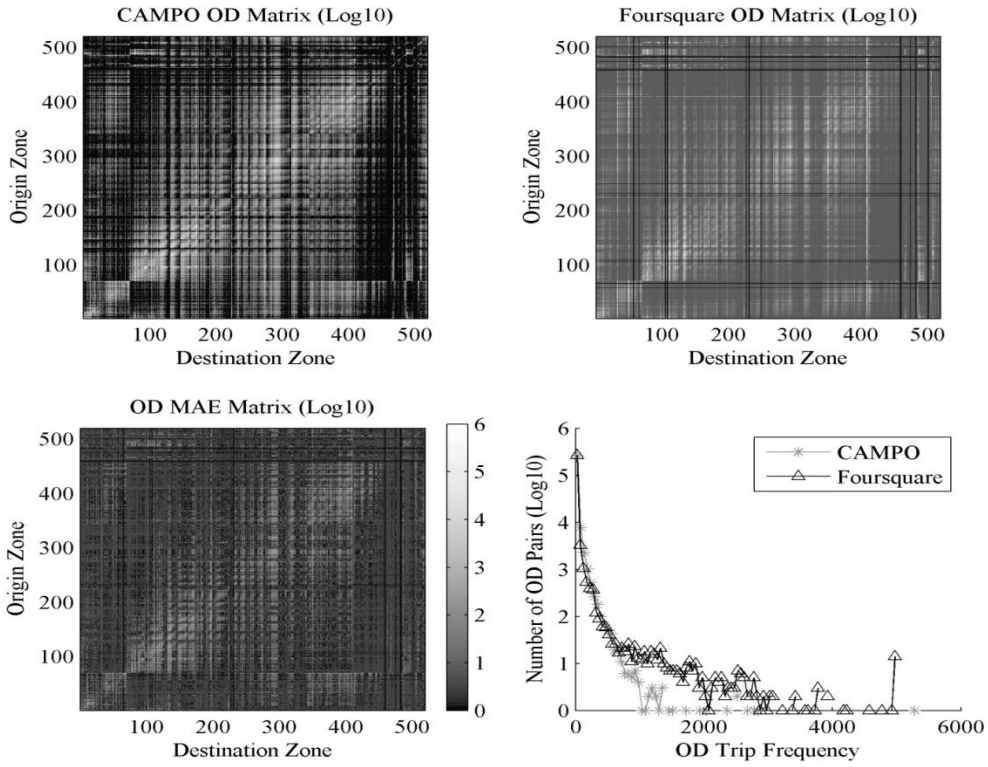
We use the OD heat maps below to provide an illustration of the distribution of trip intensities among TAZs. Each grid (i, j) in the OD heat map indicates the I_{ij} value from zone i to zone j with higher values illustrated by lighter colors. A light horizontal or vertical band indicates a high production or attraction zone and light areas indicate heavily-interacting zones. A difference diagram is also depicted to show how the estimation error distributes among different OD pairs. Overall, the heat maps provide a more detailed description of OD patterns and error distributions as compared to a single performance measure.

Singly-Constrained Gravity Model Based Method

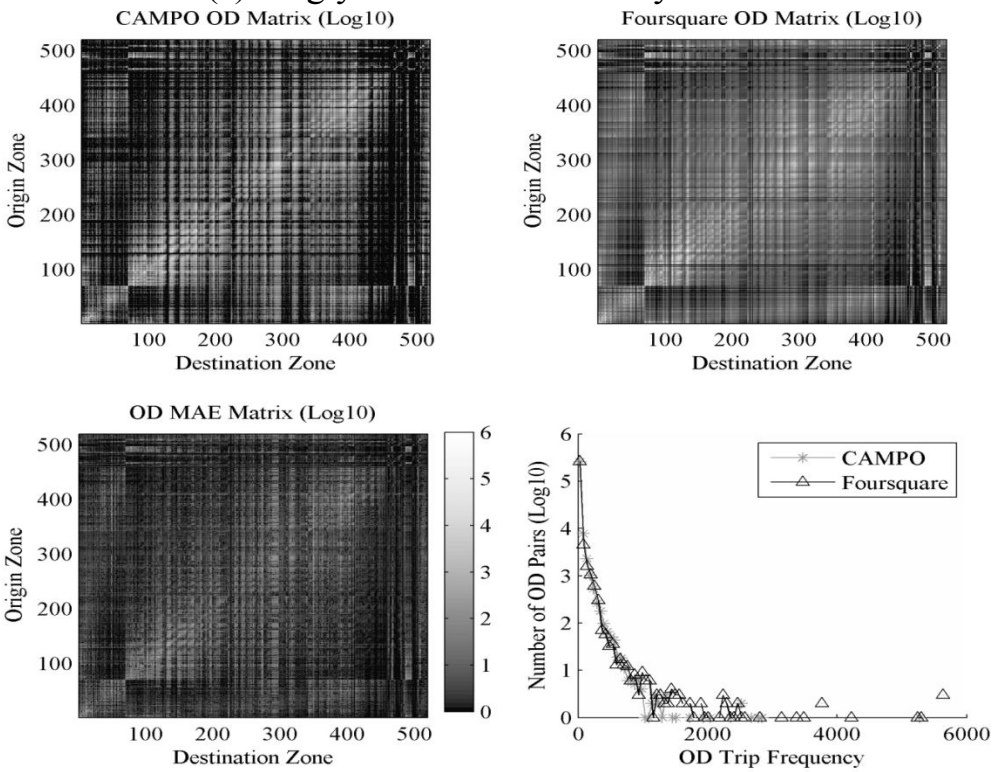
Figure 5a compares the OD flow patterns between the CAMPO OD and the singly-constrained gravity model matrices. Comparing the CAMPO and Foursquare matrices, the flow patterns demonstrate similarities between the two models. While the areas of higher flow are reasonably consistent in the Foursquare model, the areas with low flow are not as prevalent. This is consistent with the less variegated productions and attractions shown within Figures 4. Additionally, the mean absolute error (MAE) matrix is provided to demonstrate how closely the estimate Foursquare matrix matches the C**A**MPO matrix.

Doubly-Constrained Gravity Model Based Method

Figure 5b compares the OD flow pattern between the CAMPO OD and the doubly-constrained gravity model matrices. Comparing these matrices, the flow patterns demonstrate similarities between the two models consistent with what was shown in the singly-constrained model. The doubly-constrained model shows greater flow along the inter-zonal diagonal when compared to both the CAMPO reference OD and the singly-constrained model output. Additionally, the doubly-constrained model has a more variegated color pattern through the diagram, which is consistent with the reference CAMPO OD pattern and coincides with the coincidence ratio for the doubly-constrained model being closer to one than the singly-constrained model. Additionally, the MAE matrix demonstrates how closely the estimate Foursquare doubly-constrained matrix matches the CAMPO matrix. The proposed method still has significant error at the diagonals of the OD matrix indicating issues with intra-zonal trip intensity estimation.



(a) Singly-Constrained Gravity Model.



(b) Doubly-Constrained Gravity Model.

FIGURE 5 Gravity Model OD Matrix Comparison.

CONCLUSION

This paper investigates the feasibility of using the location-based social networking (LBSN) data to analyze the urban travel demand pattern using a doubly-constrained gravity model. Check-in data from Foursquare, a leading LBSN provider, was used to create production and attraction rates for the singly- and doubly-constrained gravity models, which were used in conjunction with the CAMPO OD matrix to examine the predictability of the proposed methodology.

In comparison to the traditional methods used for OD estimation, this study shows that LBSN data has potential. LBSN data is a low-cost option for updating OD matrix since the only cost comes from the purchasing of historical data from Foursquare, Twitter, and/or other LBSN data vendors. The OD matrix can be updated annually, monthly, or weekly depending on the MPO's requirement. Compared with the prevailing secondary data methods based on GPS, Bluetooth, and Cellphone, the LBSN data has user-confirmed trip purposes and destinations eliminating the need for conducting reverse-geocoding and recurrent trip pattern recognition. Furthermore, due to its intensive spatial and temporal coverage, LBSN data has the potential to become a promising dynamic travel demand data source for Active Traffic and Demand Management (ATDM) solutions *(37)*.

LBSN data also has its bias for different venue types (e.g. residential areas, recreational locations, and tourist attractions). In comparison to the existing singly-constrained gravity model based method, the proposed doubly-constrained model based method demonstrates better learning capabilities. There are some limitations with the proposed methodology that should be examined in future research. The model results still indicate some geographical bias for tourist regions (i.e. the northwest region in Figure 4) and residential areas. Additionally, the estimated OD matrix still has significant errors for intra-zonal trips (the diagonal in Figure 5). Further examination into the temporal aspects of the models as well as specific trip purposes should be researched to further validate this proposed methodology.

ACKNOWLEDGEMENT

The authors would like to thank Foursquare® for allowing the research team to obtain data through their developer API as well as Capital Area Metropolitan Planning Organization (CAMPO) for providing the 2010 OD data used within this study.

REFERENCES

1. Abrahamsson, T. Estimation of Origin-Destination Matrices Using Traffic Counts–a Literature Survey. *IIASA Interim Report IR-98-021/May*, Vol. 27, 1998, pp. 76.
2. Fisk, C. Trip Matrix Estimation from Link Traffic Counts: The Congested Network Case. *Transportation Research Part B: Methodological*, Vol. 23, No. 5, 1989, pp. 331-336.
3. Fisk, C. S. and D. E. Boyce. A Note on Trip Matrix Estimation from Link Traffic Count Data. *Transportation Research Part B: Methodological*, Vol. 17, No. 3, 1983, pp. 245-250.
4. Van Zuylen, H. J. and L. G. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. *Transportation Research Part B: Methodological*, Vol. 14, No. 3, 1980, pp. 281-293.
5. 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.

1 6. Giaimo, G., R. Anderson, L. Wargelin and P. Stopher. Will It Work? *Transportation Research*
2 *Record: Journal of the Transportation Research Board*, Vol. 2176, No. 1, 2010, pp. 26-34.
3 7. Bricka, S., J. Zmud, J. Wolf and J. Freedman. Household Travel Surveys with Gps.
4 *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2105, No.
5 1, 2009, pp. 51-56.
6 8. Wolf, J., R. Guensler and W. Bachman. Elimination of the Travel Diary: Experiment to Derive
7 Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record:*
8 *Journal of the Transportation Research Board*, Vol. 1768, No. 1, 2001, pp. 125-134.
9 9. Cho, E., S. A. Myers and J. Leskovec. Friendship and Mobility: User Movement in Location-
10 Based Social Networks. *Proceedings of the 17th ACM SIGKDD international conference on*
11 *Knowledge discovery and data mining*, 2011, pp. 1082-1090.
12 10. Yang, F., Peter J. Jin, Yang Cheng, and Bin Ran. Origin-Destination Estimation for Non-
13 Commuting Trips Using Location-Based Social Networking Data. *International Journal of*
14 *Sustainable Transportation*, Accepted. 2014, pp.
15 11. Bricka, S. Non-Response Challenges in Gps-Based Surveys. *Resource paper prepared for*
16 *the International Steering Committee on Travel Survey Conference, Available online:*
17 *http://ganymede. nustats.*
18 *com/nustats_dot_com/templates/yet_again_newmenu/docs/great_reads/Nonresponse_GPS_Base*
19 *dSurveys. pdf*, 2008,
20 12. Erlander, S., S. Nguyen and N. F. Stewart. On the Calibration of the Combined Distribution-
21 Assignment Model. *Transportation Research Part B: Methodological*, Vol. 13, No. 3, 1979, pp.
22 259-267.
23 13. Fontaine, M. D. and B. L. Smith. Investigation of the Performance of Wireless Location
24 Technology-Based Traffic Monitoring Systems. *Journal of transportation Engineering*, Vol.
25 133, No. 3, 2007, pp. 157-165.
26 14. Caceres, N., J. Wideberg and F. Benitez. Deriving Origin Destination Data from a Mobile
27 Phone Network. *Intelligent Transport Systems, IET*, Vol. 1, No. 1, 2007, pp. 15-26.
28 15. Pan, C., J. Lu, S. Di and B. Ran. Cellular-Based Data-Extracting Method for Trip
29 Distribution. *Transportation Research Record: Journal of the Transportation Research Board*,
30 Vol. 1945, No. 1, 2006, pp. 33-39.
31 16. Herrera, J. C., D. B. Work, R. Herring, X. J. Ban, Q. Jacobson and A. M. Bayen. Evaluation
32 of Traffic Data Obtained Via Gps-Enabled Mobile Phones: The *Mobile Century* Field
33 Experiment. *Transportation Research Part C: Emerging Technologies*, Vol. 18, No. 4, 2010, pp.
34 568-583.
35 17. Blogg, M., C. Semler, M. Hingorani and R. Troutbeck. Travel Time and Origin-Destination
36 Data Collection Using Bluetooth Mac Address Readers. *Australasian Transport Research Forum*
37 *(ATRF), 33rd, 2010, Canberra, ACT, Australia*, 2010,
38 18. Brennan Jr, T. M., J. M. Ernst, C. M. Day, D. M. Bullock, J. V. Krogmeier and M.
39 Martchouk. Influence of Vertical Sensor Placement on Data Collection Efficiency from
40 Bluetooth Mac Address Collection Devices. *Journal of Transportation Engineering*, Vol. 136,
41 No. 12, 2010, pp. 1104-1109.
42 19. Hainen, A. M., J. S. Wasson, S. M. Hubbard, S. M. Remias, G. D. Farnsworth and D. M.
43 Bullock. Estimating Route Choice and Travel Time Reliability with Field Observations of
44 Bluetooth Probe Vehicles. *Transportation Research Record: Journal of the Transportation*
45 *Research Board*, Vol. 2256, No. 1, 2011, pp. 43-50.

20. Barceló, J., L. Montero, L. Marquès and C. Carmona. Travel Time Forecasting and Dynamic Origin-Destination Estimation for Freeways Based on Bluetooth Traffic Monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2175, No. 1, 2010, pp. 19-27.
21. Bricka, S. Travel Behavior Data to Support Transportation Planning. *Urban Transportation Planning*, 2013,
22. Utsunomiya, M., J. Attanucci and N. Wilson. Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1971, No. 1, 2006, pp. 119-126.
23. Tornero, R., J. Martínez and J. Castelló. A Multi-Agent System for Obtaining Dynamic Origin/Destination Matrices on Intelligent Road Networks. *Proceedings of the 6th Euro American Conference on Telematics and Information Systems*, 2012, pp. 157-164.
24. Li, N. and G. Chen. Analysis of a Location-Based Social Network. *Computational Science and Engineering, 2009. CSE'09. International Conference on*, 2009, pp. 263-270.
25. Zheng, Y., X. Xie and W.-Y. Ma. Geolife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.*, Vol. 33, No. 2, 2010, pp. 32-39.
26. Karimi, H. A. Genetic Location-Based Social Networks (G-Lbsn). *Proceedings of the 3rd International Workshop on Location and the Web*, 2010, pp. 9.
27. Backstrom, L., E. Sun and C. Marlow. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. *Proceedings of the 19th international conference on World wide web*, 2010, pp. 61-70.
28. 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.
29. Cheng, Z., J. Caverlee, K. Lee and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. *ICWSM*, Vol. 2011, 2011, pp. 81-88.
30. Scellato, S., A. Noulas, R. Lambiotte and C. Mascolo. Socio-Spatial Properties of Online Location-Based Social Networks. *ICWSM*, Vol. 11, 2011, pp. 329-336.
31. Jin, P. 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. *Transportation Research Board 92nd Annual Meeting*, 2013,
32. Martin, W. A., N. A. McGuckin, N. A. McGuckin and N. A. McGuckin *Travel Estimation Techniques for Urban Planning*. National Academy Press Washington, DC, 1998.
33. Us Census Bureau (Uscb) Accessed 12 July.
34. Clrsearch Accessed 12 July.
35. Ignite Social Media Accessed 28 June
36. Howden, L. M. and J. A. Meyer *Age and Sex Composition: 2010*. US Department of Commerce, Economics and Statistics Administration, US Census Bureau, 2011.
37. Schreffler, E. N. *Integrating Active Traffic and Travel Demand Management: A Holistic Approach to Congestion Management*. 2011.