The Thesis Committee for Meredith Kimberly Cebelak
Certifies that this is the approved version of the following thesis:

Location-based Social Networking Data: Doubly-Constrained Gravity Model Origin-Destination Estimation of the Urban Travel Demand for Austin, TX

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

C. Michael Walton

Peter J. Jin

Location-Based Social Networking Data: Doubly-Constrained Gravity Model Origin-Destination Estimation of the Urban Travel Demand for Austin, TX

by

Meredith Kimberly Cebelak, B.S.C.E.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

August 2013

Dedication

To Braunson Cebelak, for all the long nights and weekends where you have patiently waited for me to conclude my day you are forever owed.

Acknowledgements

I would like to acknowledge Peter Jin, for the origination and development of the concept of using check-in data for the creation of origin-destination matrices as well as his assistance in developing the MATLAB code used within this analysis.

Abstract

Location-Based Social Networking Data: Doubly-Constrained Gravity Model Origin-Destination Estimation of the Urban Travel Demand for Austin, TX

Meredith Kimberly Cebelak, M.S.E.

The University of Texas at Austin, 2013

Supervisor: C. Michael Walton

Populations and land development have the potential to shift as economies change at a rate that is faster than currently employed for updating a transportation plan for a region. This thesis uses the Foursquare location-based social networking check-in data to analyze the origin-destination travel demand for Austin, Texas. A doubly-constrained gravity model has been employed to create an origin-destination model. This model was analyzed in comparison to a singly-constrained gravity model as well as the Capital Area Metropolitan Planning Organization's 2010 Urban Transportation Study's origin-destination matrices through trip length distributions, the zonal origin-destination flow patterns, and the zonal trip generation and attraction heat maps in an effort to validate the methodology.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Trip distribution is a significant portion of the four-step transportation planning process. The data used for the creation of origin-destination (O-D) matrices conventionally come from the traditional household survey. These surveys are often time consuming and can be an expensive endeavor. The use of traffic counts is another data collection method that can be implemented, but requires a detector infrastructure that can be expensive to install and maintain. This method is an O-D matrix updating method that assumes the existence of a baseline O-D matrix generated by other methods. Additionally, to properly gather the data needed, detectors would need to be on all viable routes between O-D pairs. Current technological advances in positioning technologies allow for use of global positioning systems (GPS), cell phones, and Bluetooth for data collection. However, these technologies have limitation as well. GPS is limited by sample size biases, concerns about privacy, as well as the time and labor costs for data collection. Cell phone based tracking is subject to significant privacy concerns from which only anonymous location information is available. This issue, in combination with the location error, makes it very difficult to confirm trip purposes for a cell phone trajectory. Bluetooth requires a dense number of readers and suffers from a low sampling rate penetration attributed to the willingness of users to turn on or off the technology.

In tandem with the current technological advances in positioning technologies, location-based services features have become available with smartphones and tablets. These devices have seen an increase in accessibility and affordability to a variety of income levels in recent years. Additionally, the rapid development of social networking

sites, like Twitter© and Facebook©, has led to the availability of location-based social networking (LBSN) data. Recent research efforts have studied the spatial patterns of LBSN user behavior through the mining of social networking sites. LBSN data has the potential to provide O-D estimates with a higher temporal resolution at a lower cost when compared to traditional methods. This data also has the ability to confirm trip purposes making it superior to pure trajectory based methods.

1.1 PROBLEM STATEMENT

Populations and land development have the potential to shift as economies change at a rate that is faster than rate for updating a transportation plan for a community. However, current methodologies for data collection used in transportation planning often take a year to gather into a useable format and can be costly efforts.

This thesis will use Foursquare location-based social networking check-in data to analyze the origin-destination travel demand for the urban area in Austin, Texas. Foursquare is an application (or app) that allows individuals to share the places that are visited with friends via checking-in. Using the check-in data collected, a doubly-constrained gravity model will be used to estimate the origin-destination demand. The evaluation of this effort will be conducted using businesses and other locations throughout the area in and around the city of Austin, Texas

The results of the doubly-constrained gravity model will be compared to the Capital Area Metropolitan Planning Organization's (CAMPO) 2010 Urban Transportation Study's origin-destination data through the analysis of the trip length distributions, the zonal origin-destination flow patterns, and the zonal trip generation and attraction heat maps. The results of this analysis will verify the functionality of the use

of the singly- and doubly-constrained gravity model methodologies for demand generation from check-in data.

1.3 THESIS SUMMARY

This thesis uses a fourth-generation programing language, MATLAB (Matrix Laboratory) for the calculations of the origin-destination demand based on Foursquare check-ins for the city of Austin, Texas to evaluate the validity of using location-based social networking data via a doubly-constrained gravity model for the creation of O-D matrices. Chapter 2 provides an introduction into the background of the trip generation models as well as existing and proposed data collection methods. Chapter 3 describes the experimental framework including the procedures and methods for the data processing, the preliminary analysis of the characteristics of the check-ins collected, and the calibration of the model. Chapter 4 discusses the resulting origin-destination demand matrix, examines the origin-destination patterns, and compares the resulting singly- and doubly-constrained matrices to the CAMPO matrix for evaluation of the methodology. Chapter 5 provides the conclusion of this effort.

Chapter 2: Background/Literature Review

2.1: TRANSPORTATION PLANNING

Transportation planning has a fundamental role in the future of a community; the earliest highway planning grew out of a need for information on the increased usage of automobiles and trucks (Weiner, 1999). In the era prior to World War II, the focus of transportation planning was on the gathering and analyzing of factual information. Most urban areas did not use urban travel studies in their planning efforts and post-World War II saw a boom in demand for automobiles. However, the existing highway infrastructure was ill equipped to the increased demand. This boom in demand, the lack of improvements in the existing highway system, and the growth of suburban developments lead to renewed efforts in transportation planning. Trip origins and destinations studies were needed to address the complex urban street systems which lead to the development of the home-interview origin-destination survey in 1944 (Weiner, 1999).

The 1950s brought new ideas and techniques to urban transportation planning including M. Earl Campbell's *Route Selection and Traffic Assignment* and Thomas Fratar's computer method, known as the Fratar method, used the distribution of future origin-destination travel data using growth factors. In 1954, Robert B. Mitchell and Chester Rapkin established a link between travel and activities in a landmark study that called for a thorough framework for and explorations into travel behavior (Mitchell, 1954). From this study, an initial development of a trip generation, distribution, and diversion model lead to the first application of the four-step model which was first applied in 1950 in the Chicago Area Transportation Study (McNally, 2008).

- Trip generation: This step measures trip frequency providing the magnitude of total daily travel in the form of productions and attractions within the system at the zonal and household level for the trip purposes.
- Trip Distribution: This step distributes the productions to match the attractions distribution reflecting the underlying travel impedance resulting in trip tables of person-trip demands for each trip purpose.
- Mode Split: This step factors the trip tables from trip distribution to produce mode-specific trip tables reflecting the choice probabilities of individual trip makers. The disaggregate results must be aggregated to the zonal level prior to the traffic assignment step.
- Traffic Assignment: This step applies the mode split trip matrices to the modal network under the assumption of user equilibrium where all paths utilized have equal impedance.

Federal legislation in the 1960s required "continuous, comprehensive, and cooperative" urban transportation planning (McNally, 2008). In the 1970s, environmental concerns and multimodal elements were brought into planning efforts.

The modern day transportation planning process is designed to analyze potential strategies through an evaluation process that incorporates the viewpoints of transportation-related agencies, organizations including Metropolitan Planning Organizations (MPO) and the general public (The Transportation Planning Process, 2007). The process includes steps to monitor existing conditions, forecast the future population and employment growth including accessing project land uses and identifying major growth corridors, identifying current and future transportation problems and needs, the development of long-range plans and short-range programs, the estimation of the

impacts of recommended future improvements, and the development of a financial plan for the implementation of strategies. Figure 1 provides FHWA's view of the overall process.

Figure 1: FHWA's Transportation Planning Process (The Transportation Planning Process, 2007).

To aid in the decision making process, transportation planning models are employed by MPOs. These models are used to simulate the impacts of changes to the transportation system (i.e. adding a roadway, increasing populations) based on "real world" conditions. Models used include the traditional land use, emissions, activity-based, and four-step models. Land use models are often employed for forecasting future development patterns, while emissions models examine the key pollutants emitted within the exhaust of vehicles. Activity-based models have been gaining popularity recently in the United States (US); they examine travel through multiple trip legs by chaining trips into tours. This approach is more disaggregated in time, space, and activities which are better suited for analyzing complex policy alternatives, such as flexible work hours and

variable pricing schemes. The traditional four-step model, which this thesis will utilize, is still commonly used at MPOs for the prediction of the demand for transportation services and is made up of the following four steps: trip generation, trip distribution, mode split, and network assignments. The trip generation and trip distribution steps within the four-step model will be the focus areas for this thesis.

2.1.1: Trip Generation

Trip generation is the first step within the four-step model and measures trip frequencies, which are developed by providing the inclination to travel. The earliest trip generation effort was in 1948 in San Juan, Puerto Rico (Weiner, 1999). This transportation study used trip generation to forecast trips and rates were developed for a series of land use categories that were arranged by location, intensity, and type of activity. Following this initial effort, in 1955, the Detroit Metropolitan Area Traffic Study developed trip generation rates by land use category for each zone within their study area.

The Trip Generation Committee of the Institute of Transportation Engineers (ITE) developed a report on trip generation rates in 1972. This committee was tasked with collecting existing trip generation rate data to be compiled into a single source. The first edition of this endeavor was published in 1976 and contained data from nearly 80 sources. The Trip Generation's 5th edition (1991) was considered the first and most comprehensive data base and contained generation rates for 121 land use categories from over 3,000 studies. The current edition, the 9th edition published in 2012, contains data from more than 5,500 studies and 172 land uses (ITE, 2013). The ITE Trip Generation reports are the most widely used reference for transportation professionals for trip generation data for site level planning and analysis (Weiner, 1999).

Within the trip generation process, the goal is to determine the magnitude of total daily travel at the household and zonal level for the trip purposes included within the study. These trip purposes typically include a minimum of three types: home-based work, home-based non-work, and non-home-based (McNally, 2007). Trip end points are modeled as either productions or attractions for use within the model.

2.1.2: Trip Distribution

After attaining the trip generations for the study area, a process to recombine the production and attraction rates for each traffic analysis zones (TAZs) into trips is undertaken. This process is called trip distribution and creates a model, or matrix, of the number of trips occurring between each origin and destination zone (McNally, 2007). To recombine the productions and attractions, models are employed. These models include but are not limited to logit, entropy, growth and gravity. Gravity models will be discussed in depth within the next sections.

The logit models include the multinomial logit destination choice model used commonly for activity-based models and the singly constrained logit destination choice models. In 1976, Wilson established a new method of constructing distribution models called the entropy maximizing method. This effort was further clarified in Wilson's "The Use of Entropy Maximising Models in the Theory of Trip Distribution, Mode split and Route Split" (1969) which relates the probability of the distribution of trips occurring in an origin-destination (O-D) pair to the number of states of the system. The growth models are comprised of two variations: uniform and doubly constrained. For the uniform growth model, the only information needed is a general growth rate for the entire study area, which is a simplistic method useful for short-term planning but is limited by the assumption that the growth factor is the same for all zones and attractions. The doubly

constrained growth factor model uses information on the growth of the number of trips originating and terminating in each zone, thus different factors can be implemented. Similar to the uniform model, this model has the advantages of simplicity, but is heavily dependent on observed trip patterns and does not include changes in travel costs within its trip distribution.

2.1.2.1: Gravity Model

The aggregate gravity model originated from an analogy with Newton's gravitational law (Mathew, 2006). Newton's gravitational law states that force (F) is related to the gravitational constant, G, the masses of two objects (m_1, m_2), and the distance between the objects (d). The formula for the relationship is as follows:

$$F = \frac{G * m_1 * m_2}{d} \quad (eqn. 1)$$

This formula is analogous the trip distribution formula, as shown below in the general form, with the number of trips per O-D pair (T_{ij}) component relating to force (F), the C relating to the gravitational constant (G), the productions from zone i (O_i) and the attractions from zone j (D_j) relating to the mass entries (m_1, m_2), and the travel cost between O-D pairs (c_{ij^n}) relating to the distance between objects (d).

$$T_{ij} = \frac{C * O_i * D_j}{c_{ij^n}} \quad (eqn. 2)$$

In an effort to ensure that the total number of productions and attractions are equal, a balancing factor (b) is added to the previous formula to either the productions or attraction factors making the equation (below) a singly constrained model, which attempts to preserve zonal inputs for the productions only (TMIP, 2010).

$$T_{ij} = b * O_i * D_j * f(c_{ij}) \quad (eqn. 3)$$

Additionally, a friction function ($f(c_{ij})$) to de-incentivize travel based on time via distance or cost increases replaces the general travel cost term (c_{ij^n}). Friction functions are discussed in more detail in a subsequent section of this chapter.

2.1.2.2: Doubly Constrained Gravity Model

The doubly constrained gravity used within this thesis attempts to preserve zonal inputs for the productions and attractions (TMIP, 2010). This model builds upon the singly constrained model and encompasses balancing factors for both the productions and the attractions (Mathew, 2006):

$$T_{ij} = \beta_i * O_i * \alpha_j * D_j * f(c_{ij}) \quad (eqn. 4)$$

The balancing factor for the productions is defined by β_j, while the balancing factor for the attractions is defined by α_i. Using the principle that the sum of the total trips for each destination is equal to the sum of the combination of productions, attractions, the balancing factors, and friction functions for each destination, the above formula can be manipulated to give formulas for the balancing factors as shown.

$$\sum_i T_{ij} = \sum_i \beta_i * O_i * \alpha_j * D_j * f(c_{ij}) \quad (eqn. 5)$$

which is equivalent to

$$\sum_i T_{ij} = D_j \quad (eqn. 6)$$

Thus,

$$D_j = \beta_i * D_j \sum_i \alpha_j * O_i * f(c_{ij}) \quad (eqn. 7)$$

from which the balancing factors (β_i, α_j) can be found

$$\beta_i = \frac{1}{\sum_j \alpha_j * A_j * f(t_{ij})} \quad (eqn. 8)$$

Similarly, for each destination zone, the balancing factors become the following.

$$\alpha_j = \frac{1}{\sum_i \beta_i * P_i * f(t_{ij})} \quad (eqn. 9)$$

Using these factors and separate singly constrained models, the following formulas can be used to find the T_{ij} for each O-D pair from this model.

$$T_{ij} = O_i * \frac{D_j * f(t_{ij})}{\sum_j D_j * f(t_{ij})} \quad (eqn. 10)$$

$$T_{ij} = A_j * \frac{P_i * f(t_{ij})}{\sum_i P_i * f(t_{ij})} \quad (eqn. 11)$$

These formulas can be solved via an iteration process similar to the Furness method (Mathew, 2006).

2.1.2.3: Atomistic Gravity Model

The atomistic gravity model, as currently implemented by CAMPO, is a triply-constrained model with constraints on the productions, attractions, and the trip length frequency. The trip length frequency constraint makes the model self-calibrating for both intra-zonal and inter-zonal trips. Within this model, the TAZs are represented by an abstract discrete spatial surface made up of 400 "atoms" that are evenly disbursed throughout the TAZ (TTI 2001). The model uses travel time rather than distance for the

zonal radii, which are then used along with the zonal centroid-to-centroid travel times for the estimation of the spatial distribution of the atom pairs. The basic formula used for the atomistic model is as follows:

$$T_{ij} = O_i * \frac{\sum_{v=1}^{Mi} \sum_{q=1}^{Mj} p_{i_v} * a_{j_q} * F_{d_{vq}} * K_{s_{ij}}}{\sum_{x=1}^{N} \sum_{n=1}^{Mj} \sum_{m=1}^{Mx} p_{i_n} * a_{x_m} * F_{d_{nm}} * K_{s_{ix}}} \quad (eqn. 12)$$

For this equation p_{i_v} is the trips produced by atom v of zone i, a_{j_q} is the relative attraction factor for atom q of zone j, $F_{d_{vq}}$ is the relative trip length factor for the estimated separation between atom pair vq, $K_{s_{ij}}$ is the bias factor for sector pair containing zones i and j, M_y is the number of atoms in zone y, and N is the number of zones. In order to calculate the O_i, the following formula is used:

$$O_i = \sum_{m=1}^{Mi} p_{i_m} \quad (eqn. 13)$$

2.1.2.4: Friction Functions

A friction function ($f(c_{ij})$) de-incentivizes travel based on time via distance or cost increases and is included within the gravity models previously discussed. This "deterrence function" (Mathew, 2006) can use a variety of formulations to appropriately calculate the impedance. The potential formulas include the linear function, negative exponential, power, and gamma function as shown in generic form below (Bossard, 1993, Mathew, 2006).

$$\text{Linear:} \quad f(c_{ij}) = \text{A} + \text{B} * d_{ij} \quad (eqn. 14)$$

$$\text{Negative exponential:} \quad f(c_{ij}) = \text{A}e^{-\text{B}*d_{ij}} \quad (eqn. 15)$$

$$\text{Power:} \quad f(c_{ij}) = {d_{ij}}^{-n} \quad (eqn. 16)$$

$$\text{Gamma: } f(c_{ij}) = \text{A} * {d_{ij}}^{\text{B}} * e^{\gamma * d_{ij}} \quad (eqn. 17)$$

where A is a positive scaling factor that controls the overall range of the function values, B is either a positive or negative constant value that affects the distribution of shorter trips, *n* is a positive or negative constant value that affects the distribution of trips, γ is a parameter of transport friction relating to the efficiency of the transportation system between two locations and always negatively affects the distribution of longer trips, and d_{ij} is the Manhattan distance between the centroids of origin zone *i* and destination zone *j* in miles.

2.2: ORIGIN-DESTINATION DATA COLLECTION

Data collection is a fundamental component for the creation of O-D matrices. Currently, there are a variety of methodologies employed for data collection. These include the traditional household survey, traffic counts, position technology, "big" data and LBSN sources.

2.2.1: Household Survey Methods

Conventionally, data for O-D matrices has come from traditional household travel behavior surveys. This survey type collects data on trip purpose, mode of transportation used, duration of trip, time of day and day of the week the trip took place, vehicle occupancy for personal vehicles, as well as personal information which includes age, sex, employment status, income, and education level (NHTS, 2013). This data can be collected via personal home interviews, telephone interviews, by mail, or by internet.

Personal home interviews require an interviewer to visit the respondent's home or office to administer questions in a face-to-face interview (Sharp, 2005). In comparison

to other methods covered within this section, it provides the most complete data set with the highest response rate, 60-70% according to Giaimo et al (2010). While the data attained from this method is of the highest quality compared to other methods, conducting this type of survey is the most expensive and time consuming.

The telephone interviews require interviewers to contact individuals and administer the survey over the telephone. Coverage for this method is limited to only households with telephones inducing a sample bias. For this method, the response rate, reported as 25-40% by Giaimo et al. (2010), is intermediate as is the quality of the data and the cost.

The mail survey format requires a questionnaire to be mailed out to respondents with the results returned either by mail or telephone. While the coverage for this method is similar to that of the personal home interview method, it has the lowest response, 20-30% (Giaimo, 2010), and data quality rates for this survey method. The cost for this method is one of the least expensive of the deployment methods for the household surveys.

The internet method of survey deployment is similar to the mail format, but places the survey on the internet for respondents to complete. However, only households with internet access are able to participate once again limiting the coverage of the survey. The response rate is similar to that of the mail survey methodology and the data quality is intermediate. Costs for this method are low, but there are higher startup costs in comparison to the other data-collection cost.

2.2.2: Traffic Count Methods

The use of traffic counts is another data collection method that be implemented for data collection. According to Abrahamsson (1998), the assignments within an O-D

matrix gives the traffic volumes for each transportation link. However, Abrahamsson concedes that there are a large number of different O-D matrices that can reproduce the observed traffic counts. In 1979, Erlander, Nguyen, and Stewart were able to show that if traffic counts were available for all links within the study area a unique O-D matrix could be calculated. This method requires the deployment of detector infrastructure throughout the study area on all viable routes between O-D pairs.

Further research was conducted throughout the 1980s within this area. Fisk and Boyce (1983), in recognizing only a sample of traffic count data is typically available, proposed a procedure for estimating the link cost functions and formulated a doubly constrained distribution-assignment model which was then extended by Kawakami, Lu, and Hirobata (1992) to include two travel modes. Fisk (1989) further expanded her research within this area to examine the congested network scenario, which examined three different formulations for the O-D matrix formulation. O-D matrix estimation from observed traffic data was conducted with applications on networks having more than 70 links by Van Zuylen and Willumsen (1980) and by LeBlanc and Farhangian (1982). In 1988, Cascetta and Nguyen utilized traffic counts for both cars and transit to estimate an O-D matrix. The study used classical and Bayesian statistical inference techniques providing a framework for models and algorithms to be developed from.

While in recent years this method has been shown to be feasible in practice (Watson 2006, Doblas 2005), to achieve this feasibility, detectors would need to be installed for full coverage of the network to prevent large data gaps leading to operation and maintenance costs, which are an expensive in the long term commitment. Detectors provide data at fixed locations and there are concerns with the accuracy of the estimated traffic conditions between detectors (Fontaine, 2007). Additionally, as stated in

Abrahamsson's work, a "target" O-D matrix, typically from prior information on the anticipated or existing O-D matrix, is often need to verify this methodology.

2.2.3: Positioning Technology Methods

Advances in position technologies have made GPS, cellular phone, and Bluetooth technologies viable data sources for traffic flow monitoring, traveler information provision, and advanced traffic and demand management. These technologies have been employed by survey researchers both in simulation efforts and field deployments. Currently, research has begun to examine the use of credit card data, navigation services, and the potential for vehicle-to-infrastructure to be used as viable data sources.

2.2.3.1: Global Positioning Systems

The US Military initiated an effort for a satellite-based positioning system in the 1970s, which became the fully operational GPS in 1995 (Sen and Bricka, 2009). Rapid improvements along with the relative low-cost, high accuracy solution to the positioning requirements for travel surveys within the technology lead to the quick adoption by researchers both in the US and internationally (Bricka 2008). The first travel survey effort that utilized GPS was the 1996 proof-of-concept survey in Lexington, Kentucky (Murakami and Wagner, 1999). This study utilized Personal Digital Assistants (PDA) with GPS capabilities to capture vehicle based daily trip information for 100 households over a six day period. The intent of the study effort was to identify an alternative to trip diaries that was cost effective and to determine the willingness of individuals to participate in this data collection strategy. The study demonstrated the additional ability to collect information on route choice as well as travel speed.

Since the initial proof-of-concept effort, numerous studies have examined the use of GPS technology as a data collection tool. Two studies have specifically examined the feasibility of using GPS to replace the traditional data collection survey efforts. Wolf et al. (2001) noted that while several studies had used GPS as a supplement to the traditional data collection efforts, the effort to examine the feasibility of using GPS as a replacement data collection effort had not been done. Their small scale proof-of-concept effort collected personal vehicle travel data via GPS and used a geographic information system (GIS) to derive the traditional diary elements. This GPS data was then compared with the paper diary data elements and was found to match or surpass the paper diary elements indicating the feasibility of the method. The second effort was the 2010 proof-of-concept for the Greater Cincinnati Household Travel Survey (Giaimo et. al.). This study examined the replacement of travel diaries with a large-scale multiday GPS survey and was the first effort of this kind. The survey was made up of a fully representative sample (household size, number of automobiles, income, age, geographic region, etc.) that recorded data for up to three days of travel. The results of this effort showed that completion rates of the survey via GPS were adequate as well as representative and participants were not burdened to carry the devices. It was noted that significant incentives, as well as additional efforts, were needed to conduct the survey effort. Furthermore, there were a number of logistical issues identified. These included the timely retrieval of GPS units, the GPS unit loss rate of 2.7% which was noted mostly among low-income, urban households, and battery outages that resulted in incomplete data. The effort also noted that there were limitations based on the software used with the data. The map editing process required the training of staff to review the data to ensure it was incorporated appropriately. Some discrepancies were the inclusion of

stops within a trip that were misidentified as a single trip, and trips that incorrectly split into two trips due to a loss in GPS signal.

In addition to these pilot studies, research has been conducted using GPS to determine the characteristic of underreporting that occurs within traditional surveys. A study by Bricka and Bhat (2006) conducted a comparative examination of GPS with the traditional household travel survey to determine the likelihood and the level of underreporting. This study concluded that individuals under 30, men, individuals with less than a high school education, unemployed individuals, individuals who make many trips, travel long distances, and trip chain were affiliated more with underreporting. Bricka (2008) identified additional challenges with non-response in GPS surveys, specifically the burden to respondents. These burdens include the length of the survey since the units can be deployed for longer survey durations, privacy concerns, and equipment complications; for appropriate deployment, these burdens would need to be mitigated to the greatest degree possible. Non-responses were found to be from older, less educated, and lower income participants which follows the trend of those more likely to accept technology being male, young, highly educated, and of a higher income status. This non-response has the potential to lead to a sampling bias of younger and nonminority participants, as was shown in the Oregon Household Travel Survey test pilot, suggesting other methods of data collection may be more appropriate (Bricka et al., 2009). This same study also indicated that costs for a GPS-based survey were over twice as expensive as the traditional survey method. However, the study did note that these costs would be expected to decrease as the data collection process was streamlined, new technology became available, and development costs were able to be allocated to a larger sample within a full study.

2.2.3.2: Cellular Phones

As of December 2012, there are 326.4 million wireless active devices, which included smartphones, tablets, and hotspots, within the US (U.S. Wireless Quick Facts, 2013). Moreover, as of May 2013, 91% of US adults age 18 and older have a cellular phone (Brenner, 2013). In addition, wireless location technologies (WLT) are available from wireless carriers. WTL fall into two main categories: mobile based, where the location is determined from signals received from base stations or from GPS, and network based, which relies on an existing network to determine the location by measuring signal parameters at the base station (Sayed, 2005). The Federal Communication Commission (FCC) mandated E911 for all cellular carriers which required that carriers be able to provide a 911 caller's phone number for return calls as well as the location of callers via WTL (Revision, 1997). This ability has inspired transportation researchers to investigate the feasibility of extracting traffic data from the location data of cellular phones.

Yim (2003) examined cellular probe technologies noting that the use of E911 for probe activities introduced privacy concerns and additionally suggested improvements in cellular geolocation technologies would need to be conducted to fully realize the capabilities of the technology. Pan et al. (2006) examined a cellular-based data-extracting method for trip distribution, which was theoretically proven to be feasible as well as experimental feasible. The study showed the method was advantageous in its ability to directly attain the spatiotemporal information about travelers via mobile carrier thus requiring minimal costs and labor. Similarly, Caceres, Wideberg, and Benitez (2007) developed a technique to use the global system for mobile communications (GSM) to derive O-D data. The study used simulated data and produced reasonably

precise estimation results. However, the study did note that data collection can only be done while the phone is powered on. In 2007, Fontaine and Smith found that the simulated effects of using WTL for monitoring traffic overestimated the capabilities of the system when two case studies were simulated. Dense networks with mixed congestion and free flow were show to have mixed results when simulated. The study also noted that WLT data collection would need to be tailored to the specific localized parameters including frequency rates for sampling which may need to be adjusted to account for traffic conditions.

In 2010, two studies showed further promise of WTL. Schlaich, Otterstätter, and Fiedrich developed a method to generate time-space trajectories for travelers through the analysis of cellular phone data from location-area-updates. These location-area-updates are recorded from mobile phones while in the standby-mode. One limitation with using location-area data is that short trips cannot be detected as they may not move between location areas. Another limitation noted within the study was the need to attain the data from only one mobile service providers, which may create a skew within the data. Finally, the trajectories that were able to be produced from this study are only representative of SIM-cards and not vehicles, of particular importance since a vehicle may have any number of SIM-cards onboard. The second study examined the deployment of GPS-enabled mobile phones for traffic monitoring as a proof-of-concept effort (Herrera et al., 2010). With an average penetration rate between 2 and 3% in the study, higher accuracy for velocities were achieved in comparison to loop detectors and it was noted that the data collected could also be used for planning purposes. The penetration rates seen in the study would be sufficient to achieve the spatiotemporal coverage of a network since sensors would be moving throughout the system.

Additionally, it was noted that the costs associated with installation and maintenance were minimal.

Recently, cellular probe data has evolved rapidly over the last decade with upgrades in the wireless communication standards into 3G and 4G, the rise of market domination of smartphones, and integration within social media and cloud computing. Companies, such as AirSage, have partnered with cellular companies to receive wireless signals using them to anonymously determine location (AirSage, 2013). This aggregated location data is time- and date-stamped which can be used to model, evaluate, and analyze the movements and flows of commuters for almost every city in the nation. While this data has good temporal and special coverage, the cost of the data may make it cost prohibitive for usage by many municipalities. Additionally, trip purpose information is not available for this data.

2.2.3.3: Bluetooth

The Bluetooth technology capitalizes on the short-range personal wireless connectivity technologies that allow personal devices to communicate with each other directly without the need for the line of sight requirements of radio frequency based connectivity (Bisdikian, 2001). Since its development 1998, Bluetooth technology has been noted to be a low cost, user friendly way of collecting data (Blogg et al., 2010, Brennan et al., 2010, Hainen et al., 2011). The technology uses a unique media access control (MAC) for each device that makes device tracking possible. These MACs are a unique 48-bit, 12 alphanumeric character address that is assigned to each device by the device's manufacturer, which are not affiliated with a user alleviating privacy concerns.

Research has been conducted to determine the effectiveness of using Bluetooth technology to collect data that can be utilized for O-D matrices. Blogg et al. (2010)

found that the technology was effective in collecting O-D data in small and controlled networks demonstrating the technology compared favorably to video and automated number plate recognition data. This result was substantiated by the 2011 study by Hainen et al., which compared the technology with license plate matching. The Blogg et al. study did note that there were cautions and limitations with the technology, including the requirement of appropriate detector placement, the short ping cycle (approximately 0.1 second), which could lead to a device being detected multiple times as it passed a single detector, and the possibility that multiple MAC addresses could be within a single vehicle.

The 2010 study by Brennan et al. looked to address the concerns with detector placement. The study noted that there were no existing design guidelines for detector placement and the variation in placement locations, both in height and distance from the facility to be monitored, lead to high variance in the number of addresses captured. The study noted that between 5 and 10% of the vehicle population had discoverable MAC addresses that were able to be collected. Additionally, it was noted that there was no relationship between the traffic volume and the efficiency of collection, rather the height of the detector influenced the collection efficiency.

Barceló et al. (2010) examined using Bluetooth technology for dynamic O-D estimations for freeways. This study noted that the variability of the Bluetooth sample collected yielded objectionable expansion errors and subsequently using the technology independent of other methods was too risky to create a straightforward estimation of an O-D matrix.

2.2.4: Additional Methods

Bricka (2013) stated that synthetic data, "Big" data, smartphone data sources were potential sources for travel survey data. Additional areas that are currently being examined include credit card, navigation technologies, and vehicle-to-infrastructure (V2I) data.

Synthetic data is data that is comprised of synthetic households that are generated typically for a block group within a census tract (TMIP, 2013). Each synthetic household is made up of "individuals" that have an associated set of demographics that closely match to the demographics of real households using census data. This data can then be used to estimate trips within the study area.

"Big" data includes transactions, interactions, and observations (Bricka, 2013). Transaction data includes credit card data in the form of purchase and payment records, product/service logs, and dynamic pricing data. Interaction data includes support contacts, web logs, and social interactions and feeds. Observation data includes sensors, spatial and GPS coordinates, and user click stream. It has been noted that this data source will have challenges for incorporation into transportation planning, namely with data capture, sharing, and management and storage. Additionally, Bricka (2013) noted that the data attained may have biases, and investigations into its function as a complement or substitute would need to be done.

In 2006, Utsunomiya et al. examined the potential use of automated fare collection systems via transaction data, similar to credit card data, to improve transit planning. This study noted that the current market penetration was insufficient to provide a representative sample of the population with a bias toward higher income and younger individuals being noted. Additionally, privacy concerns as well as errors with

transactions and/or routing assignments were acknowledged with this method. Similarly, the 2012 study by Macfarlane and Garrow, considered the use of targeted marketing records that are sold by credit reporting agencies as a data source to estimate regional travel demand. The study showed that use of the targeted marketing data in conjunction with vehicle registration data was successful when considering age, noting that additional variables should be considered in the future.

In 2010, INRIX issued a press release concerning their partnering with the Ford Motor Company and MapQuest who would be implementing their SpeedWaves™ technology. This technology falls under the navigation umbrella and would provide increased accuracy of more than 70% for real-time travel time information through the partners systems for over 260,000 miles of roadways. This data capability could provide a potentially useful platform for the creation of O-D matrices. However, this method is limited by sampling biases.

The 2012 Tornero, Martinez, and Castello study explored the potentials of the V2I communication technologies as a new data source to be used in the creation of O-D matrices. The study states that the use of dedicated short-range communication connecting vehicles via on-board units to infrastructure via roadside units would have the potential to collect data on every vehicle connected to the system creating an accurate, instantaneous, and dynamic O-D matrix in real-time. This would effectively eliminate the need for an estimation O-D matrix method. The study did note that there were potential privacy concerns with this method. Moreover, V2I has not been deployed, with the exception of testbeds, and is not currently a viable option for data collection.

2.2.5: Location-Based Social Networking

Within the study of sociology, a social network is defined as a social structure comprised of a set individuals or organizations and the interaction ties between these individuals or organizations (Wasserman, 1994). Social networking sites, traditionally web-based, build up these actor/interaction ties by building networks or relations among individuals with similar interest, activities, backgrounds, and other types of connections. According to eBizMBA.com, popular social networking sites include Facebook©, Twitter©, and LinkedIn®. Facebook©, founded in 2004, is currently ranked as the number one site for global web traffic and second for the United States according to Alexa, a website analytics company, and has over one billion active users (Key Facts, 2013).

Location-based services (LBS) are location and time data used as control features within computer programs. LBS can be used in everything from a cell phone to control systems to smart weapons (Wikipedia, 2013). Location-based social networking (LBSN) is the combination of social networking and LBS. According to Zheng (2012),

> A location-based social network (LBSN) does not only mean adding a location to an existing social network so that people in the social structure can share location-embedded information, but also consists of the new social structure made up of individuals connected by the interdependency derived from their locations in the physical world as well as their location-tagged media content, such as photos, video, and texts. Here, the physical location consists of the instant location of an individual at a given timestamp and the location history that an individual has accumulated in a certain period. Further, the interdependency includes not only that two persons co-occur in the same physical location or share similar location histories but also the knowledge, e.g., common interests, behavior, and activities, inferred from an individual's location (history) and location-tagged data.

LBSN services include geo-tagged-media-based, point-location-driven, and trajectory-centric. Geo-tagged-media-based services are media focused and include

Twitter. Point-location-driven services are focused on point location providing instant real-time information and include Foursquare. While trajectory-centric services are focused on trajectories providing rich data and include GeoLife.

Recently, researchers have begun to data mine LBSN sites to comprehend the spatial patterns of users. The 2009 study by Li and Chen was the first large-scale quantitative analysis of a real-world commercial LBSN. The study examined Brightkite, which allows users to share their location, post notes, and upload photos, and used the Markov-based location predictor to determine future locations of users resulting in a median accuracy of 49%. Following this initial effort, Cho, Myers, and Leskovec (2011) investigated the relationship between geographic human movement, the temporal dynamics of human movement, and the social network ties using Brightkite and Gowalla, which allows users to check-in to their current locations. This study showed that while short-ranged travel was periodic in nature both spatially and temporally, long-distance travel was influenced by social networking ties. Additionally, it was shown that social relationship could explain approximately 10 to 30% of human movements and a model was proposed to predict the locations and dynamics of future human movements. Other research includes Zheng, Xie, and Ma (2010) and Karimi (2010), which explored GeoLife and Genetic Location-Based Social Networks (G-LBSM), respectively.

Backstrom, Sun, and Marlow (2010) determined that a network of a user's Facebook friends could be used to predict the user's location within 25 miles for 69.1% of the users with 16 or more friends. Concurrently, Cheng, Caverlee, and Lee (2010) used Twitter to estimate a user's city-level location using only the content of the user's tweets. The study was able to place 51% of the users within 100 miles of their actual location.

2.3.4.1: Foursquare Data

Foursquare is an application (or app), created by Dennis Crowley and Naveen Selvadurai in 2008 and launched in Austin, TX at the annual South by Southwest (SXSW) Interactive event in March 2009, allowing individuals to share and save the places that they visit with friends via check-ins. As of January 2013, Foursquare had over 30 million users worldwide with over three billion check-ins. Users of the app include individuals as well as businesses. Check-ins are encouraged by the app with "badges" begin able to be earned through frequent checking-in at locations. Through the app, businesses can engage with customers promoting news, events, and discounts.

Research has been conducted using Foursquare recently, demonstrating its viability as a data source. These efforts include the 2011 study by Cheng et al. that used check-ins from various LBSN sites, the majority coming from Foursquare, to examine human mobility patterns quantitatively across spatial, temporal, social, and textual aspects. Contemporarily, Scellato et al. (2011) conducted the first large-scale effort to unravel the socio-spatial properties of LBSN users among three main LBSN sites (Brightkite, Foursquare, and Gowalla). The study found that the socio-spatial structure could not be explained solely from the geographic factors or social mechanisms, but proposed gravity models as a more accurate way of modeling the networks. Yang et al. (Yang, 2014) used Foursquare LBSN data to estimate an O-D matrix for non-commuting trips in the Chicago urban area resulting in the promising potential for the methodology. Jin et al. (2013) examined using check-in data from Foursquare to analyze the O-D demand for Austin, TX using a singly-constrained gravity model with a two regime friction factor. This study was able to illustrate the potential of using LBSN data for travel demand analysis and monitoring.

Chapter 3: Methodology

3.1: EXPERIMENTAL DESIGN

3.1.1: Study Area

According to the United States Census Bureau, Austin, TX had an estimated population of 842,592 in 2012 and is the fourth largest city in Texas and the 11th largest city in the US. Austin serves as the capital city for the state of Texas, and is home to the University of Texas at Austin. Additionally, many Fortune 500 companies are headquartered or have an office within the Austin metropolitan region including Dell, Whole Foods Market, and Advanced Micro Devices Inc. (CNN Money, 2013). Simultaneously, Austin is known as "The Live Music Capital of the World" and as such hosts more than 250 music venues and festivals, including SXSW, and Austin City Limits. These music festivals and other social events, comprising of the Circuit of the Americas Formula 1™ and the Austin Film Festival, bring to Austin over 19 million visitors annually (Austin, 2013).

The city of Austin has an area of 272 square miles (U.S. Census Bureau, 2013). For the purpose of this thesis, the study area included the 520 TAZs within the City of Austin's jurisdictions, Figure 2. These boundaries are the same that were used in the Jin et al. study (2013).

Figure 2: Map of Study Area. (Bing, 2012)

3.1.2: CAMPO

In 2010, Austin's metropolitan planning organization, CAMPO, published their Urban Transportation Study. This document provided the 2005 base year Travel Demand Model (TDM) calibration summary for the updating of the 2035 Long Range Plan which was recalibrated and validated in 2009. The study area for this report includes Williamson, Travis, Hays, Bastrop, and Caldwell counties. The internal TAZs used in the study totaled 1,413. In addition to these internal TAZ, 49 external stations were included within the study for a grand total of 1462 zones.

3.1.2.1: CAMPO Origin-Destination Data

The 2005 Travel Demand Model (TDM) was comprised of seventeen trip purposes made up of eleven "person" trips (1-11) and six "vehicle" trips (12-17). These trips were defined as:

1. Home Based Work Person Trips Direct
2. Home Based Work Person Trips Strategic
3. Home Based Work Person Trips Complex
4. Home Based Non-work Retail Person Trips
5. Home Based Non-work Other Person Trips
6. Home Based Non-work Primary Education Person Trips
7. Home Based Non-work University/College Person Trips
8. Home Based Non-work UT-Austin Education Person Trips
9. Home Based Non-Work/Non-home Based (non-work) Airport Person Trips
10. Non-home Based Work-related Person Trips
11. Non-home Based Other Person Trips
12. Non-home Based External Commuter/Visitor Vehicle Trips
13. Commercial Truck/Taxi Vehicle Trips
14. External Local Auto Vehicle Trips
15. External Local Truck Vehicle Trips
16. External Through Auto Vehicle Trips
17. External Through Truck Vehicle Trips

For their study, CAMPO split the Home Based Work trips into three categories. The Direct Home Based trip is defined as a trip that is made up of both direct home-to-work and direct work-to-home trips. The Strategic Home Based trip is defined as a trip that

includes an intermediate destination involving the dropping off or picking up of a child at a childcare facility, which includes daycare, nursery school, babysitter, pre-school, elementary or secondary school. The Complex trip is defined as a trip that involves an intermediate stop at any destination between the home and work places.

To collect the necessary data for the TDM, CAMPO performed travel surveys. These consisted of household surveys that included 1,500 household samples, workplace surveys made up of 210 business samples, a special generator survey-component of workplace survey, external survey that included vehicle classification sites, a commercial vehicle survey, an on-board transit ridership survey via Capital Metro, and a roadway congestion analysis report. Table 1 provides a description of the 2005 demographic data for the study area.

Population	1,458,641
Households	559,423
Household Size	2.61
Median Household Income	$53,627
Automobiles Owned	859,021
Automobiles per Household	1.50
Automobiles per Person	0.60
Total Employment	713,136
Basic Employment	181,361
Retail Employment	133,952
Service Employment	299,936
Educational Employment	43,279
Airport Employment	17,804
Population/Employment Ratio	2.05

Table 1: CAMPO 2005 Demographic Data.

The CAMPO model structure will be used to create the base or ground truth for the proposed model from this thesis. The first 13 categories from the listed 17

categories will be combined to attain the trips taken with the Home Based Work trips combined into one category for this study.

3.1.3: Data Collection

Due to the popularity, comprehensive functionality, and easy of data attainment, Foursquare was selected as the main data source for this thesis. Figure 3 shows the comparative demographic characteristics of Foursquare users (Ignite Social Media, 2013), Austin, TX, (USCB, 2013, CLRSearch, 2012) and the US (Howden, 2011). 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 higher 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.

Figure 3: Comparative Demographic Characteristics

Jin et al. (2013) provided the initial framework for the data collection, which will be documented in detail within this section. Data was collected from Tuesday, June 11 through Tuesday, July 2, 2012 for the calibration and analysis of the model.

3.1.2.1: Trolling Algorithm

The developers of Foursquare provide free access to their data through an API (Application Programming Interface) via OAuth2 (Open Authorization 2.0). Through

the API, venues endpoints, which include homes, business, parks, and other sites, are able to be accessed without user authentication. The real-time API provides real-time views into check-ins. Foursquare has a venue push API that pushes, or sends, data each time a user checks into a venue; however, a venue manager must authorize the push request for each location. This limitation makes this feature impractical for data collection. Instead, venues were identified within the study area and a computer program was developed that would create a snapshot of the total number of check-ins for each of the venues. The computer programs snapshots are taken at 45 to 50 minute intervals and an hourly rate is calculated using the following formula:

$$C_{hr} = \left(\frac{x_2 - x_1}{t_2 - t_1}\right) * 60 \quad (eqn. 18)$$

where C_{hr} is the check-ins per hour, x_i is the number of check-ins at the two time intervals, and t_i is the time in minutes for each of the intervals.

Using this trolling method, the data collected included: venue ID, venue name, category, latitude, longitude, the number of check-in per hour, and the number of unique users. Data collection was done for each hour during the study period using the trolling method.

3.1.2.2: Preliminary Analysis of Collected Data

For the data collected June 11 through July 2 an initial analysis was conducted to verify the type of data that was collected. For use within the doubly-constrained gravity model, the data was examined to confirm that categories were assigned via Foursquare to each of the check-ins. When venues are created within Foursquare, a category can be assigned as well as subcategories. Categories that Foursquare uses include: Arts & entertainment, College & University, Food, Professional & Other Places, Nightlife Spots,

Residences, Great Outdoors, Shops & Services, and Travel & Transport. Foursquare does provided secondary and tertiary categories, but these were not included in the study as the primary category classification provides the information needed. It should be noted that categories are assigned by venue creator and is optional, for some of the venues included within the study, no category was assigned. For these venues, a key word search was performed to assign the appropriate primary category where possible.

Table 2 shows a breakdown of the number of venues and check-ins collected during the study period for each category. The majority of the venues within the study area come from the Shops & Services category, while the least number of venues can be found in the Nightlife Spots category. Check-ins are most commonly 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.

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 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 2: Foursquare Category Venue and Check-in Statistics.

The average number check-ins was also calculated for each venue 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 initially unexpected, which may be explained by the frequent check ins at these locations by visitors. Additionally, due to the low percentage (1.5%) of check-ins for the Unclassified category, it was determined to be a small enough to be removed from the study without impacting the travel demand analysis results negatively.

Using the latitude and longitude data, venues were assigned to their respective TAZ. The 19,170 venues are shown in Figure 4 individually and in Figure 5 by density. These figures demonstrate the spatial coverage within Austin for the LBSN data. Moreover, the figure shows the venues are concentrated within the central business district (CBD) located near the center of the map. Additionally, the majority of the venues are located within the densely populated areas of the study area, as is to be expected. Furthermore, venues with check-ins exist in almost every TAZ within the study area, only the three highlighted TAZ shown in Figure 4 do not have any venues, demonstrating the spatial coverage available by this method.

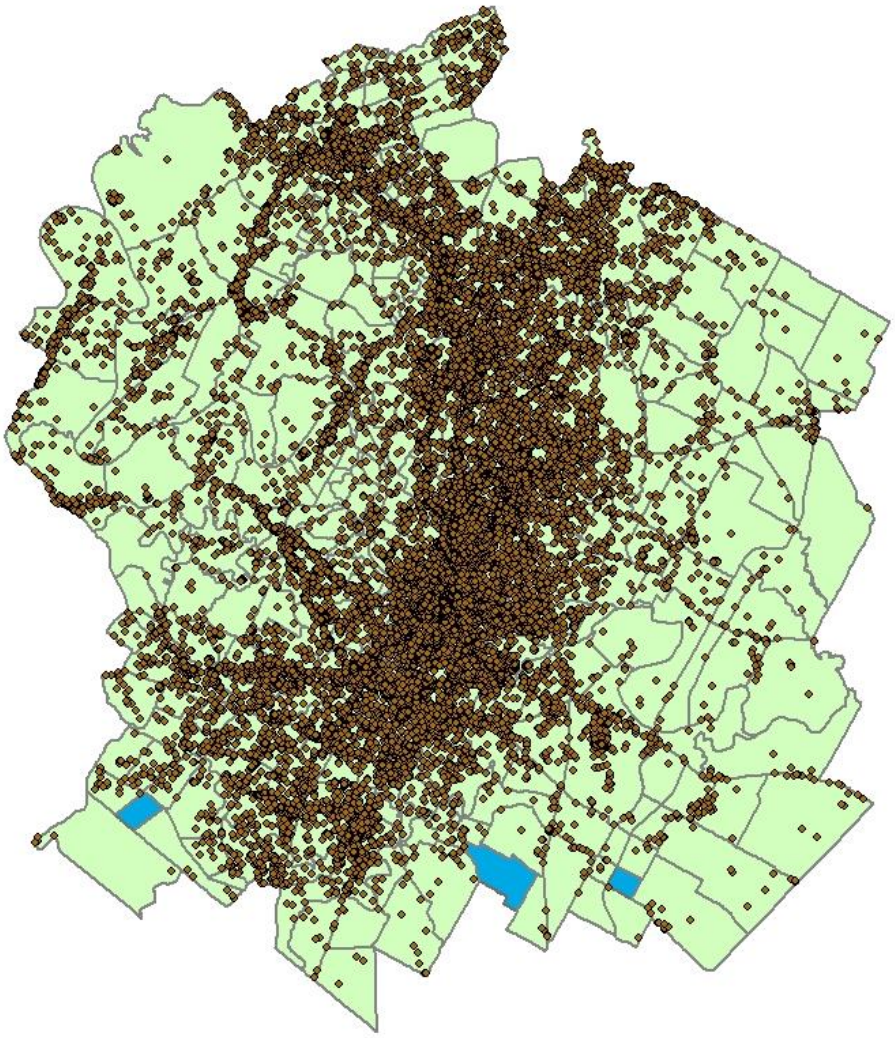


Figure 4: Venue Locations

Figure 5: Venue Locations by Density

To better explore the temporal characteristics of the check-in data, check-ins in the various categories were aggregated for weekdays and weekends. Figure 6 shows the most frequent check-ins during weekdays were in the categories of Food, Shops & Services, and Professional & Other Places. For the weekends, the most frequent check-ins occurred in the Food, Shops & Services, and Nightlife Spots. Additionally, the data shows weekday check-ins are significantly more predominant when compared to weekend check-ins for Colleges & Universities, Professional & Other Places and Residence. In should also be noted that Shops & Services and Nightlife Spots have significantly more frequent check-ins on the weekends than on the weekdays.

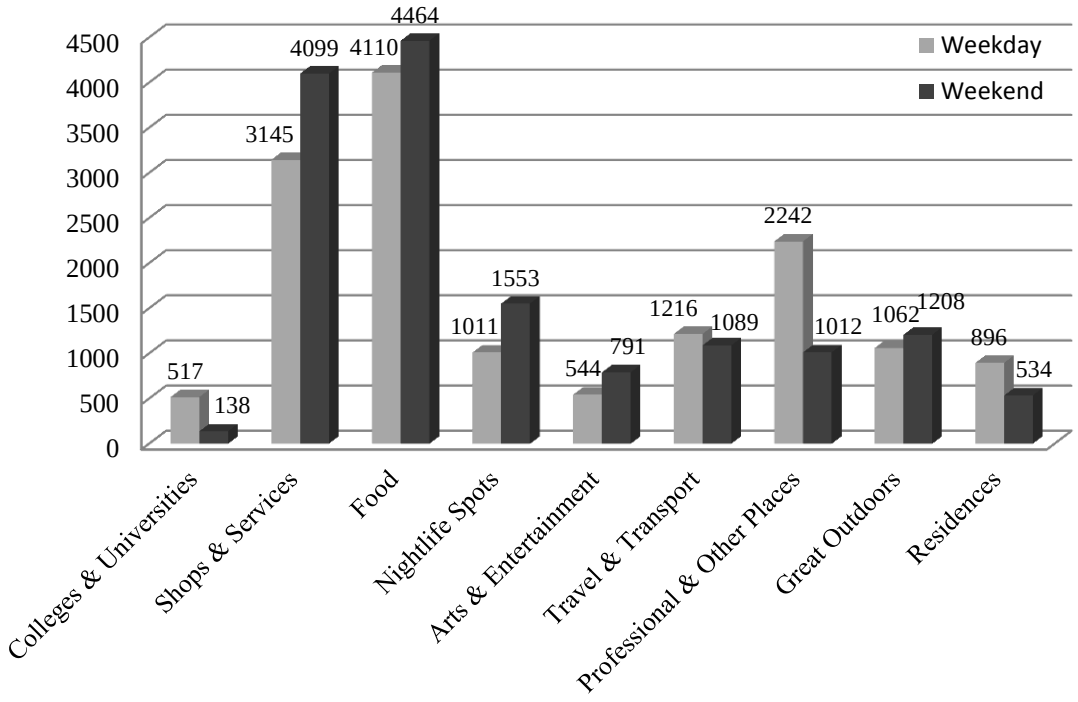


Figure 6: Weekday and Weekend Check-in Comparison by Category

3.1.4: Origin-Destination Modeling

Jin et al. (2013) examined the use of Foursquare check-in data for the Austin area for use in the creation of an O-D matrix and used the singly-constrained gravity model. This thesis looks to continue the effort and further validate the use of check-in data by utilizing the doubly-constrained gravity model. MATLAB software was used to create and evaluate the model. The MATLAB codes used in this effort have been included in the Appendices of this thesis.

3.1.4.1: Trip Generation Modeling

As previously mentioned, the CAMPO O-D data from the CAMPO Urban Transportation Study (2010) was used as the ground truth. This data, first, had to be

manipulated to only include the 520 City of Austin TAZs that are within the boundaries of this study. To accomplish this effort, a text file was created that contained the numbers of the TAZs to be included called tazid.txt. Using the tazid.txt file and the 2010 Person Trip Table.txt, MATLAB assigned the trip table data to each of the 13 categories as previously defined, see Appendix A for the code. The focus of this study eliminated the Home-based Non-work Primary Education and University/College categories in an effort to focus exclusively on the University of Texas at Austin trips. The result of this effort was the assignment of the CAMPO trip data into 520x520 matrices with the following 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 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)

For the weekday Foursquare check-in data, a 10x1462 matrix was created with the weekday Foursquare check-in data. This file, called 'weekday', was used to create the 'tripdist', 'totalOD', and 'checkins' matrices via MATLAB, see Appendix B for code. The 'tripdist' matrix used the Manhattan distance between the centroids of origin zone *i* and destination zone *j* to create a 520x520 matrix. The 'totalOD' matrix combined the CAMPO trips per category into a single 520x520 matrix. This matrix will serve as the ground truth matrix for future comparisons. The 'checkins' matrix restructured the

'weekday' matrix removing the data associated with TAZs not included within this study creating a 10x520 matrix.

The next step was to obtain the production and attraction rates from the check-in data. The 'checkins' matrix was employed and the first nine of the ten rows were assigned to the following categories: Professional, Shops, Universities, Residence, Travel spots, Entertainment, Food, Nightlife, and Outdoor. The tenth row was the uncategorized data and was eliminated from the study as described previously. Productions and attractions for each zone were calculated by the following formulas for the singly- and doubly-constrained models:

Singly-Constrained -

$$O_i = \gamma * x_i \quad (eqn. 19)$$

$$D_j = (\varepsilon * x_i) + \left[\frac{1}{N}\sum_i (\gamma - \varepsilon)x_i\right] \quad (eqn. 20)$$

Doubly-Constrained -

$$O_i = \gamma * x_i \quad (eqn. 21)$$

$$D_j = (\varepsilon * x_i) + \frac{{x_i}^{\eta}}{\sum_i {x_i}^{\eta}}\sum_i (\gamma - \varepsilon)x_i \quad (eqn. 22)$$

where x_i is the total check ins in zone i found by $\sum_i (Professional + Shops + Universities + Residence + Travelspots + Entertainment + Food + Nightlife + Outdoor)$, γ and ε are adjustment factors created from a genetic optimization algorithm in an effort to appropriately scale the trip productions to Foursquare check-ins, $\frac{1}{N}\sum_i (\gamma - \varepsilon)x_i$ and $\frac{{x_i}^{\eta}}{\sum_i {x_i}^{\eta}}$ are the residual terms that guarantees the total number of

productions are equal to the total number of attractions. The genetic optimization algorithm and its results will be discussed in more details in the next section.

3.1.4.2: Trip Distribution Modeling

After attaining the productions and attractions, the process to recombine the productions and attractions into trips via a trip distribution model is undertaken. The previously studied singly-constrained gravity model used in the Jin et al. study demonstrated the effectiveness of the methodology. However, there were limitations in this previous effort. The zonal production and attraction model typically resulted in symmetric patterns from the extremely small residual term which may be a factor of taking an average values and not zone-specific values. Additionally, the singly-constrained model only adjusts the zonal attractions and converges comparatively slowly. For this thesis, the doubly-constrained gravity model will be employed for further validation of the methodology and will be compared to the singly-constrained model's results.

3.1.4.2: Gravity Models

The TDM used for the CAMPO study is the ATOM2, an atomistic model which is a triply constrained model. While the singly-constrained gravity model used in the Jin et al. study showed the check-in method for O-D matrix creation was functional, the doubly-constrained gravity model is a near approximation to the modeling effort used in the CAMPO study for the estimation of the O-D matrix. The doubly-constrained gravity model constrains both the productions and attractions.

As discussed in Chapter 2, the singly- and doubly-constrained model uses the following formulas:

Singly-Constrained -

$$T_{ij} = O_i * \frac{D_j * f(c_{ij})}{\sum_j D_j * f(c_{ij})} \quad (eqn. 23)$$

Doubly-Constrained -

$$T_{ij} = \beta_i * O_i * \alpha_j * D_j * f(c_{ij}) \quad (eqn. 24)$$

To calculate the $f(c_{ij})$, the two-regime function found by Jin et al. was used. This function accounts for the different trends for short and long distance trips and is structured as such:

$$F_{ij}(d_{ij}) = F_{ij}{}^{(s)}(d_{ij}) I_{d_{ij} \leq T_d} + F_{ij}{}^{(l)}(d_{ij}) I_{d_{ij} > T_d} \quad (eqn. 25)$$

where the $I_{[clause]}$ is an indicator function for a logic clause that equals one if true and zero otherwise, *s* and *l* are the short and long distance trip regime, and T_d is the threshold that determines the regime. Jin et al. explored various combinations of the linear, negative exponential, and gamma friction factor functions and found the linear model achieved the best results for short distance trips, while the negative exponential model was best for long distance trips:

$$F_{ij}(d_{ij}) = \begin{cases} \theta + \lambda * d_{ij}, & d_{ij} < \sigma \\ \mu * e^{-\rho * d_{ij}} d_{ij} \geq \sigma \end{cases} \quad (eqn. 26)$$

To calculate the α_j and β_i factors that constrain the productions and attractions, at MATLAB was used, see Appendix E for code.

3.2: 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 (MATLAB, 2013). 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. Meanwhile, between two generations, each "individual" is allowed to "mutate" at a given probability which ensures the ability of jumping out of local optimal. This process is repeated and the population eventually "evolves" toward an optimal solution. Table 3 compares the genetic algorithm to a classical algorithm highlighting two main differences between the algorithms.

Genetic	Classic
Generation of a population of points for each iteration with the best point within the population approaching an optimal solution.	Generation of a single point for each iteration with the sequence of points approaching an optimal solution
Next population is selected by computation using random number generators.	Next point in sequence is selected via a deterministic computation.

Table 3: Algorithm Comparison (MATLAB, 2013).

The genetic algorithm was used to obtain parameters for the friction function, the productions, and the attraction calculations that would in turn minimize the mean absolute error (MAE) between the modeled O-D matrix and the ground truth CAMPO O-D matrix. To evaluate the performance of these parameters, a coincidence ratio (CR) was used. The CR measures the percent of the area that "coincides" for the two curves/distributions that are being compared (Martin, 1998). The CR uses the following formula:

$$CR = \frac{\sum_i min\left(p_i^M, p_i^O\right)}{\sum_i max\left(p_i^M, p_i^O\right)} \quad (eqn. 27)$$

where p_i^M is the percentage of trips within the interval *i* in the predicted trips from the check-in data, and p_i^O is the percentage of trips within the interval *i* in the survey trips from CAMPO. The value for the CR ranges from zero, when the distributions are completely different, and one, when the distributions are exactly the same. The goal for the model validation is to have a CR that is close to one.

Using the genetic algorithm, MATLAB produced nine parameters that are used within the singly- and doubly-constrained gravity model; see Appendices D, E, F, and G for code. The attraction formulas accounts for the adjusted symmetry of the distribution of check-ins by adding the exponential portion of the equation, which includes a location-based ratio for the doubly-constrained model. This location based ratio assigns the residuals based on the check in intensity. The genetic algorithm parameters for the equations for the attractions, production, and friction factors can be found in Table 4. Additionally, the genetic algorithm calculated the CR and MAE for singly- and doubly-constrained models. It should be noted that under the standard evaluation framework for machine learning algorithms, the calibration dataset and the evaluation dataset should be separated which mimics the use of those learning algorithms in practice. However, in our study, only the 2010 CAMPO O-D matrix is available and dividing it to fit the calibration-evaluation framework is difficult. Therefore, calibration and evaluation is based on the same matrix. The actual model performance may vary with separated calibration and evaluation datasets.

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
d_{ij}	11.18205	5.22909
CR	0.7456	0.9523
MAE	15.9348	10.2134

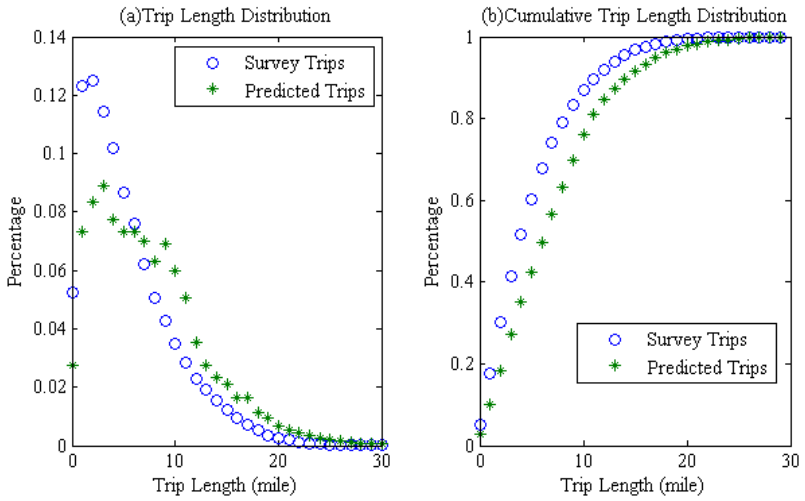
Table 4: Genetic Optimization Parameters

Chapter 4: Results and Analysis

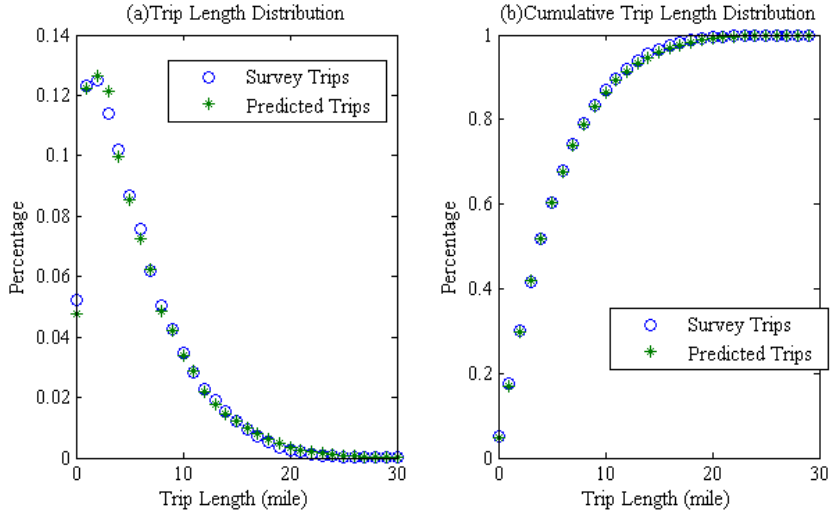
4.1: O-D Matrix Comparison

Using the genetic algorithm factors discussed in Chapter 3, a Foursquare check-in based O-D matrices were created for the singly- and doubly-constrained gravity models. The matrices were compared to the CAMPO O-D matrix to determine the validity of the proposed doubly-constrained model. This comparison was done with three different methods trip length distribution comparison, zonal trip generation and attraction heat maps, zonal O-D flow patterns.

The first effort compares the trip length distributions for the Foursquare check-in gravity model matrices with the CAMPO matrix. Figure 7 shows the trip length distributions (a) and the cumulative trip length distributions (b) for the singly-constrained (7a) and doubly-constrained (7b) models. Examination of the Trip Length Distribution portion of figure 7b, shows the doubly-constrained model is relatively constant with respect to the general curvature. However for Figure 7a, under estimation is occurs for short trips and slight over estimation occurs for long trips. For the cumulative distribution figure, slight under estimation is consistently shown for the doubly-constrained 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

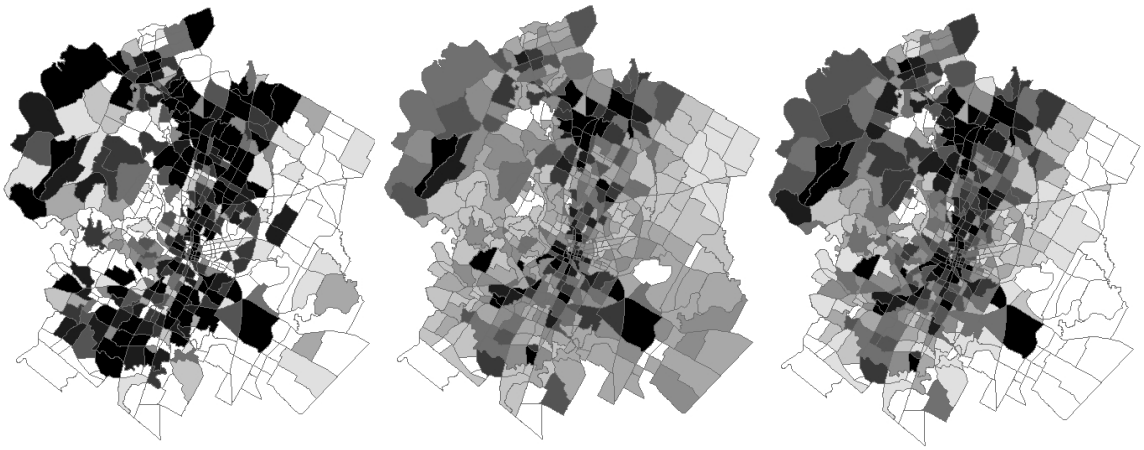


(b) Doubly-constrained Model Trip Length Frequency Results

Figure 7: Trip Length and Cumulative Trip Length Distributions

The next comparison examined was the zone based production and attraction heat maps. These maps use a gradient based color scheme to differentiate the number of productions and attractions generated for each zone for each of the models. In an effort to keep the comparisons consistent, all zonal maps use the same scaling factor for the

gradients. Figure 8 compares productions generated from the CAMPO model, and from the Foursquare for the singly- and doubly-constrained gravity models demonstrating where the methodology excels and where there are limitations. Using the CAMPO production map as the ground truth model, 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 with similar magnitude to the CAMPO map through the study region. 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



CAMPO Singly-Constrained Doubly-Constrained

Figure 7: Production Comparison Maps

Figure 8 compares the attractions of the two models with respect to the CAMPO OD matrix and highlights where the methodology excels and where there are limitations. Once again using the CAMPO production map as the ground truth model, the singly-

constrained model predicts attraction rates similar to the CAMPO model for many areas, but suffers again 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.

CAMPO Singly-Constrained Doubly-Constrained

Figure 8: Attraction Comparison Maps

The final comparison effort between the two models examines the flow patterns between the origins and destinations for both the CAMPO ground truth O-D matrix and the singly- and doubly-constrained gravity model O-D matrices. The figures below show the zonal comparison of the O-D patterns for the CAMPO matrix, the singly- (Figure 8) and doubly-constrained (Figure 9) Foursquare matrices using the Log10 intensity, which is calculated using the following formula.

$$I_{ij} = log_{10}\left(\frac{T_{ij}}{\sum_i \sum_j T_{ij}}\right) \quad (eqn. 28)$$

Using the intensity formula, the graphic uses the horizontal axes as the origins and the vertical axes as the production zones in ascending order. Additionally, the O-D MAE matrices for the singly- and doubly-constrained models are included.

Figure 9 demonstrates the similarities between the singly-constrained model and the CAMPO model. The darker the color within the figure, the higher the O-D flow. While the areas of lighter flow, shown with the lighter coloring, are reasonably consistent in the Foursquare model, the areas with higher flow are not as prevalent. This is consistent with the less variegated productions and attractions shown within the singly-constrained model as shown above. Additionally, the MAE matrix is provided to demonstrate how closely the estimate Foursquare matrix matches the CAMPO matrix.

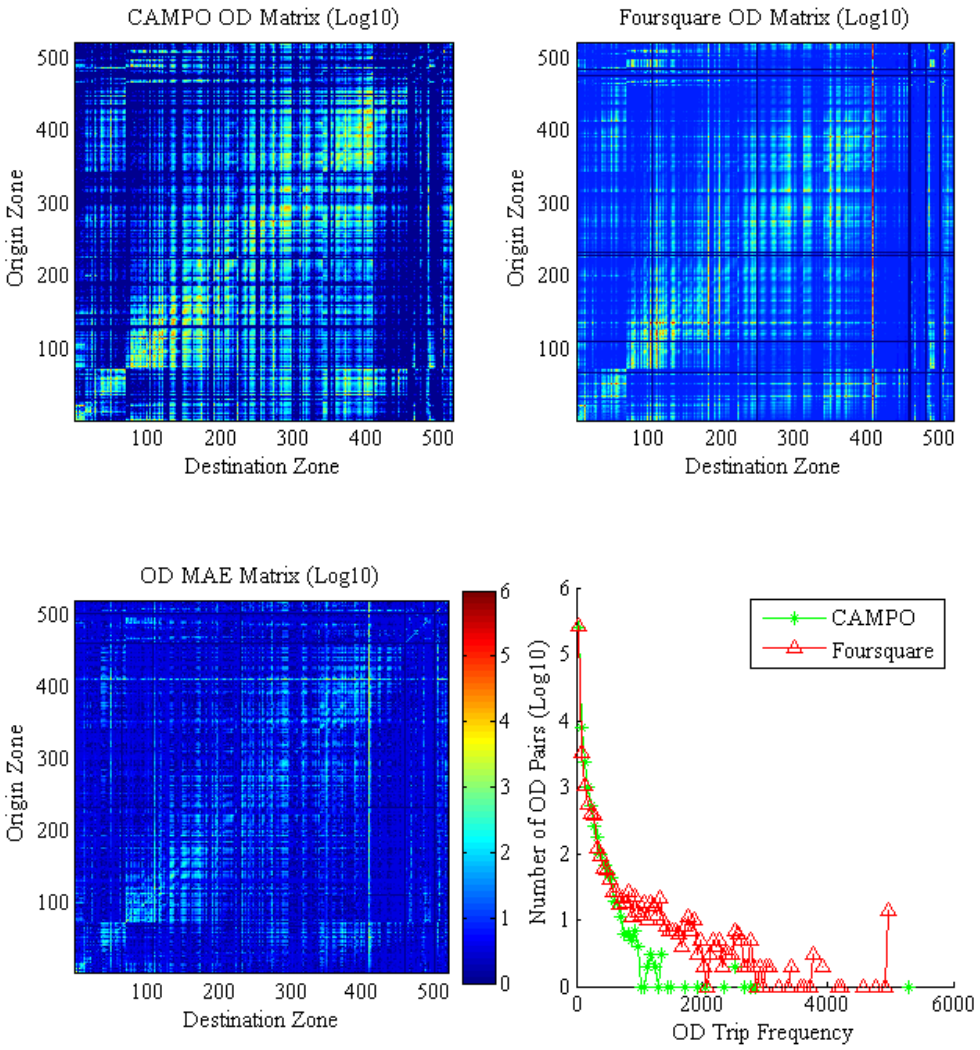


Figure 9: Zonal Comparison of Singly-Constrained Gravity Model O-D

Figure 9 compares the O-D flow pattern between the CAMPO O-D matrix and the doubly-constrained gravity model matrix. Comparing the CAMPO and Foursquare 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 less flow along the inter-zonal 45º line when compared to both the CAMPO ground truth and the singly-constrained model. Additionally, the doubly-constrained

model has a more variegated color pattern through the diagram, which more accurately reflects the ground truth model. This coincides with the CR for the doubly-constrained model being closer to one than the singly-constrained model, which were 0.9523 and 0.7456, respectively. Similar to the singly-constrained model, the MAE matrix is provided to demonstrate how closely the estimate Foursquare doubly-constrained matrix matches the CAMPO matrix.

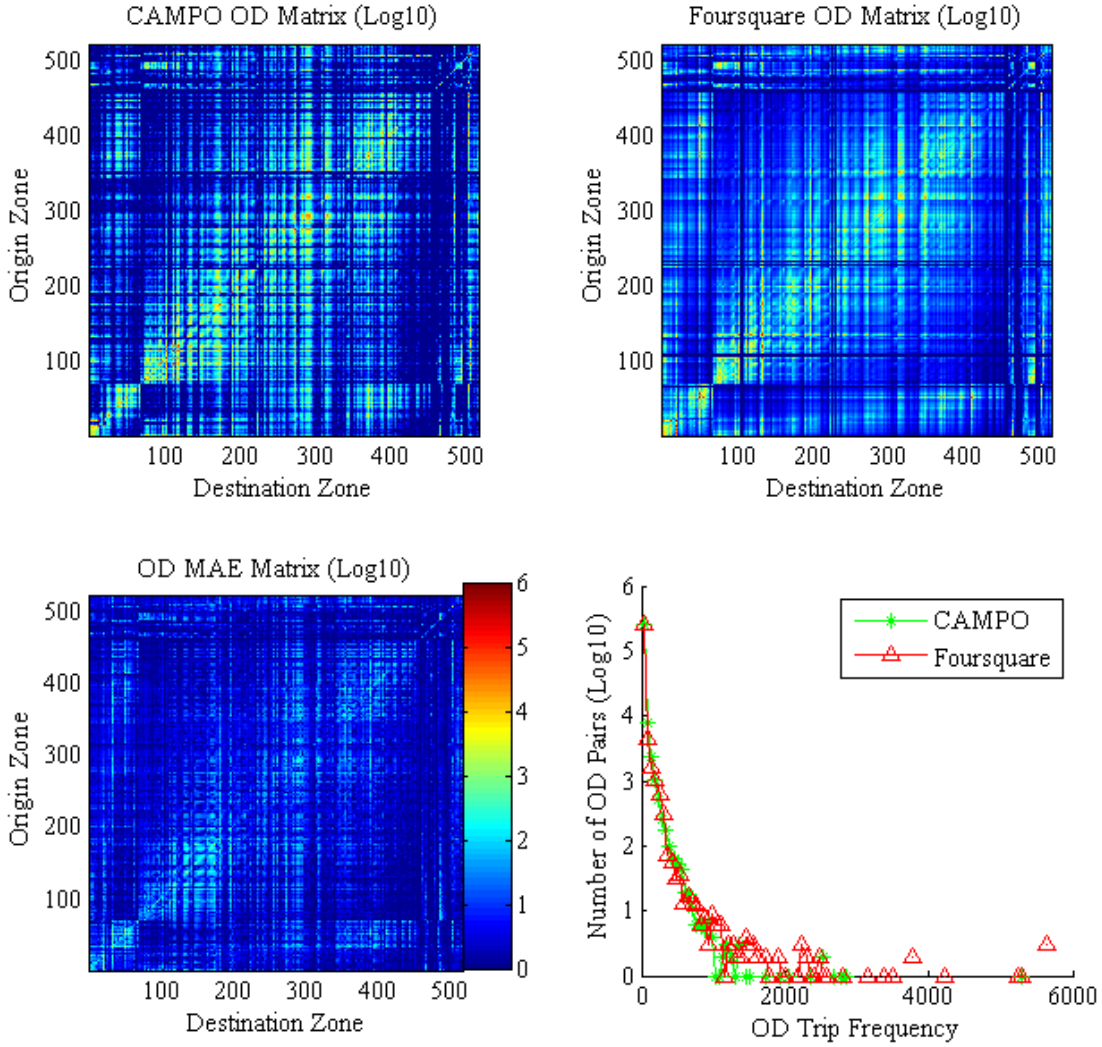


Figure 10: Zonal Comparison of Doubly-Constrained Gravity Model O-D

Chapter 5: Conclusion

5.1: CONCLUSION

Trip distribution is a significant portion of the transportation planning process. Traditionally, the data used for the creation of O-D matrices comes from the household surveys. This thesis examines the effectiveness of using LBSN data in the form of check-ins from the Foursquare application to calculate an O-D matrix for the Austin area. Previous studies by Yang et al. (2014) and Jin et al. (2013) explored the use of check-in data to create an O-D matrix using singly-constrained gravity model. These previous efforts demonstrated the validity of the method. As an expansion, this thesis explores the use of a doubly-constrained gravity model for the creation of an O-D matrix. Check-in data from Foursquare, a leading LBSN provider, was used to create production and attraction rates for the doubly-constrained model as well as the singly-constrained gravity model which was used in conjunction with the CAMPO ground truth matrix to examine the predictability of the proposed methodology. Additionally, location-based ratios were used to calculate the productions and attractions for the singly- and doubly-constrained models and a genetic optimization algorithm was employed to attempt to reach a global optimization for the models. Furthermore, a combination friction factor that was comprised of a linear friction function for short distances (defined as less than 11.18205 for the singly-constrained and 5.22909 for the doubly-constrained models for this effort) and a negative exponential function for long distances.

In comparison to the traditional methods used for O-D estimation, this study shows that LBSN data has the potential to provide better spatial coverage, and benefits from build-in user verification, real-time updating potentials, and a significantly lower data collection cost compared to other methodologies. To further evaluate the model, a

comparison of trip length distributions, zonal trip generation and attraction heat maps, zonal O-D flow patterns were conducted. This effort related the Foursquare singly- and doubly-constrained models with the ground truth CAMPO model. Comparing the CAMPO ground truth O-D matrix with the Foursquare singly- and doubly-constrained matrix, a coincidence ratios of 0.7456 and 0.9523, respectively, were established. Additionally, mean absolute errors of 15.9348 and 10.2134 were calculated. Moreover, in comparison to the singly-constrained gravity model, the doubly-constrained model demonstrates better learning capabilities.

There are some limitations with the proposed methodology that should be examined in future research. Currently the matrices are still quite symmetric (although not exactly symmetric) and as shown in the intensity plots, there are deficiencies in the one to one comparisons between the Foursquare models and the CAMPO model. Further examination into the temporal aspects of the models as well as specific trip purposes should be researched to further validate this proposed methodology.

Despite these initial short comings of the model, these initial efforts demonstrate that the use of LBSN check-in data is a viable method for creating an O-D matrix. Future research should examine different functions for associating the check-in data to productions and attractions with the intent of distributing the check-ins in a manner that is proportional to density of venues within the zone. Additionally, the use of the triply-constrained atomistic model with the check-in data may produce O-D matrix that more nearly resembles the CAMPO matrix.

Appendix A

CAMPO Data Processing (campoProcess.m)

```
functioncampoProcess
closeall
data = csvread('2010 Person Trip Table.txt');
totalZone = 1462;

zoneIdx = csvread('tazid.txt');

% 1. Home Based Work Person Trips Direct (HBW-Direct)
HBWD = reshape(data(:,1+2),totalZone,totalZone)';
% 2. Home Based Work Person Trips Strategic (HBW-Strategic)
HBWS = reshape(data(:,2+2),totalZone,totalZone)';
% 3. Home Based Work Person Trips Complex (HBW-Complex)
HBWC = reshape(data(:,3+2),totalZone,totalZone)';
HBW = HBWD+HBWS+HBWC;

% 4. Home Based Non-work Retail Person Trips (HBNW-R)
HBR = reshape(data(:,4+2),totalZone,totalZone)';
% 5. Home Based Non-work Other Person Trips (HBNW-O)
HBO = reshape(data(:,5+2),totalZone,totalZone)';
% 6. Home Based Non-work Primary Education Person Trips (HBNW-E1)
HBEdu = reshape(data(:,6+2),totalZone,totalZone)';
% 7. Home Based Non-work University/College Person Trips (HBNW-E2)
HBUniv = reshape(data(:,7+2),totalZone,totalZone)';
% 8. Home Based Non-work UT-Austin Education Person Trips (HBNW- UT)
HBUT = reshape(data(:,8+2),totalZone,totalZone)';
% 9. HBNW/NHB (Non-work) Airport Person Trips (NW-Airport)
NWAir = reshape(data(:,9+2),totalZone,totalZone)';
% 10. Non-home Based Work-related Person Trips (NHB-W)
NHBW =  reshape(data(:,10+2),totalZone,totalZone)';
% 11. Non-home Based Other Person Trips (NHB-O)
NHBO =  reshape(data(:,11+2),totalZone,totalZone)';
% 12. Non-home Based External Commuter/Visitor Vehicle Trips (NHB-Exlo)
NHBE =  reshape(data(:,12+2),totalZone,totalZone)';

HBW = HBW(zoneIdx,zoneIdx);
HBR = HBR(zoneIdx,zoneIdx);
HBO = HBO(zoneIdx,zoneIdx);
HBUT = HBUT(zoneIdx,zoneIdx);
NWAir = NWAir(zoneIdx,zoneIdx);
NHBW = NHBW(zoneIdx,zoneIdx);
NHBO = NHBO(zoneIdx,zoneIdx);
NHBE = NHBE(zoneIdx,zoneIdx);

save('campo.mat','HBW','HBR','HBO','HBUT','NWAir','NHBW','NHBO','NHBE')
;
```

Appendix B

Check-in Data Processing (foursquaredata.m)

```
clearall;
loadcampo
loadcentroids
loadweekday

zoneIdx = csvread('tazid.txt');

fori=1:520
for j=1:520
    distance(i,j)=(abs(lat(i)-lat(j))+abs(lng(i)-lng(j)))*100;
end
end
tripdist=distance+5.*eye(size(distance,1));

totalOD=HBO+HBR+HBUT+HBW+NHBE+NHBO+NHBW+NWAir;

weekday = reshape(weekday(:,:),10,1462)';

checkin=weekday(zoneIdx,:);

checkins=reshape(checkin(:,:),520,10)';
```

Appendix C

Doubly-Constrained Gravity Model Optimization (doubleGravityOpt.m)

```
function doubleGravityOpt
clear all

load campo
load centroids
load weekday
% global tripdist totalOD checkins
zoneIdx = csvread('tazid.txt');
for i=1:520
    for j=1:520
        distance(i,j)=(abs(lat(i)-lat(j))+abs(lng(i)-lng(j)))*100;
    end
end
tripdist=distance+5.*eye(size(distance,1));
totalOD=HBO+HBR+HBUT+HBW+NHBE+NHBO+NHBW+NWAir;
weekday = reshape(weekday(:,:),10,1462)';
checkin=weekday(zoneIdx,:);
checkins=reshape(checkin,520,10)';

algCells = {@dgravity,@sgravity};
algNames = {'Doubly','Singly'};

%%%%%%%%%%%%kp   ka   pow alpha beta alpha1 beta1 TD TUpperBd TLowBd
adjMid
lowerDBds = [1e-3 1e-3 0.1 1e-3  1e-3 1e-3   1e-3  5  5000    300    .1
];
upperDBds = [10   10   4   5     5    10     10    15 10000    1500   .8  ];

%%%%%%%%%%%%kp   ka   bp alpha beta alpha1 beta1 TD TUpperBd TLowBd
adjMid
lowerSBds = [1e-3 1e-3 1e-3 1e-3  1e-3 1e-3   1e-3  5  5000    300   .8
];
upperSBds = [10   10   1e-3 10    10   10     10    15 5000    300   .8  ];

lowerBdsCell = {lowerDBds,lowerSBds};
upperBdsCell = {upperDBds,upperSBds};

for a=2:2
    alg = algCells{a};
    algName = algNames{a};
    lowerBds = lowerBdsCell{a};
    upperBds = upperBdsCell{a};

nVars = length(lowerBds);
n = length(totalOD);
```

```
% options = gaoptimset('PlotFcns',
@gaplotbestf,'Generations',100,'TolFun',1e-12);
CR=0;
swapRatio=0;
params=[];
while abs(CR)<.7
    [params,fav,exitflag,output] = ga(@(x)
eva(x,checkins,'CR',totalOD,tripdist,n,alg),nVars,[],[],[],[],lowerBds,
upperBds)
%     CR = -fav;
%     swapRatio = fav;
    [swapRatio,CR,FR,MAE,ME] =
eva(params,checkins,'SwapRatio',totalOD,tripdist,n,alg)
end
truTSum = sum(sum(totalOD));
predictedTrips = dgravity(params,checkins,tripdist,n,truTSum);
truT = reshape(totalOD,1,[]);
algT = reshape(predictedTrips,1,[]);

save(['res_' algName
'_allpurpose.mat'],'params','CR','FR','ME','MAE','swapRatio','totalOD',
'predictedTrips');

csvwrite(['res_' algName '_ProAttHeatMap.csv'],[zoneIdx sum(totalOD,2)
sum(predictedTrips,2) sum(totalOD)' sum(predictedTrips)']);

display('optimization done.')

fig1=figure(1)
interval=100;
totalLength=3000;
CR =
compareTripLengthDist(totalOD,predictedTrips,tripdist,interval,totalLen
gth)
saveas(fig1,'cr.fig')

end

end
```

Appendix D

Evaluation of MAE, ME, and CR (eva.m)

```
function [z,CR,FR,MAE,ME] = eva(x,checkins,obj,totalOD,tripdist,n,alg)
% clc
truTSum = sum(sum(totalOD));
predictedTrips = alg(x,checkins,tripdist,n,truTSum);
% predictedTrips = predictedTrips/algTSum*truTSum;
truT = reshape(totalOD,1,[]);
algT = reshape(predictedTrips,1,[]);

          if sum(isnan(algT) | (algT)<0)>0
             z = 999999999;
          else

switch obj
       case 'MAE'
          MAE = mean(reshape(abs(truT-algT),1,[]));
          z = MAE;
       case 'Theil'
          p=sqrt((sum((truT-algT).^2))/length(truT));

q=sqrt((sum((truT).^2))/length(truT))+sqrt((sum((algT).^2))/length(truT
));
           z = p/q;
    case 'HybridAE'
       z = 1/2*mean(reshape(abs(truT-
algT),1,[]))+1/2*max(reshape(abs(truT-algT),1,[]));
       case 'CR'
          z = -CoincidentRatio(totalOD,predictedTrips,tripdist);
  case 'SwapRatio'
          z = swapRatio(truT,algT);
    case 'FreqRatio'
          z = -
frequencyRatio(truT,algT,ceil(max(max(truT))/1000)*1000);
    case 'CRFR'
          z = -min(CoincidentRatio(totalOD,predictedTrips,tripdist),...

4/5*frequencyRatio(truT,algT,ceil(max(max(truT))/1000)*1000));
    otherwise
          MAE = mean(reshape(abs(truT-algT),1,[]));
          z = MAE;
end
          end
MAE = mean(reshape(abs(truT-algT),1,[]));
ME = mean(reshape(algT-truT,1,[]));
CR = CoincidentRatio(totalOD,predictedTrips,tripdist);
SR = swapRatio(truT,algT);
FR = frequencyRatio(truT,algT,ceil(max(max(truT))/1000)*1000);
```

```
% display(MAE)
display(['z=' num2str(z) ', fr=', num2str(FR),  ', sr=' num2str(SR) ',
params=' num2str(x(1)) ',' num2str(x(2)) ',' num2str(x(3)) ','
num2str(x(4)) ', freAdjParams=' num2str(x(end-2)) ',' num2str(x(end-1))
',' num2str(x(end)) ])
% display(['z=' num2str(z) ', mae=' num2str(MAE) ', params='
num2str(x(1)) ',' num2str(x(2)) ',' num2str(x(3)) ',' num2str(x(4))])

function fr = frequencyRatio(truOD,algOD,ub)
    bin = 0:50:ub;
    truHist = hist(reshape(truOD,1,[]),bin);
    algHist = hist(reshape(algOD,1,[]),bin);
    truPercent=truHist./sum(truHist);
    algPercent=algHist./sum(algHist);
    fr=sum(min(truPercent,algPercent))/sum(max(truPercent,algPercent));
end



function cr = CoincidentRatio(trips,predictedTrips,tripdist)
    interval=100;
    totalLength=3000;
    m=0:interval:totalLength;

    y1=zeros(length(m),1);
    y2=zeros(length(m),1);

    for k=0:interval:totalLength
      y1(k/interval+1)=y1(k/interval+1)+sum(sum(trips(tripdist>=k &
tripdist<k+interval)));

y2(k/interval+1)=y2(k/interval+1)+sum(sum(predictedTrips(tripdist>=k &
tripdist<k+interval)));
    end

    tripsPercent=y1./sum(y1);
    predictedTripsPercent=y2./sum(y2);

cr=sum(min(tripsPercent,predictedTripsPercent))/sum(max(tripsPercent,pr
edictedTripsPercent));
end

end
```

Appendix E

Doubly-constrained Gravity Model (dgravity.m)

```
function [Tcur]=dgravity(x,checkins,tripdist,n,truTSum)

alpha=x(4);
beta=x(5);
alpha1=x(6);
beta1=x(7);
TD = x(8);
TUpperBd = x(9);
TLowBd = x(10);
adjMid = x(11);

kp = x(1);
% bp = x(2);
ka = x(2);
pow = x(3);

professional=checkins(1,:);
shops=checkins(2,:);
universities=checkins(3,:);
residence=checkins(4,:);
travelspots=checkins(5,:);
entertainment=checkins(6,:);
food=checkins(7,:);
nightlife=checkins(8,:);
outdoor=checkins(9,:);

friction=(alpha+beta.*tripdist).*(tripdist<TD)+(alpha1*exp(-
beta1.*tripdist)).*(tripdist>=TD);

inputCheckins =
professional+residence+universities+entertainment+nightlife+shops+food+
travelspots+outdoor;
production=inputCheckins.*(ka+kp);

attraction=inputCheckins.*ka;
ba = inputCheckins.^pow/sum(inputCheckins.^pow)*(sum(production)-
sum(attraction));
attraction = attraction + ba;


alphaj=ones(1,n); %1*n
betai=ones(1,n); %1*n

prevAlphaj = zeros(1,n);
```

```
prevBetai = zeros(1,n);

AS=alphaj.*attraction; %1*n
PS=betai.*production; %1*n

Tcur=(AS'*PS).*friction; %

prevDif = 1;
curDif = 0;
stepCnt = 0;
while abs(prevDif-curDif)>1e-3 && stepCnt<=20
    Pi=sum(Tcur,2)'; %1*n
    Aj=sum(Tcur,1); %1*n

    prevAlphaj = alphaj;
    prevBetai = betai;

    betai=1./((alphaj.*Aj)*friction'); %summation over j
    alphaj=1./((betai.*Pi)*friction);
    AS=alphaj.*attraction; %1*n
    PS=betai.*production; %1*n
    Tcur=(AS'*PS).*friction;
    prevDif = curDif;
    curDif = max(max(abs(alphaj-prevAlphaj)),max(abs(betai-prevBetai)));
    stepCnt = stepCnt+1;
end

Tcur = Tcur/sum(sum(Tcur))*truTSum;

%%Frequency Bias Adjustment%%
%for high frequency values
orgTcur = Tcur;
highIdx = Tcur>=TLowBd;
Tcur(highIdx)= adjMid*Tcur(highIdx);
%for extreme values
Tcur(Tcur>TUpperBd)=TUpperBd;
%obtain difference and redistribute
dif = sum(sum(orgTcur(highIdx)-Tcur(highIdx)));
Tcur(~highIdx) =
Tcur(~highIdx)+dif*Tcur(~highIdx).^pow/sum(sum(Tcur(~highIdx).^pow));
```

Appendix F

Swap Ratio (swapRatio.m)

```
functionsr = swapRatio(x,y)

x=reshape(x,1,[]);
y=reshape(y,1,[]);

nonzeroX = x(x>0 | y>0);
nonzeroY = y(x>0 | y>0);

srVector = abs(atan2(nonzeroY,nonzeroX)/pi*180-45);

sr = mean(srVector);
```

Appendix G

Trip Length Distance Comparison (compareTripLengthDist.m)

```
function
[CR]=compareTripLengthDist(trips,predictedTrips,tripdist,interval,total
Length)

% global y1 y2

m=0:interval:totalLength;

y1=zeros(length(m),1);
y2=zeros(length(m),1);

for k=0:interval:totalLength
    for i=1:size(trips,1)
        for j=1:size(trips,1)
            if tripdist(i,j)>=k && tripdist(i,j)<=k+interval && i~=j
                y1(k/interval+1)=y1(k/interval+1)+trips(i,j);
                y2(k/interval+1)=y2(k/interval+1)+predictedTrips(i,j);
            end
        end
    end
end

tripsPercent=y1./sum(y1);
predictedTripsPercent=y2./sum(y2);

tripsPercentCum=zeros(length(tripsPercent),1);
predictedTripsPercentCum=zeros(length(predictedTripsPercent),1);

for mm=1:length(y1)
    for nn=1:mm
      tripsPercentCum(mm)=tripsPercentCum(mm)+tripsPercent(nn);

predictedTripsPercentCum(mm)=predictedTripsPercentCum(mm)+predictedTrip
sPercent(nn);
    end
end
subplot(1,2,1)
CR=sum(min(tripsPercent,predictedTripsPercent))/sum(max(tripsPercent,pr
edictedTripsPercent));
plot(m,tripsPercent,'o',m,predictedTripsPercent,'*')
xlabel('Trip Length (mile)')
ylabel('Percentage')
set(gca,'XTickLabel',str2num(get(gca,'XTickLabel'))/100);
hleg1 = legend('Survey Trips','Predicted Trips');
title('(a)Trip Length Distribution');
```

```
subplot(1,2,2)
plot(m, tripsPercentCum,'o',m,predictedTripsPercentCum,'*')
xlabel('Trip Length (mile)')
ylabel('Percentage')
axis([0 3000 0 1]);
set(gca,'XTickLabel',str2num(get(gca,'XTickLabel'))/100);
hleg1 = legend('Survey Trips','Predicted Trips');
title('(b)Cumulative Trip Length Distribution');
```

Bibliography

Abrahamsson, Torgil. "Estimation of origin-destination matrices using traffic counts–a literature survey." *IIASA Interim Report IR-98-021/May* 27 (1998): 76.

Alexa.Alexa Internet, Inc. n.d. Web. Visited: 28 June 2013

AirSage. AirSage, Inc. n.d. Web. Visited: August 11, 2013.

Austin. Austin Texas. Austin Convention & Visitors Bureau. n.d. Web. 12 July 2013.

Backstrom, Lars, Eric Sun, and Cameron Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

Barceló, Jaume, et al. "Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring." *Transportation Research Record: Journal of the Transportation Research Board* 2175.1 (2010): 19-27.

Blogg, Miranda, et al. "Travel time and origin-destination data collection using Bluetooth MAC address readers." *Australasian Transport Research Forum (ATRF), 33rd, 2010, Canberra, ACT, Australia*. 2010.

Bing. Bing. n.d. Web. 20 May 2012.

Bisdikian, Chatschik. "An overview of the Bluetooth wireless technology."*Communications Magazine, IEEE* 39.12 (2001): 86-94.

Bossard, Earl G. "RETAIL: Retail trade spatial interaction." *Spreadsheet models for urban and regional analysis* (1993): 419-448.

Brennan Jr, Thomas M., et al. "Influence of vertical sensor placement on data collection efficiency from bluetooth MAC address collection devices." *Journal of Transportation Engineering* 136.12 (2010): 1104-1109.

Brenner, Joanna. “Pew Internet: Mobile” *Pew Internet & American Life Project*.Pew Research Center. 6 June 2013. Web.Visited: 8 July 2013.

Bricka, Stacey. "Non-response challenges in GPS-based surveys." Resource paper prepared for the International Steering Committee on Travel Survey Conference. 2008.

Bricka, Stacy. “Travel Behavior Data to Support Transportation Planning.” Urban Transportation Planning. University of Texas, Austin. 6 February 2013. Lecture.

Bricka, Stacey, and Chandra R. Bhat. "Comparative analysis of global positioning system-based and travel survey-based data." *Transportation Research Record: Journal of the Transportation Research Board* 1972.1 (2006): 9-20.

Bricka, Stacey, Johanna Zmud, Jean Wolf, and Joel Freedman. "Household Travel Surveys with GPS." *Transportation Research Record: Journal of the Transportation Research Board* 2105.1 (2009): 51-56.

Caceres, N., J. P. Wideberg, and F. G. Benitez. 2007. "Deriving origin-destination datafrom a mobile phone network." *Intelligent Transport Systems* no. 1 (1):15-26.

CAMPO. Urban Transportation Study: Report of 2005 Base Year Travel Demand Model Calibration Summary for Updating the 2035 Long Range Plan. Austin: CAMPO, March 2010.

Cascetta, Ennio, and Sang Nguyen. "A unified framework for estimating or updating origin/destination matrices from traffic counts." *Transportation Research Part B: Methodological* 22.6 (1988): 437-455.

Cheng, Zhiyuan, James Caverlee, and Kyumin 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*. ACM, 2010.

Cheng, Zhiyuan, et al. "Exploring Millions of Footprints in Location Sharing Services." *ICWSM* 2011 (2011): 81-88.

Cho, Eunjoon, Seth A. Myers, and Jure Leskovec. "Friendship and mobility: user movement in location-based social networks." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.

CLRSearch. CLRChoice, Inc. n.d. Web. Visited: 10 November 2012.

CNN Money. Time Warner. 21 May 2013. Web. Visited: 12 July 2013.

Doblas, Javier, and Francisco 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* 39.7 (2005): 565-591.

eBizMBA. eBizMBA Inc. n.d. Web. Visited: 28 June 2013.

Erlander, Sven, Sang Nguyen, and Neil Frederick Stewart. "On the calibration of the combined distribution-assignment model." *Transportation Research Part B: Methodological* 13.3 (1979): 259-267.

Fisk, Caroline. S. "Trip matrix estimation from link traffic counts: the congested network case." *Transportation Research Part B: Methodological* 23.5 (1989): 331-336.

Fisk, Caroline S., and David E. Boyce. "A note on trip matrix estimation from link traffic count data." *Transportation Research Part B: Methodological* 17.3 (1983): 245-250.

Fontaine, Michael D., and Brian L. Smith. "Investigation of the performance of wireless location technology-based traffic monitoring systems." *Journal of transportation Engineering* 133.3 (2007): 157-165.

Foursquare. Foursquare. n.d. Web. Visited: 6 July 2013.

Giaimo, Greg, et al. "Will It Work?." Transportation Research Record: Journal of the Transportation Research Board 2176.1 (2010): 26-34.

Hainen, Alexander M., et al. "Estimating Route Choice and Travel Time Reliability with Field Observations of Bluetooth Probe Vehicles." *Transportation Research Record: Journal of the Transportation Research Board* 2256.1 (2011): 43-50.

Herrera, Juan C., et al. "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment." *Transportation Research Part C: Emerging Technologies* 18.4 (2010): 568-583.

Howden, Lindsay M., and Julie A. Meyer. *Age and sex composition: 2010*. US Department of Commerce, Economics and Statistics Administration, US Census Bureau, 2011.

Ignite Social Media. Ignite Social Media. 31 July 2012. Web. Visited: 28 June 2013.

ITE. Institute of Transportation Engineers. Washington, DC. 2013. Web. Visited: 2 July 2013

ITE. *Trip Generation: An Informational Report.* 5th ed. Washington DC: Institute of Transportation Engineers, 1991.

"INTRIX Technology Breakthrough Significantly Improves Accuracy of Real-Time Traffic Information for Navigation on Arterials, City Streets and Secondary Roads." *INRIX*.com. n.p. 6 January 2010. Web. Visited: 9 July 2013.

Jin, Peter J, Fan Yang, Meredith Cebelak, Bin Ran, and C. Michael Walton. Urban Travel Demand Analysis for Austin TX USA Using Location-Based Social Networking Data. Poster session presented at: *Transportation Research Board92nd Annual Meeting*; January 13-17, 2013; Washington, DC.

Karimi, Hassan A. "Genetic Location-Based Social Networks (G-LBSN)."*Proceedings of the 3rd International Workshop on Location and the Web*. ACM, 2010.

Kawakami, Shogo, Huapu Lu, and Yasuhiro Hirobata."Estimation of Origin--Destination Matrices from Link Traffic Counts Considering the Interaction of the Traffic Modes." *Papers in Regional Science* 71.2 (1992): 139-151.

"Key Facts" *Facebook*. Facebook, 2013.Web. 29 June. 2013.

LeBlanc, Larry J., and Keyvan Farhangian."Selection of a trip table which reproduces observed link flows." *Transportation Research Part B: Methodological* 16.2 (1982): 83-88.

Li, Nan, and Guanling Chen. "Analysis of a location-based social network." *Computational Science and Engineering, 2009.CSE'09.International Conference on*. Vol. 4.IEEE, 2009.

Macfarlane, Gregory and Laurie Garrow. Estimating a Vehicle Ownership Model on Targeted Marketing Data. Poster session presented at: Household Travel Survey Symposium, *Travel Surveys: Moving from Tradition to Practical Innovation*; November 8-9, 2012; Dallas, TX.

Martin, William A., and Nancy A. McGuckin. *Travel estimation techniques for urban planning*. Vol. 365. Washington, DC: National Academy Press, 1998.

Mathew, Tom V., and K. V. Krishna Rao."Introduction to Transportation engineering." *Civil Engineering–Transportation Engineering. IIT Bombay, NPTEL* (2007).

MATLAB. The Math Works, Inc. n.d. Web. Visited 25 July 2013.

McNally, Michael G. *The Four Step Model*. Center for Activity Systems Analysis, Institute of Transportation Studies, University of California Irvine, Irvine, CA. November 2008.

Mitchell, Robert Buchanan, and Chester Rapkin."Urban Traffic--A Function of Land Use." (1954).

Murakami, Elaine, and David P. Wagner. "Can using global positioning system (GPS) improve trip reporting?." *Transportation Research Part C: Emerging Technologies* 7.2 (1999): 149-165.

NHTS.FHWA. 2009. Web. Visited: 6 July 2013.

Pan, Changxuan, et al. "Cellular-based data-extracting method for trip distribution." *Transportation Research Record: Journal of the Transportation Research Board* 1945.1 (2006): 33-39.

"Revision of the Commission's Rules to Ensure Compatibility with Enhanced 911 Emergency Calling Systems."11 FCC 18676. 1997

Sayed, Ali H., Alireza Tarighat, and Nima Khajehnouri. "Network-based wireless location: challenges faced in developing techniques for accurate wireless location information." *Signal Processing Magazine, IEEE* 22.4 (2005): 24-40.

Scellato, Salvatore, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo."Socio-Spatial Properties of Online Location-Based Social Networks." *ICWSM* 11 (2011): 329-336.

Schlaich, Johannes, et al. "Generating trajectories from mobile phone data." Proceedings of the 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies. 2010.

Sen, SB Sudeshna, and S. Bricka." Data Collection Technologies–Past, Present, and Future." *International Conference on Travel Behaviour Research*. 2009.

Sharp, Joy, and Elaine Murakami. "Travel surveys: Methodological and technology-related considerations." *Journal of Transportation and Statistics* 8 (2005): 97.

The Transportation Planning Process: Key Issues. Transportation Planning Capacity Building Program, FHWA. September 2007. FHWA-HEP-07-039.

TMIP.FHWA, 2013.Web. 10 July 2013.

TMIP. *Travel Model Validation and Reasonableness Checking Manual*. 2nd Edition. Federal Highway Administration, Washington DC. 24 September 2010.

Tornero, Rafael, Javier Martínez, and Joaquín 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*. ACM, 2012.

TTI.*ATOM2 User Manual.* Austin, TX: TxDOT, Transportation Planning & Programming Division, February 2001.

"U.S. Wireless Quick Facts." *CTIA -The Wireless Association*. n.d. Web. Visited: 8 July 2013.

U.S. Census Bureau (USCB). USCB. 27 June 2013. Web. Visited: 12 July 2013.

Utsunomiya, Mariko, John Attanucci, and Nigel Wilson. "Potential uses of transit smart card registration and transaction data to improve transit planning." *Transportation Research Record: Journal of the Transportation Research Board*1971.1 (2006): 119-126.

Van Zuylen, Henk J., and Luis G. Willumsen. "The most likely trip matrix estimated from traffic counts." *Transportation Research Part B: Methodological*14.3 (1980): 281-293.

Wasserman, Stanley. *Social network analysis: Methods and applications*. Vol. 8.Cambridge University Press, 1994.

Watson, James R., and Panos D. Prevedouros. "Derivation of origin-destination distributions from traffic counts: Implications for freeway simulation." *Transportation Research Record: Journal of the Transportation Research Board*1964.1 (2006): 260-269.

Weiner, Edward. Urban transportation planning in the United States: An historical overview. Greenwood Publishing Group, 1999.

Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 22 July 2004. Web. 10 Aug. 2004.

Wilson, A. G. "A Statistical Theory of Spatial Distribution Models." Transportation Research, Volume 1, Issue 3. November 1967: 253-269.

Wilson, A. G. "The Use of Entropy Maximising Models in the Theory of Trip Distribution, Mode Split and Route Split." Journal of Transport Economics and Policy, Volume 3, Number 1. January 1969: 108-126.

Wolf, Jean, Randall Guensler, and William 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* 1768.1 (2001): 125-134.

Yang, Fan, Peter J. Jin, Yang Cheng, and Bin Ran. "Origin-Destination Estimation for Non-Commuting Trips Using Location-based Social Networking Data." *International Journal of Sustainable Transportation*. Accepted. 2014

Yim, Youngbin. "The state of cellular probes." California Path Program, Institute of Transportation Studies. Berkeley, CA. July 2003.

Zheng, Yu. "Tutorial on location-based social networks."*WWW2012* (2012).

Zheng, Yu, Xing Xie, and Wei-Ying Ma. "GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory." *IEEE Data Eng. Bull.* 33.2 (2010): 32-39.