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

Transportation Planning via Location-Based Social Networking Data: Exploring Many-to-Many Connections

Committee:

C. Michael Walton, Supervisor

Chandra Bhat

Stephen D. Boyles

Jennifer Duthie

Keri Stephens

Daniel Yang

Transportation Planning via Location-Based Social Networking Data: Exploring Many-to-Many Connections

by

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

Dissertation

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

Doctor of Philosophy

The University of Texas at Austin

August 2015

Acknowledgements

I would like to acknowledge the following individuals, without who my success would not have been possible: Dr. C. Michael Walton,, who served as my advisor for during my time at the University of Texas at Austin, my dissertation committee members, Dr. Chandra Bhat, Dr. Steve Boyles, Dr. Jen Duthie, Dr. Keri Stephens, Dr. C. Michael Walton, and Dr. Daniel Yang, who all provided great insight into this dissertations topic and helped to steer the effort with their vision; Dr. Peter Jin, who provided the initial inspiration for this dissertation and has been an invaluable asset throughout the process; Dr. Katie Kam, who provided assistance with the GIS component of the analysis; Dr. Jia Li, who provided hours of assistance with the coding component, helping throughout the de-bugging; and my family, especially my parents, who provided endless support and encouragement.

Transportation Planning via Location-Based Social Networking Data: Exploring Many-to-Many Connections

Meredith Kimberly Cebelak, Ph.D.

The University of Texas at Austin, 2015

Supervisor: C. Michael Walton

Today's metropolitan areas see changes in populations and land development occurring at faster rates than transportation planning can be updated. This dissertation explores the use of a new dataset from the location-based social networking spectrum to analyze origin-destination travel demand within Austin, TX. A detailed exploration of the proposed data source is conducted to determine its overall capabilities with respect to the Austin area demographics. A new methodology is proposed for the creation of origin-destination matrices using a peer-to-peer modeling structure. This methodology is compared against a previously examined and more traditional approach, the doubly-constrained gravity model, to understand the capabilities of both models with various friction functions. Each method is examined within the constructs of the study area's existing origin-destination matrix by examining the coincidence ratios, mean errors, mean absolute errors, frequency ratios, swap ratios, trip length distributions, zonal trip generation and attraction heat maps , and zonal origin-destination flow patterns.

Through multiple measures, this dissertation provides initial interpretations of the robust Foursquare data collected for the Austin area. Based upon the data analytics performed, the Foursquare data source is shown to be capable of providing immensely detailed spatial-temporal data that can be utilized as a supplementary data source to

traditional transportation planning data collection methods or in conjunction with other data sources, such as social networking platforms. The examination of the proposed peer-to-peer methodology presented within this dissertation provides a first look at the potential of many-to-many modeling for transportation planning. The peer-to-peer model was found to be superior to the doubly-constrained gravity model with respect to intrazonal trips. Furthermore, the peer-to-peer model was found to better estimate productions, attractions, and zone to zone movements when a linear function was used for long trips, and was computationally more proficient for all models examined.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

This dissertation expands on the Master's thesis, "Location-Based Social Networking Data: Doubly-constrained Gravity Model Origin-Destination Estimation of the Urban Travel Demand for Austin, TX," (Cebelak 2013) concept of using location-based social networking for transportation planning. While the thesis examined the doubly-constrained gravity model, this effort proposes using many-to-many connections models as another viable transportation planning model for determining origin-destination patterns.

MOTIVATION

Transportation planning, an essential part of a community's ability to plan for the future, has been studied since the 1940's (Weiner 1986). The goal of the planning effort is to understand where trips begin and end, what modes are being used, and what roadways and pathways are being utilized. Early efforts to determine human mobility patterns used the growth factors to distribute future origin-destination travel data in what is known as the Fratar method (Brokke and Mertz 1958). This in turn lead to the now commonly used four step model (McNally 2008), which is comprised of trip generation, distribution, mode split, and traffic assignment. This dissertation will focus on the trip generation and distribution steps of the four step model and will present a novel approach that make use of many-to-many concepts.

Within relational database theory, the "many-to-many" connections concept refers to relationships between a "parent" or entity row and several "child" or characteristic rows as well as the relationship between a "child" row and several "parent" rows, and has been described as a "mirror of the real-life relationship between the objects" (Janssen 2014). The concept has been applied to a variety of disciplines including business,

marketing, technology based industry, as well as anthropology. The movements of individuals can be influenced by business marketing and social anthropology, and thus can be analyzed under the spectrum of many-to-many connections. There are three many-to-many modeling structures that focus in these areas: business-to-customer, peer-to-peer, and social forces.

Business-to-customer (B2C), where the "parent" role is filled by businesses and the "child" role is filled by customers, has been researched with respect to the transportation field; one of the earlier examples was in 2001. This effort (TRIP 2001) examined the relationship between the economy and the transportation system within the US, with respect to freight movement, noting the higher levels and greater reliability for freight transport for business-to-business and business-to-customer exchanges. After this initial effort, two additional research efforts were done in the early part of this century. Both used B2C to further analyze trends within the freight industry, specifically the parcel component of the industry (Pagano 2001, Rabah and Mahmassani 2002). Research in this area had a brief hiatus until 2006, after which a handful of related research was conducted examining the relationships between suppliers and customers with respect to logistic services (Davis and Mentzer 2006, Leinbach 2007, Park, Min, and Park 2011), aviation (Franke 2007), and personal vehicles (Aboltins and Rivza 2014).

With regard to the transportation industry, the research within the social forces genre of many-to-many connections in recent years focuses primarily on pedestrian interactions. In 1991, Helbing proposed a mathematical model to describe the movement of pedestrians that became the basis for his later effort that related behavioral changes in pedestrians to social forces, which were defined as external influences or the environment, public opinion, and social norms and trends (Helbing 1994). This effort lead to studies in crowd dynamics (Helbing et al. 2005), bottleneck flow for pedestrians

(Kretz, Hengst, and Vortisch 2008), prediction and simulation of pedestrian movements (Rudloff et al. 2011, Deroo and Auberlet 2012, Duives, Daamen, and Hoogendoorn 2013), and pedestrian route choice (Werberich et al. 2014).

Within the peer-to-peer modeling dynamic, transportation research trends are similar to those seen in the social forces modeling. Where social forces research has focused mainly on pedestrians, in recent years, peer-to-peer transportation modeling has concentrated in the carsharing spectrum, but to a lesser degree (Hampshire and Gaites 2011, Rivasplata et al. 2012, Chen, McNeil, and Dill 2014, Ballús-Armet et al. 2014, Dill, Howland, and McNeil 2014). Prior to this, one of the first transportation related (via supply chain management) research efforts in peer-to-peer modeling examined the use of an e-supply chain portal to overcome large number of peer-to-peer relations for complex organizations focusing on business modeling (Boyson, Corsi, and Verbraeck 2003). Additionally early efforts focused on vehicle-to-vehicle information sharing (Yang and Recker 2006, Yang and Recker 2008), signals (Coplen 2007, Sabra and Riniker 2009), and college age driver incidents (Tisdale 2013).

Data collection is a critical part of any modeling effort. Conventionally, the four step model has employed the traditional household survey, traffic counts, and position technologies (i.e., GPS, Bluetooth) as data collection methods. Recent research has explored the opportunities to use smartphones (INRIX 2010, AirSage 2014), and other data sources including vehicle to infrastructure (V2I) communications (Tornero, Martínez, and Castelló 2012). While all of these data collection methods have pros and cons, the most notable con for them is the cost of the data. Data collection through social networking can range from free to a few thousand dollars, while the conventional methods range from tens of thousands of dollars to hundreds of thousands of dollars.

Location-based services (LBS), which use location and time data, have four distinct areas of concentration: maps/navigation, tracking, information, and applications. Location-based social networking (LBSN) falls within the applications category of LBS combining it with social networking sites like Facebook, Twitter, and Foursquare. As tablets and smartphones are owned and used by more of the population, this data source has becomes a more population representative data source. Thus, researchers have begun to mine this data source to better understand user's spatial patterns, geographic movements, temporal dynamics, networking ties, and location predictions. The first efforts to explore spatial patterns of users of LBSN used Markov-based location predictors to determine future locations of users (Li 2009). Further studies of user location prediction based on a user's friends (Backstrom, Sun, and Marlow 2010) and a user's content (Cheng, Caverlee, and Lee 2010) were examined in subsequent years and proved valuable. Consequently, research explored LBSN's data relationships between geographic movements (Cho, Myers, and Leskovec 2011), human movement's temporal dynamics (Zheng, Xie, and Ma 2010), as well as the links of social networking (Karimi 2010).

The most popular LBSN site is Foursquare, which has over 55 million users and over seven billion check-ins (Foursquare 2015), has users that include business and individuals. Researchers have recently begun to explore this site's data set due to its popularity, high penetration rate, and large sample size, specifically to explore mobility patterns across spatial, temporal, and social aspects among users (Cheng, Caverlee, and Lee 2011, Scellato et al. 2011). Additionally, the site's data has been explored more recently for its transportation planning application. The 2011 study by F. Yang et al. (2014) specifically used Foursquare data to estimate an origin-destination matrix for the Chicago urban area, and was among the first to demonstrate the data's potential for use in

transportation planning. Continuing this effort, the modeling technique was applied to the Austin, TX area using a singly-constrained gravity model (Jin et al. 2013). To further analyze the use of Foursquare data for transportation planning, a doubly-constrained gravity model was proposed for the Austin area and demonstrated better learning capabilities when compared to the singly-constrained gravity model (Jin et al. 2014). Finally, exploration of the data with respect to mode choice revealed the data set's potential to provide information on select modes (airplane, bus, rail, and bicycle), but could not provide any insight on the walk or automobile modes (Cebelak, Jin, and Walton 2014). While these efforts have utilized the data from Foursquare, an in-depth and detailed exploration of Foursquare as a data source has not been done to date. This has limited the realization of the data source's potential with respect to transportation planning. This dissertation will include a detailed examination that will consist of identifying day-to-day as well as time dependent trends, a first for the industry.

RESEARCH QUESTIONS

Due to the lack of exploration in many-to-many modeling with respect to transportation planning and the novel nature of the use of LBSN as a data source for transportation planning, this dissertation attempts to apply these modeling concepts in conjunction with the LBSN database to answer the questions:

1. Will many-to-many modeling provide a more insightful origin-destination matrix for the Austin area when compared to the doubly-constrained gravity model method while employing LBSN data using the local MPOs origin-destination model as a base comparison?
2. How impactful is LBSN as a data sources? How well does the data set represent the existing demographic of the study area with respect to land

use? Can it be used as a stand-alone data set or are additional complementary data sets needed?

In answering these questions, this dissertation will provide a first in-depth examination of a Foursquare dataset through the analysis of check-in data characteristics over the spatial-temporal range. This analysis will reveal the data sources strengths, limitations, and potential for usage within the transportation planning due to the richness of the data available. Further, the foray into peer-to-peer modeling will be the first exploration pertaining to the field of transportation planning and its comparison to the doubly-constrained gravity model will disclose the model's capabilities and potential usage for industry practitioners.

ORGANIZATION OF DISSERTATION

The remaining sections of this document are organized as follows. Chapter 2 presents the literature review, which covers the research conducted within the areas of transportation planning, location-based social networking, and many-to-many modeling. Chapter 3 explores the location-based social networking dataset used within this dissertation with respect to category, time of day, and day of week for venue check-ins, user demographics, and the relationship between land use and venue categories. Chapter 4 provides details on the methodologies used to analyze variations of doubly-constrained gravity and peer-to-peer models using the location-based social networking data. Chapter 5 gives the results of a case study examined using the proposed methodologies from Chapter 4. Chapter 6 offers the conclusion of the dissertation efforts and presents future areas of exploration.

Chapter 2: Literature Review

This chapter examines the literature reviewed in an effort to explore the relevant topics of transportation planning and modeling approaches, trends in data collection for the transportation planning process, the area of social media and location-based social media, and the many-to-many model. The endeavor aims to provide a foundation for the relevance and value of this dissertation.

TRANSPORTATION PLANNING

Since the 1940's, Transportation planning has been studied as an essential part of a community's ability to plan for the future (Weiner 1986). Prior to this time period, the focus of transportation planning was limited to the collection and analysis of existing information with little thought given to the future. Not until the post-World War II boom in automobile demand, which the existing infrastructure was not equipped to handle, and the city dwellers migration to the suburbs was there a need and interest to further the efforts in transportation planning. In 1944, the need to understand the complexities of urban street systems from a trip origin and destination perspective led to the development of the home-interview origin-destination (OD) survey (Weiner 1999).

The 1950s brought new ideas and techniques to urban transportation planning to determine human mobility patterns using growth factors to distribute future OD travel data in what is known as the Fratar method (Brokke and Metz 1958). Following this effort, Robert B. Mitchell and Chester Rapkin (1954) established a link between travel and activities in an effort to create a thorough framework for exploration into travel behavior. Based upon the work of Mitchell and Rapkin, the now commonly used four-step model (McNally 2008), which is comprised of trip generation, distribution, mode split, and traffic assignment, was first comprehensively applied in the Chicago Area

Transportation Study in the 1950s. Further discussion of this model is provided in the next portion of this chapter. The 1960's brought Federal legislation that required "continuous, comprehensive, and cooperative" urban transportation planning and in the 1970s environmental concerns and multimodal elements were included in the requirement (McNally 2008)

Modern day efforts in transportation planning incorporate the views of various transportation agencies and the general public within the analysis of potential strategies (The Transportation Planning Process 2007). These efforts include the monitoring of existing conditions, forecasting future populations and employment growth through the assessment of projected land uses and identification of major growth corridors, identifying both current and future transportation issues and needs, developing long-range plans and short-range programs, estimating impacts from recommended future improvements, as well as developing financial plans for the implementation of strategies. The Federal Highway Administration's (FHWA) overview of their Transportation Planning Process is shown in Figure 1.

Figure 2.1: FHWA's Transportation Planning Process (The Transportation Planning Process 2007).

Metropolitan Planning Organizations (MPOs) use transportation planning models to simulate the impacts of changes to their system and aid in their decision making process. The models employed include the traditional land use models, the emissions models, the four-step models, and the activity-based models. The land use models, used often for forecasting future development patterns, and the emissions models, used for examination of key pollutants from vehicle exhaust, are outside the focus of this effort and will not be further discussed. The four-step and activity-based models will be discussed further in the subsequent sections of this chapter.

Four-Step Travel Demand Model

For the prediction of demand for transportation services, the four-step model is one of the most commonly used models by MPOs comprising of the four-steps of trip

generation, trip distribution, modal split, and network assignments. The first two steps are the main focus of this dissertation.

The first step of the four-step model, trip generation, measures the frequency of trips. The earliest effort of trip generation was in San Juan, Puerto Rico in 1948 in a study that developed rates for land use categories based on location, intensity, and type of activity (Weiner 1999). Following this effort, the Detroit Metropolitan Area Traffic Study in 1955 developed trip generation rates for each zone within the study area by land use category. In 1972, the Institute of Transportation Engineers' (ITE) Trip Generation Committee was tasked with the collection and compilation of existing trip generation rate data. Published in 1976, the first edition of this effort, *The Trip Generation*, contained data from nearly 80 sources. The 1991 version and 5th edition was considered the most comprehensive database containing trip generation rates for 121 land use categories from over 3,000 studies. The most current edition is the 9th edition and was published in 2012 containing rates for 172 land uses based on over 5,500 studies (ITE 2013). For transportation professionals, the ITE Trip Generation reports are the most widely used reference for trip generation data with respect to site level planning and analysis (Weiner 1999).

For MPOs, the trip generation process's goal is to determine the magnitude of total daily travel at the household and traffic analysis zonal level for all trip purposes included within the study. These trip purposes typically include at a minimum three types: home-based work, home-based non-work, and non-home-based (McNally 2008). The trip end points are modeled as either productions or attractions within this transportation demand model.

Upon attaining the trip generations for the study area, the second step of the model, trip distribution, is undertaken. This process recombines the production and

attraction rates for each traffic analysis zone (TAZ) and creates a matrix of the number of trips occurring between each origin and destination TAZ (McNally 2008). The recombination effort is done using models which include, but are not limited to, logit, entropy, growth, and gravity. Within the logit models, the multinomial logit destination choice model is commonly used with activity-based models (Bhat and Koppelman 1999). The entropy maximizing method was established by Wilson in 1976 and was able to relate the probability of the distribution of trips occurring in an OD pair to the number of states of the system (Wilson 1967). Growth models have two variations: uniform, which only requires a general growth rate for the study area, and constrained, which uses information on the growth of the number of trips that originate and terminate within each zone allowing for different factors to be utilized (O'Flaherty 1997). While both versions of the growth model benefit from simplicity, the uniform method suffers from the assumption of a single growth factor for all zones and attractions, and the doubly-constrained suffers from its heavy dependence on observed trip patterns and the lack of inclusion of changes in travel costs within its trip distribution. Gravity models will be discussed in depth within the following section.

Gravity Models

The aggregate gravity model originated from Newton's gravitational law (Mathew and Rao 2007), which states that force, F, is related to the gravitational constant, *G*, the masses of two objects, m_1 and m_2, and the distance between the objects, *d*, and is formulated as follows:

$$F = \frac{G * m_1 * m_2}{d} \quad (Eqn. 2.1)$$

The trip distribution formula is analogous with this Newtonian formula and is shown below in the general form, with the following relationships:

- 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).

$$T_{ij} = \frac{C * O_i * D_j}{c_{ij^n}} \quad (Eqn. 2.2)$$

To ensure the total number of productions and attractions are equal, a balancing factor (b) is added to Equation 1.2 to either the productions or attraction factors for the singly-constrained gravity model (Equation 2.3), which attempts to preserve zonal inputs for the productions only (TMIP 2010). Additionally, within this model the general travel cost term, $\left(c_{ij^n}\right)$ from Equation 2.2, is replaced by a friction function ($f(c_{ij})$) to de-incentivize travel based on time via distance or a cost increases. Further details on friction functions are discussed in a subsequent section of this chapter.

$$T_{ij} = b * O_i * D_j * f\left(c_{ij}\right) \quad (Eqn. 2.3)$$

The doubly constrained gravity builds upon the singly-constrained gravity model and attempts to preserve zonal inputs for the productions and attractions (TMIP 2010). This model encompasses balancing factors for both the productions and the attractions and its equation is shown below (Mathew and Rao 2007). Within this equation, the balancing factor for the productions is defined by β_i, and α_i defines the balancing factor for the attractions.

$$T_{ij} = \beta_i * O_i * \alpha_j * D_j * f(c_{ij}) \quad (Eqn.\ 2.4)$$

The sum of the total trips for each destination should equal the sum of the combination of productions, attractions, balancing factors, and friction functions for each destination. Using this principle, Equation 4 can be manipulated into Equation 2.5.

$$\sum_i T_{ij} = \sum_i \beta_i * O_i * \alpha_j * D_j * f(c_{ij}) \quad (Eqn.\ 2.5)$$

The sum of trips in any specific row or column of the OD matrix should equal the total number of trips produced in that zone as shown in Equations 2.6 and 2.7.

$$\sum_j T_{ij} = O_i \quad (Eqn.\ 2.6)$$

$$\sum_i T_{ij} = D_j \quad (Eqn.\ 2.7)$$

From Equation 4 and 7, balancing factors (β_i, α_j) can be found (Equation 2.8 and 2.9).

$$\beta_i = \frac{1}{\sum_j \alpha_j * A_j * f(t_{ij})} \quad (Eqn.\ 2.8)$$

$$\alpha_j = \frac{1}{\sum_i \beta_i * P_i * f(t_{ij})} \quad (Eqn.\ 2.9)$$

Using equation 2.8 and 2.9 with separate singly constrained models, the following formulas can be used to find the T_{ij} for each O-D pair from this model using an iteration process similar to the Furness method (Mathew and Rao 2007).

$$T_{ij} = O_i * \frac{D_j * f(t_{ij})}{\sum_j D_j * f(t_{ij})} \quad (Eqn.\ 2.10)$$

$$T_{ij} = A_j * \frac{P_i * f(t_{ij})}{\sum_i P_i * f(t_{ij})} \quad (Eqn.\ 2.11)$$

As used in the singly- and doubly-constrained gravity models, the friction function ($f(c_{ij})$) de-incentivizes travel based on the increase in time via distance or the cost. This "deterrence function" (Mathew and Rao 2007) can use a variety of formulations to appropriately calculate the impedance including the linear function, negative exponential, power, and gamma function, (Bossard 1993, Mathew and Rao 2007), which are shown in the equations below.

$$\text{Linear: } f(c_{ij}) = \alpha + \beta * d_{ij} \quad (Eqn. 14)$$

$$\text{Negative exponential: } f(c_{ij}) = \alpha e^{-\beta * d_{ij}} \quad (Eqn. 15)$$

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

$$\text{Gamma: } f(c_{ij}) = \alpha * {d_{ij}}^{-\beta} * e^{-\gamma * d_{ij}} \quad (Eqn. 17)$$

In the above equations, α is a positive scaling factor that controls the overall range of the function values, β is a 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.

The triply-constrained gravity model, the atomistic gravity model, has constraints on the productions, attractions, and the trip length frequency. The addition of the constraint on the trip length frequency makes the model self-calibrating for both intra-zonal and inter-zonal trips. For this model, the TAZs are represented by an abstract discrete spatial surface made up of "atoms" that are disbursed throughout the TAZ.

Additionally, the model uses travel time instead of distance for the zonal radii, which are then used 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 with p_{i_v} represents the trips produced by atom *v* of zone i, a_{j_q} the relative attraction factor for atom q of zone j, $F_{d_{vq}}$ the relative trip length factor for the estimated separation between atom pair vq, $K_{s_{ij}}$ the bias factor for sector pair containing zones i and j, M_y the number of atoms in zone y, and N represents the number of zones.

$$T_{ij} = O_i * \frac{\sum_{v=1}^{Mi} \sum_{q=1}^{Mj} p_{i_v} * a_{j_q} * F_{d_{vq}} * K_{s_{ij}}}{\sum_{x=1}^{N} \sum_{n=1}^{Mj} \sum_{m=1}^{Mx} p_{i_n} * a_{x_m} * F_{d_{nm}} * K_{s_{ix}}} \quad (Eqn.\ 2.12)$$

O_i, the total trips produced in zone i, is calculate using the following formula.

$$O_i = \sum_{m=1}^{Mi} p_{i_m} \quad (Eqn.\ 2.13)$$

Activity-Based Travel Demand Model

Recently, activity-based travel demand models have gained popularity in the United States (US). This model type generates activities, identifies destinations for the activities, determines the mode used for travel, and predicts the particular network route used (Castiglione, Bradley, and Gliebe 2014). These models examine travel through "trip chaining" where multiple trip legs are chaining trips into tours. In comparison to four-step models, activity-based models are able to represent the realistic constraints of time and space as well as the connections between activities and travel for individuals as well as multiple people within a household (Castiglione, Bradley, and Gliebe 2014). This is done through the understanding of behavior theories with respect to how individuals decide to or not to participate in activities. Decisions on where and when to participate in

activities are also included within the analysis. Table 1 shows a comparison between trip-based and activity-based models.

Model Type	Spatial/ Temporal Detail	Person/ Household Detail	Sensitivity to Policy	Run Time	Cost
Four-step	Low to Medium	Medium	Medium	Medium	Medium
Activity-based	Medium to High	High	Medium to High	Medium	Medium

Table 2.1: Model Comparison Between Four-step and Activity-based (Castiglione, Bradley, and Gliebe 2014)

The activity-based approach is more disaggregated in time, space, and activities; thus, the models are better suited for analyzing complex policy alternatives, i.e., flexible work hours and variable pricing schemes (Bhat and Koppelman 1999). Additionally, the models can be used to produce detailed performance metrics that can be used to support equity analysis, regional planning, as well as regional air quality, transit, and transportation demand management forecasting. For these reasons, many municipalities have begun to move toward using the activity-based model. With this in mind, this dissertation will explore how the location-based social networking data may be used to aid municipalities towards this modeling structure and how the data may be used within an activity-based approach in a latter chapter.

DATA COLLECTION METHODS

One of the fundamental components for the creation of an OD matrix, regardless of model type, is the data collection. There are a variety of methodologies employed for data collection including the more traditional methods of household surveys and traffic counts, the increasingly utilized technology-based methods of using global position and

cellphones, and future methods that are being researched and include innovative data sources and connected vehicle technologies.

Traditional Methods

Household Surveys

Data has conventionally come from traditional household travel behavior surveys for the creation of OD matrices. These surveys collect data that includes trip purpose, transportation mode used, trip duration, time of day as well as the day of the week the trip took place, vehicle occupancy when personal vehicle is used, and personal demographic information including age, sex, employment status, income, and education level (NHTS 2013). Surveys data can be collected via personal home interviews, telephone interviews, by mail, or by internet. For the personal home interviews, an interviewer is required to visit the respondent's home or office to administer questions in a face-to-face interview (Sharp 2005). This method provides one of the most complete data sets with the highest response rates of 60-70% (Giaimo et al. 2010) when compared to other methods covered within this section. Despite this high quality data, the conduction of this method is the most expensive and time consuming.

The telephone interview method requires interviewers to contact individuals via telephone to administer the survey. Sample bias exists for this method since it limits participants to only those households with telephones and response rates are intermediate in quality of data and cost and range from 25 to 40% (Giaimo et al. 2010). For the mail survey format, a questionnaire is mailed out to respondents with the results returned either by mail or telephone. The coverage for this method is similar to that of the personal home interview method; however, it has the lowest response rates of 20-30% (Giaimo et al. 2010), and low data quality rates. The method does have the advantage of beings one

of the least expensive methods for the household surveys. The final method of survey deployment is the internet, which is similar to the mail format of survey deployment, but places the survey on the internet for respondents to complete. Similar to the sample bias discussed for the telephone method, only households with internet access are able to participate. The response rates for this method are similar to those of the mail method with intermediate data quality and while the costs are lower for this method, there are higher startup costs associated with the uses of survey platforms.

Traffic Counts

In 1979 a study by Erlander, Nguyen, and Steward were the first to demonstrate the ability to create a unique OD matrix if traffic counts for all links were available. However, the method required that detector infrastructure would be deployed throughout the study area on all viable routes between OD pairs, which is typically cost prohibitive for municipalities. This study was followed by additional research in the 1980s. Using observed traffic data, OD matrix estimation was conducted for a networks having more than 70 links (Van Zuylen and Willumsen 1980, LeBlanc and Farhangian 1982). Fisk and Boyce recognized that traditionally only a sample of traffic count data is available and proposed a method for estimating link cost functions and formulated doubly-constrained distribution assignment model for this data (Fisk and Boyce 1983). The formulated model by Fisk and Boyce was extended to include two travel modes in the work by Kawakamik, Lu, and Hirobata in 1992. Fisk furthers the 1983 effort with respect to the congested network scenario and examines three different formulations to create the OD matrices (Fisk 1989). The use of two modes of traffic counts (cars and transit) was also examined for OD matrix estimation by Cascetta and Nguyen (1988) using classical and Bayesian statistical inference techniques for the model framework and algorithm development. In

the 1998 study by Abrahamsson, traffic volumes for each link within a system were used to create an OD matrix. The author notes that many different OD matrices could be reproduced from the observed traffic counts.

Recently, this method has been shown to be possible in practice (Watson and Prevedouros 2006, Doblas and Benitez 2005). However, these works concede that detectors would need to be installed for full coverage of the network to prevent large data gaps which in turn would lead to operation and maintenance costs that would be an expensive long term commitment. Fontaine and Smith (2007) noted that there are concerns with the accuracy of estimated traffic conditions between detectors as they only provide data at fixed locations. An additional limitation of the data set, as stated in Abrahamsson's effort, is the need for a "target" OD matrix to verify the methodology, which would typically come from prior information on the anticipated or existing OD matrix.

Technology-Based Methods

Global positioning systems (GPS), cellphones, and Bluetooth technologies have benefited from advances in position technologies and have made these data sources viable for traffic flow monitoring, providing traveler information, and advanced traffic and demand management. Survey researchers have used these methods in simulation efforts and field deployments.

Global Positioning Systems

A satellite-based positioning system was initiated by the US Military in the 1970s. This system became the fully operational GPS in 1995 (Sen and Bricka 2009). The quick adoption of the technology by the domestic and international research community due to

the positioning requirement for travel surveys can be attributed to relative low-cost, high accuracy of the technology (Bricka 2008).

In 1996, the first GPS travel survey effort was done in Lexington, Kentucky as a proof-of-concept (Murakami and Wagner 1999). Personal Digital Assistants (PDAs) equipped with GPS were utilized to capture vehicle based daily trip information for 100 households over six days. The study had two goals: to identify an alternative to trip diaries that was cost effective and to determine individual participant's willingness to use this data collection methodology. In addition to these goals, the study was able to demonstrate the technologies ability to collect information on route choice and travel speed. Following this initial work, Wolf, Guensler, and Bachman (2001), noting that previous efforts had only used GPS as a supplemental data source, attempted to utilize GPS as replacement data collection source and demonstrated the ability of the technology to collected personal vehicle travel data using a geographic information system (GIS) to derive the traditional diary elements. To validate this small scale effort, Wolf et al. compared the derived travel diary data with paper diary data finding matching or superseding diary elements from the GPS data source. A large scale effort demonstrating the proof-of-concept was conducted by Giaimo et al. in 2010. This first of its kind study examined the replacement of travel diaries with a multiday GPS survey for the Greater Cincinnati Household Travel Survey and was made up of a fully representative sample (household size, income, age, geographic region, etc.) where data was recorded for up to three days of travel. Resulting in completion rates that were acceptable and representative, the method showed that participants were not additionally burdened in carrying devices. This effort, however, was not without its drawbacks. Significant incentives and logistical issues, including the timely retrieval of GPS units, GPS unit loss rates, and battery outages, were noted with these logistical issues resulting in incomplete

data. Additionally, the software used with the data had some limitations including the map editing process requiring the review of data to ensure its appropriate incorporation, and the miss identification of trips as two trips due to a stop within the trip or the loss of the GPS signal.

Research has also been conducted using GPS to determine the characteristics of underreporting which occurs with traditional survey methods. Bricka and Bhat (2006) conducted a comparative examination of GPS and the traditional household survey to determine the level and likelihood of underreporting finding that individuals under 30, males, individuals with less than a high school education, those who were unemployed, those who make many trips, traveled long distances, and those who trip chained were likely to underreport. For GPS surveys, challenges with non-responses were identified by Bricka in 2008, who noted the burdens of survey length (duration of study time), privacy concerns, and equipment complications. The study noted that non-responses were found to be associated with older, less educated, and low income participants, which followed the trends associated with technology acceptors typically being young, highly education, higher income males. This indication of a sampling bias was supported by the Oregon Household Travel Survey test pilot, which utilized GPS as its data collection method and suggests that other methods of data collection may be more appropriate and noted that the cost for the GPS-based survey was over two times as expensive as traditional methods (Bricka et al. 2009). The study did note that these costs were expected to decrease as the data collection process became streamlined and as new technology became available.

An additional limitation of GPS stems from the need of line of sight for the technology to function properly. Obstructions, which include tall buildings and trees, between GPS devices and satellites may limit the ability of data capture from the device

(Bhat 2014). Downtown areas may also have data losses for routes and segments due to the limitation of access to GPS antenna.

Cellphone

By the end of 2012, over 326.4 million wireless active devices, including smartphones, tablets, and hotspots, existed within the US (U.S. Wireless Quick Facts 2013) and by the middle of 2013, 91% of adults age 18 and older owned a cellular phone in the US (Brenner 2013). These statistics along with the availability of wireless location technologies (WLT) from wireless carriers have inspired transportation researchers to investigate the feasibility of extracting traffic data from the location data of cellphones.

There are two categories of WLT: mobile based and network based. For mobile based WLT, the location is determined from signals received from base stations or from GPS; while for network based, an existing network is relied upon to determine the location by measuring signal parameters at the base station (Sayed, Tarighat, and Khajehnouri 2005). The E911 mandate from the Federal Communication Commission (FCC) requires that all cellular carriers be able to provide a 911 caller's phone number for return calls as well as the location of said caller via WTL (Revision 1997). This mandate led to Yim's 2003 examination into cellular probe technologies where it was noted that the use of E911 for probe activities introduced privacy concerns and improvements in cellular geolocation technologies would be needed to realize the full capabilities of the technology. In 2006, Pan et al. demonstrated the theoretical and experimental feasibility of using cellular-based data-extracting methods for trip distribution showing the methods ability to directly attain traveler spatiotemporal information from mobile carriers was advantageous as it required minimal labor and costs. Correspondingly, the 2007 effort of Caceres, Wideberg, and Benitez developed a

technique that used the global system for mobile communications (GSM) for the derivation of OD data via simulated data which produced estimation results of reasonable precision. However, a limitation of the technique was the need for the cellphone to be powered on for data collection to occur. The simulated effects of using WLT for monitoring traffic was examined by Fontaine and Smith (2007) which found overestimation in the capabilities of the system for dense networks with mixed congestion and free flow conditions. The study also noted the need for WLT data collection to be tailored to specific localized parameters including frequency rates for sampling, which may need to be adjusted to account for traffic conditions.

In 2010, a method developed by Schlaich, Otterstätter, and Fiedrich to generate time-space trajectories for travelers through the analysis of cellular phone data from location-area-updates, which are recorded from mobile phones while in the standby-mode, found that while trajectories were able to be produced they were only representative of SIM-cards and not vehicles, and that short trips could not be detected since they may exist within one location area only. This first limitation is of particular importance since one vehicle may contain multiple SIM-cards onboard. Concurrently, a study by Herrera et al. (2010) used GPS-enabled mobile phones for traffic monitoring and served as a proof-of-concept of the proposed methodology. This effort found that higher accuracies for velocities were attained in comparison to loop detectors despite penetration rates of two to three percent. These minimal penetration rates were found to be sufficient to achieve spatiotemporal coverage of the network since the devices would be moving throughout the transportation system, thus making the data collection method viable for transportation planning purposes.

Further research using cellphone data to create OD trips by purpose and time of day has recently been undertaken. The work by Çolak et al. (2015) explored the use of

cellphone data within a four-step model with a focus on opportunities and limitations of the data. The effort used only cellphone data and population density for two large cities, one in the US and one international, finding the extent the cellphone data was able to accurately reflect daily travel and proposing guidance on how to utilize the data for OD generation by purpose and time of day. While the effort was successful, the authors only accounted for traditional work-life schedules (i.e., working 8 a.m. to 5 p.m., no shift work) and did not attempt to account for work locations where cellphone use is limited or prohibited (i.e., hospitals, schools). Additionally, the spatial resolution analyzed within the study only had success at town and subdistrict level indicating a considerable limitation of the current methodology. The authors also note the significant existence of mismatched data between the comparison trip data with the estimated data and concede that while the method holds promise, the data is not representative of the population at the TAZ level traditional needed by metropolitan organizations for planning purposes.

Recent evolutions in cellphone probe data had been attributed to upgrades in wireless communication standards into 3G and 4G, the market domination of smartphones, as well as the integration of social media and cloud computing. Companies like AirSage have teamed with cellular companies to receive wireless signal data which are then used to anonymously determine location (AirSage 2013). This time- and date-stamped aggregated location data can be used to model, evaluate, and analyze the movements of commuters for almost every city in the US. Despite the good spatiotemporal coverage, the data cost may make it cost prohibitive for usage by many municipalities. Furthermore, there is a lack of trip purpose information for this data type.

Additional limitations with the technology include the battery life and GPS sensors within devices. When GPS is used with a fully charged battery the expected battery life is no more than three hours (Bhat 2014), which significantly impacts data

collection. The GPS sensors used within cellphones are typically low cost and are likely to have failures. Incomplete data from cellphone GPS data collection has been attributed a loss of cellphone's GPS sensor and satellite due to being indoors, and from software start-up time often at the beginning of trips (Rasouli 2014). Miss allocation of position data due to the cellphone's GPS positioning itself in an incorrect location and then correcting the location leading to paths that do not make sense (Stopher and Speisser 2011). As mentioned above, privacy is a concern with respect to this technology. Link et al. (2014) noted that data collected via smartphones could be used for unintended purposes such as a geotagged photo including individuals that did not give consent for a study. Finally, the use of data collected via smartphones can introduce selection biases specifically when an application is used that must be downloaded from one of the application stores (Bhat 2014).

Bluetooth

Capitalizing on short-range personal wireless connectivity technologies, Bluetooth allows personal devices to have direct communication with each other without the need for line of sight, which is required for radio frequency based connectivity (Bisdikian 2001). Developed in 1998, Bluetooth has been noted as a low cost, user friendly method for the collection of data (Blogg et al. 2010, Brennan et al. 2010, Hainen et al. 2011). Tracking is done by unique media access control (MAC), a 48 bit, 12 alphanumeric character address assigned to the device by its manufacturer, for each device eliminating privacy concerns as the MACs are not affiliated with the users.

Within the last decade researchers have begun examining how this technology can be utilized of OD matrix creation. Bluetooth was found to be effective in the collection of OD data in small controlled networks in Blogg et al.'s 2010 study, which demonstrated

the favorable comparison of the data to video and automated number plate recognition data. These results were substantiated by Hainen et al. (2011) in a study that compared the technology with license plate matching.

Limitations were found for this technology, which included the need for appropriate detector placement, short ping cycles of approximately 0.1 seconds potentially leading to a single device being detected multiple times as it passes a single detector, and the potential for multiple MAC addresses existing within a single vehicle similar to the issues of WLT (Blogg et al. 2010, Yucel et al. 2013). Brennan et al. (2010) examined concerns with detector placement and noted the lack of existing design guidelines for placement and the existing variation in placement locations in both height and distance from the monitored facility led to large variances in the number of captured addresses. The effort noted between 5 and 10% of the vehicle population had discoverable MAC addresses that were able to be collected and that no relationship between traffic volume and collection efficiency existed, rather the height of the detector influenced the collection efficiency. Friesen and McLeod (2014) noted sensor deployment for appropriate coverage in urban areas was a particular challenge that needs to be addressed as the technology's estimation capability is dependent on a high penetration of devices thus increasing the reliability of the data attained.

Exploration into dynamic OD matrix estimation for freeways using Bluetooth was done by Barceó et al. (2010). Noting the variability of the sample collected yielded objectionable expansion errors, the authors indicated the use of the technology independent of other methods was too risky.

Future Methods

A study by Bricka (2013) indicated that smartphone data sources were resources for potential travel survey data. Recent research has explored the opportunities to use non-traditional data sources such as smartphones (INRIX 2010, AirSage 2014), smart cards (Pellietier, Trepanier, and Morency 2011) and vehicle to infrastructure (V2I) communications (Tornero, Martínez, and Castelló 2012).

Social Media

Social media has been examined as a data source for transportation planning recently. Details specific on the use of location-based social networking for planning will be discussed in a subsequent section of this chapter.

In 2008, Molin, Arentze, and Timmermans examined social relations and the trips made to maintain these relations using data collected on 1980's ego-centric social networks in the Netherlands. The authors' found that socio-demographic attributes have only a modest influence on the size of a participant's network, yet had a larger impact on the travel time and frequency of contact. A follow up study was performed by van den Berg, Arentze, and Timmermans (2009) that examined ego-centric social networks using a data set from 2008. This effort used regression models to explore and predict the size and distribution of the network across social categories, geographic distances and contact frequency finding the relationships to significant but not strong.

A similar effort was recently conducted by Toole et al. (2015). This work examined massive, passively collected data from communication technologies with geographic information to discovery individual visitation patterns and compare them to existing social connections and strangers to determine predictability. The author's found that the contact composition of a user's ego network correlated with mobility behavior and suggested geography as an important feature for contextualizing social relationships.

A probabilistic model was defined in the work by Alesiani, Gkiotsalitis, and Baldessari (2013) which used social networking data to describe activity patterns. Recently, Misra et al. (2014) explored the use of crowdsourcing as a method for involving participants in the transportation planning process and noted numerous successful efforts within the US. Grant-Muller et al. (2014) examined how social media data can be used in conjunction with or separate from current data sources for transportation purposes. The study focused on social media text content, which was noted to possibly suffer from coverage limitations attributed to a user's ability to choose to contribute content. The authors noted the complexity of sentiment analysis needed for this data sources viability.

Smart Cards

Smart cards are portable, plastic cards with built-in technology that can contain financial, personal, and transactionary data. The use of this type of technology deployment is seen predominately in Europe and Asia, but in recent years has begun to be implemented within North America (Pelletier, Trépanier, and Morency 2011). This work included a review of smart card data uses within the public transportation sector, noting the data's advantage in the reconstruction of user trips to examine travel behaviors. The effort also indicated concerns about the use of the data including, but not limited to, privacy concerns, the lack of confirmed trip purpose, limited knowledge of ultimate destinations, as well as market penetration rates.

With increased deployment, smart cards have become a detailed dynamic data source for public transit agencies. These uses include turnover analysis (Bagchi and White 2005), typical user type and trip habits (Agard, Morency, and Trépanier 2006), creation of future demand matrices (Park and Kim 2008), and comparison of data with

household survey datasets (Trépanier, Morency, and Agard 2009). More recently, Devillaine, Munizaga, and Trepanier (2012) presented a method for detecting and estimating transit users' location, time, duration, and purpose of activates using smart card data as well as land use and travel behavior information. The authors noted the abundance of the smart card dataset allowed for many detailed explorations of time-space travel, origin-destination matrices creation as well as user behaviors insights. However, the authors noted that the dataset suffered from a lack of socio-demographic information and its limitation to public transit movements. Similarly, the study by Munizaga and Palma (2012) explored a methodology for origin-destination matrix creation using smart card and GPS data for a large scale multimodal public transportation system. This study was able to accurately predict the alighting point for 80% of the boarding transactions included and obtain origin-destination matrices from the data; however, it was noted that the attained matrices were not the same as those obtained from the traditional origin-destination surveys. Recently, smart card transactions were explored for use in discovering and partially correcting travel survey bias (Spurr, Chapleau, and Piché 2014). This work indicated the existence of overestimated subway boardings during peak periods within the travel survey data and demonstrated the ability to adjust the weights of particular trip types to match the entry volumes at subway stations.

Connected Vehicle

With the anticipation of pilot deployments in 2015 (CV Pilots Deployment 2014) connected vehicle technologies have been explored in recent years for their potential contribution to transportation planning. In 2012 Tornero, Martinez, and Castello examined the potential of vehicle-to-infrastructure (V2I) communication technologies to create OD matrices. This study indicated that the use of dedicated short-range

communication (DSRC) for the connection of vehicles via on-board units to infrastructure via roadside units potentially would collect data from every vehicle connected to the system and would therefore be able to create an accurate, instantaneous, and dynamic OD matrix in rea-time. This ability could effectively eliminate the need for OD estimation. However, privacy concerns were noted with the use of this data collection method and any realization of the data source is still years away.

SOCIAL NETWORKING

A social network is defined as a structure made up of a set of individuals or organizations and the interaction ties between these individuals or organizations within the study of sociology (Wasserman 1994). Traditionally web-based, social networking sites build up these interaction ties through constructing networks or relationships among individuals with similar interests, activities, backgrounds, and various other types of connections. The most popular social networking sites currently include Facebook©, Twitter©, and LinkedIn® (eBizMBA DATE) with Facebook© ranked as the number two site for web traffic both globally and for the US (Alexa 2015) having over a billion active users (Company Info 2015).

Research efforts have examined social networking within various areas of transportation. Social media for marketing approaches for public transportation was explored by Morris, Robertson, and Spinks (2009). The authors examined the use of blogs, podcasts, social networking sites such as MySpace, Facebook, and Twitter, web photos and videos, wikis, virtual worlds, and Google noting how each was being used or could be used for public transportation outreach. An effort by Kaufman and Moss (2014) suggested that the co-monitoring of social media sites by transit agencies could lead to a better understanding of user opinions.

Mobile navigation systems have been researched in Europe with respect to social media. Huang and Gartner (2012) examined the mobile pedestrian navigation systems recommending an intelligence-based routing method to address the lack of social navigation support found in current systems. The study used Vienna, Austria as the base network for the evaluation of the methodologies with results demonstrating that the collective intelligence-based routes had significant improvements on route quality compared to those with lower complexity. The 2012 effort by Fiorentino et al. examined the intelligent transport system for optimized urban trips or i-Tour project in Naples. The authors note that real time information is provided by transportation operators creating a traveler information system composed of personalized location based services that support the user community with details on traffic events, tourist information, points of interests, and recommendations. In recent years, social networks have been used to understand how personal attitudes and information diffusion with respect to activity and travel behavior choices have been explored (Chen, Frie, and Mahmassani 2014). This conceptual work noted the importance of social networks as a source of information for travel behavior researchers to gain better insight into the future travel behaviors. In 2015, this effort was continued by Chen, Talebpour, and Mahmassani using agent-based modeling to probe social influences on route choice in an effort to develop a practical tool for encompassing travel behavior patterns not addressed by traditional methods. The authors noted the insight offered by the method did have limitations, specifically the assumption that all drivers within the system share their opinions about route choice via social media daily.

Social networking has been used for conducting surveys for transportation purposes. The study by Efthymiou and Antoniou (2012) used Facebook for survey dissemination to elicit opinions on carsharing, bikesharing, electric vehicles, and travel

patterns for young Greeks. The authors noted that the medium did have associated sample bias with the majority of the participants between the age of 18 and 35. In addition to Facebook, the authors examined Twitter feeds for key words related to the study. In the United Kingdom (UK) a specially designed website, Cycology, was used to explore social processes of commuting cyclists (Bartle, Avineri, and Chatterjee 2013). This effort found that the sharing of information provided social and functional roles; however, the authors noted that the non-naturally occurring online community and relatively small sample size limit the value of the effort. While all of the previously discussed data collection methods have pros and cons, one the most notable con is the data cost. Data collection through social networking can range from free to a few thousand dollars, while more conventional methods range from tens of thousands of dollars to hundreds of thousands of dollars.

Location-Based Social Networking

As a specific subset of social networking, location-based social networking (LBSN) falls under the umbrella of location-based services (LBS). LBS, which use location and time data, has four distinct areas of concentration: maps/navigation, tracking, information, and applications. Figure 2 shows a depiction of this breakdown. Under the maps/navigation group, common applications include Waze, Google Maps, and Metr0. Waze is a navigation application that allows drivers to connect with one another to share real-time traffic and road information such as existing incidents and police traps (Waze 2014). Similar to Waze, Google Maps provides users with navigation for travel via vehicle, transit, or by walking, and offers satellite imagery and street views (Google Maps 2014). Metr0 provides public transportation guidance for over 400 cities in 75 different countries worldwide and provides users information on places of interest for

select cities (Metr0 2014). Common tracking applications include Find My Friends, Life360, and Glympse. Find My Friends is an application that allows users to keep track of family and friends by allowing users to share their location with other users (Find My Friends 2014). Similarly, Life360 allow users to view users within their private circle on a map and has a feature that allows for group messaging (Life360 2014). Glympse is another application that allows users to share locations and to let friends know anticipated arrival times (Glympse 2014). The information based category contains applications that can assist with local searches as well as city guides; Yelp, WeatherPro, and Zillow are commonly used applications. Yelp is an application with the purpose of helping users find local businesses based on reviews of services and allows for businesses to directly address concerns of customers (Yelp 2014). Zillow (2014) has a similar objective but focuses on the real estate market by providing information to consumers about homes for sale and rent in addition to estimated market values and neighborhood information like walkability and transit access. WeatherPro is Europe's leading paid-for weather application and provides high-quality weather forecasts for over two million locations around the world (WeatherPro 2014).

The final category is applications, which contains social networking sites and context advertising and includes the popular Foursquare, Facebook, Twitter, and AdMob. Facebook is used by billions of people around the world to connect with friends, family, and the world through sharing and expressing what matters to them (Facebook 2014). Like Facebook, Twitter (2014) connects friends and other individuals by allowing users to send and read short 140-character messages or "tweets." Google's AdMob is an application that provides application developers a way to monetize, promote, and analyze their applications through the placement of relevant advertisements within the user's application (AdMob 2014). LBSN falls within this last category of LBS combining it

with social networking sites and will be the focus of this dissertation. LBSN services include geo-tagged-media-based, point-location-driven, and trajectory-centric locations. Geo-tagged-media-based services are media focused and include applications like Twitter, while point-location-driven services focus on point locations providing instant real-time information and include applications like Foursquare. A general discussion about Foursquare will be provided within this chapter with a more detailed discussion to be found within subsequent chapters of this dissertation. The trajectory-centric services are focused on trajectories providing rich data and include sites like Waze.

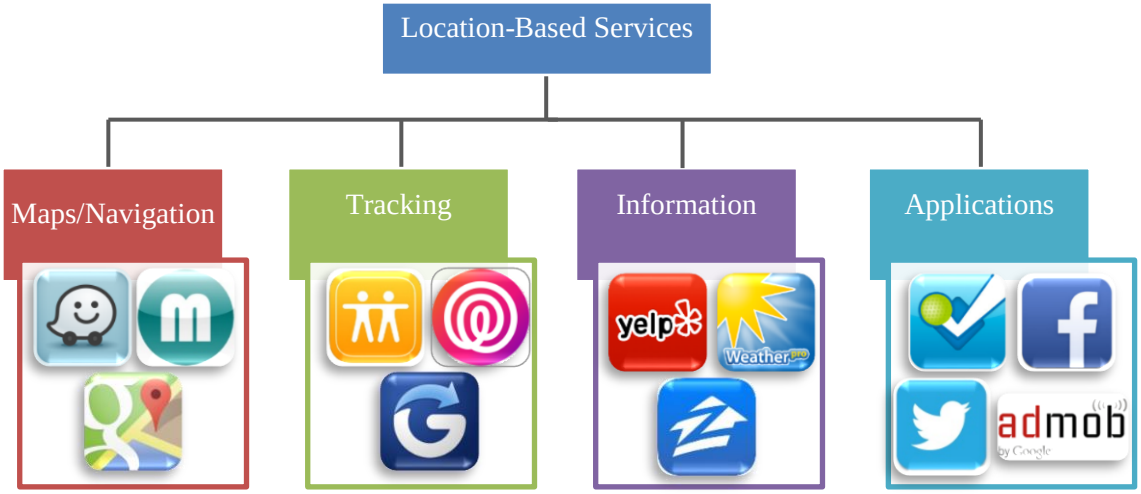


Figure 2.2: Location-Based Services Categories with Application Icons.

As tablets and smartphones are owned and used by more of the population, this LBSN data source has becomes a more population representative data source. Thus, researchers have begun to mine it to better understand user's spatial patterns, geographic movements, temporal dynamics, networking ties, and location predictions. The first efforts to explore spatial patterns of users of LBSN used Markov-based location predictors to determine future locations of users (Li and Chen 2009). This study used

Brightkite, a site that allowed users to share their locations, post notes, and upload photos, and employed a Markov-based location predictor to determine future locations of users with moderate success.

Further studies of user location prediction based on a user's friends (Backstrom, Sun, and Marlow 2010) and a user's content (Cheng, Caverlee, and Lee 2010) were examined in subsequent years and proved valuable. The effort of Backstrom, Sun, and Marlow showed that a user's Facebook friend network could be used to predict the user's location within 25 miles for over 69% of the users with 16 or more friends. Contemporarily, Cheng, Caverlee, and Lee examined Twitter data for the estimation of a user's city-level location based solely upon the content of the user's tweets with a success rate of placement of 51% of the users within 100 miles of their actual location. Following these efforts, research explored LBSN's data relationships between geographic movements (Cho, Myers, and Lseskovec 2011), human movement's temporal dynamics (Zheng, Xie, and Ma 2010), as well as the links of social networking (Karimi 2010). In Cho, Myers, and Lesckovec's work the relationship between geographic movements and social networking ties was explored using Brightkite and Gowalla, a site that allowed users to check-in to their current locations. The results indicated that while short-ranged travel was periodic spatially and temporally in nature, long-distance travel was influenced by social networking ties and social relationships could explain between 10 and 30% of human movements. The studies by Zheng, Xie, and Ma and by Karimi explored two different data sources, GeoLife and Genetic Location-Based Social Networks (G-LBSM), respectively.

To assist in the identification of dangerous intersections in Israel, Fire et al. (2012) examined Waze data to identify locations with reoccurring incidents and where there locations with high police presence without reported incidents. The results of the

analysis revealed areas where there were a large number of incidents but not enough police coverage or the reverse situation; also, it was estimated that almost 68% of the incidents did not have police intervention and the average response time to an incident was just under 29 minutes. Facebook was used in the 2014 study by Wall, Macfarlane, and Watkins to characterize individual travel behavior based on size and distribution of online social networks examining whether individuals traveled further via airline travel to reach destinations where they had friends as compared to those where they did not have friends. The authors showed that this assumption was correct; however, the sample had a significant bias toward students who traditionally have limited discretionary funds for travel and for out-of-state students, travel is likely to be to their hometowns for holidays and the summer semester.

Recently, much research has been conducted using Twitter. Wang and Taylor (2014) proposed a method for the collection and analysis of data from Twitter creating a process map for the collecting data on human mobility in New York City. Twitter posts were examined by Doran, Gokhale, and Konduri (2014) for the Metropolitan Transit Authority (MTA) in New York City to be used to improve service. The authors noted that the while there was promise there were a number of challenges including data quality concerns since the data suffers from sparsity as individuals are less willing to share unless they are pursuing trips on the network and the existence of false-positive reports. Other concerns noted data fusion issues that indicate a need to develop data fusion and mining techniques that can synthesize the information when there are quality issues, and participant availability, specifically recruiting participants and ensuring continued involvement. Gal-Tzur et al. (2014) used Twitter data from Liverpool, UK analyzing content characteristics with specific attention paid to transportation posts. The results of this study used syntax analysis that included 500 terms, showed that valuable information

for policy makers did exist within the media, and this information could be effectively harvested although done so with much difficulty. Similar to this effort, sentiment analysis was used to analyze Twitter data for the Chicago Transit Authority (CTA) with focus on the "L" system. This study by Collins, Hasan, and Ukkusuri (2013) noted the benefits of cost effectiveness, the ability to collect the data in real-time, and the meaningful insight provided by using the data source. However, the analysis did note that riders were more likely to assert a negative sentiment compared to positive ones and that the sample contained biases toward certain "L" lines, may have excluded the large Latino and Polish communities with the collection of only English based tweets, and the limited access of lower income populations to the technology that is needed for Twitter. A study of sentiment analysis via Weibo, a Chinese application similar to Twitter, also noted barriers to the analysis. The study noted the challenges the Chinese character based language presented and that the rule-based approached had difficulty identifying ironic statements (Cao et al. 2014). The work of Hasan and Ukkusuri (2014) examined the use of New York City Foursquare check-in data via Twitter to develop a methodology to understand activity patterns of individuals through the use of machine learning techniques. The authors use an activity pattern model, which was found to have promising results. A limitation noted for the effort was the lack of report of certain activities that are participated in at the home or work place.

The most popular LBSN site is Foursquare, which has over 30 million users and over three billion check-ins (Media 2013), has users that include business and individuals. Researchers have recently begun to explore this sites data set due to its popularity, high penetration rate, and large sample size, specifically to explore mobility patterns across spatial, temporal, and social aspects among users (Cheng et al. 2011, Scellato et al. 2011).

Additionally, the site's data has been explored more recently for its transportation planning application. The 2011 study by F. Yang et al. (2014) specifically used Foursquare data to estimate an origin-destination matrix for the Chicago urban area, and was among the first to demonstrate the data's potential for use in transportation planning. Continuing this effort, the modeling technique was applied to the Austin, TX area using a singly-constrained gravity model (Jin et al. 2013). Goers (2013) used Foursquare data to investigate its potential to inform planning and redevelopment decisions for the Tampa (Florida) Planning Division finding that the data source revealed morning work related check-ins and the ebb and flow of activities at restaurants. The data also showed a lack of check-in data for neighborhoods with older and poorer populations, demonstrating a bias within the source. To further analyze the use of Foursquare data for transportation planning, a doubly-constrained gravity model was proposed for the Austin area and demonstrated better learning capabilities when compared to the singly-constrained gravity model (Jin et al. 2014). Finally, exploration of the data with respect to mode choice revealed the data set's potential to provide information on select modes (airplane, bus, rail, and bicycle), but could not provide any insight on the walk or automobile modes (Cebelak 2014).

Further exploration of Foursquare for OD estimation was performed by SA et al. (2015). This study compared OD estimation from Foursquare, using the method proposed by Jin et al. 2013, and from cell phone data to an existing OD matrix that was constructed from travel surveys confirming the ability of the Foursquare data to correlate to the existing OD estimation. The authors also indicated Foursquare data was better than the cell phone data at OD estimation. Research has also been conducted to relate Foursquare data and land use data for automated travel activity inferring. Abdulazim et al. (2015) used data from the Greater Toronto and Hamilton Area to develop adaptive algorithms to

estimate activity distributions from land use with the goal of addressing survey burdens from the collection of long-term personal travel diaries. The results of the machine learning classifiers used in this study revealed that trip distance and time had a more predictive power than land use.

MANY-TO-MANY MODELING

Within relational database theory, the "many-to-many" connections concept refers to relationships between a "parent" or entity row and several "child" or characteristic rows as well as the relationship between a "child" row and several "parent" rows, and has been described as a "mirror of the real-life relationship between the objects" (Janssen 2014). The concept has been applied to a variety of disciplines including business, marketing, technology based industry, as well as anthropology. The movements of individuals can be influenced by business marketing and social anthropology, and thus can be analyzed under the spectrum of many-to-many connections. There are three many-to-many modeling structures that focus in these areas: business-to-customer, social forces, and peer-to-peer. Brief descriptions of business-to-customer and social forces efforts are given in the following sections, with the majority of the effort below given to the peer-to-peer focus of this dissertation.

Business-to-Customer

Business-to-customer (B2C), where the "parent" role is filled by businesses and the "child" role is filled by customers, has been researched with respect to the transportation field; one of the earlier examples was in 2001. This effort (TRIP 2001) examined the relationship between the economy and the transportation system within the US, with respect to freight movement, noting the higher levels and greater reliability for freight transport for business-to-business and business-to-customer exchanges. After this

initial effort, two additional research efforts were done in the early part of this century. Both used B2C to further analyze trends within the freight industry, specifically the parcel component of the industry (Pagano 2001, Rabah and Mahmassani 2002). Research in this area had a brief hiatus until 2006, after which a handful of related research was conducted examining the relationships between suppliers and customers with respect to logistic services (Davis and Mentzer 2006, Leinbach 2007, Park, Min, and Park 2011), aviation (Franke 2007), and personal vehicles (Aboltins and Rivza 2014).

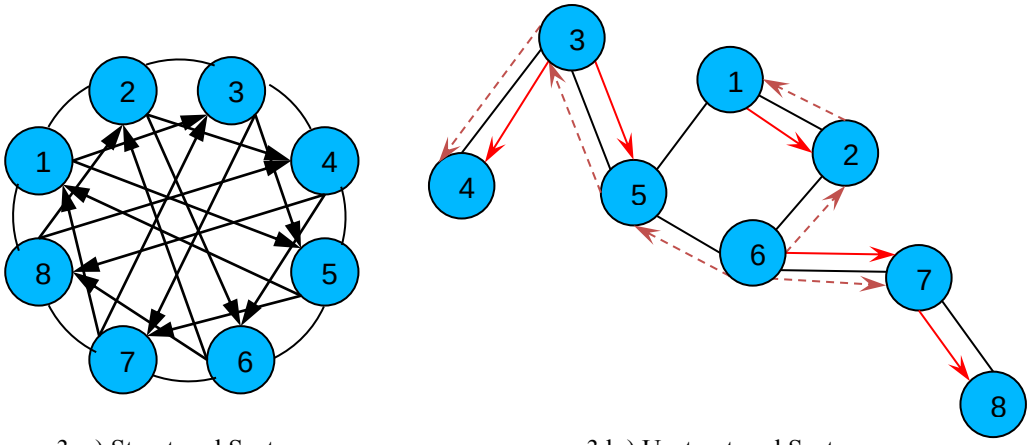
Social Forces

With regard to the transportation industry, the research within the social forces genera of many-to-many connections in recent years focuses primarily on pedestrian interactions. In 1991, Helbing proposed a mathematical model to describe the movement of pedestrians, which became the basis for his later effort that related behavioral changes in pedestrians to social forces, which were defined as external influences or the environment, public opinion, and social norms and trends (Helbing 1994). This effort lead to studies in crowd dynamics (Helbing et al. 2005), bottleneck flow for pedestrians (Kretz, Hengst, and Vortisch 2008), prediction and simulation of pedestrian movements (Rudloff et al. 2011, Deroo and Auberlet 2012, Duives, Daamen, and Hoogendoorn 2013), and pedestrian route choice (Werberich et al. 2014).

Peer-to-Peer

According to Amad et al. (2012), peer-to-peer (P2P) modeling has attracted interest in recent years due to the ability to support today's internet applications and the characteristics of scalability, fault tolerance, and robustness making it well adapted for social networks. P2P network modeling consists of unstructured and structured systems. Unstructured systems generally are based on a global index or use a flood algorithm to

locate and discover peers, while structured systems are based on the Distributed Hash Table concepts where each entity name in the system can be mapped into a single search space (i.e., ring topology, hierarchical rings) using hash functions and all entities within the system have a consistent view of that mapping. Figure 3 shows how these systems are structured. Examples of unstructured and structured systems include the peer-to-peer file audio file sharing Napster (Napster 2014) and the file sharing application BitTorrent (Xu et al. 2013), respectively.



3.a.) Structured System 3.b.) Unstructured System

Figure 2.3: P2P Network Structures.

Internet and Computer Network

P2P is most commonly used for internet and computer networks. Beginning in the early 2000s, efforts explored wireless ad-hoc networks using directional antennas (Yi, Pei, and Kalyanaraman 2003, Cain et al. 2003). In 2006, Popa et al. presented an effort to reduce congestion in wireless networks via multipath routing using mechanisms on non-greedy paths to improve energy efficiency. This effort and the efforts of 2003 were used by Medina, Hoffmann, and Ayaz (2008) to examine the topological properties of aeronautical ad hoc networks between aircrafts and ground stations where the spatial

density of nodes (aircrafts) are significantly higher. This effort explored the use of greedy forwarding, a concept where information packets selected locally optimal or greedy choices in choosing the next node to hop to which is the closest neighboring node geographically, and the use of omnidirectional antenna. Medina et al. (2010) continued this effort examining an airborne mesh network via direct air-to-air radio links in an effort to extend the coverage of broadband air-to-ground infrastructure networks (i.e., aircraft to ground station). Each node within the proposed network potentially hosts several hundred bandwidth-demanding users (i.e., passengers) and the use of multipath routing algorithms can improve network performance under these conditions. The authors propose a geographic load share routing (GLSR) algorithm that exploits path diversity thus, mitigating congestion within the multi-hop wireless network with directional antennas. GSLR extends the greedy routing algorithm by take advantage of the multiplicity of OD paths in multi-hop networks with the choice of neighbor simultaneously maximizing advancement toward the destination while minimizing queueing delay. .

With respect to computer networks, Gradowski, Mrowinski, and Kosinski (2010) presented a P2P network configuration where files are exchanged directly without the use of a central server. Work done by Sukjit (2011) presented a novel algorithm for generating logical layers of hole-free, non-overlapping rectangular grids to support the data processing needs of oriented connectivity from P2P networks. In the same year, Neumayer, Doulkeridis, and Norvag examined unstructured P2P computer networks presenting a hybrid approach that used a hierarchical overlay network for document retrieval. This hierarchical overlay is an aggregation of selected terms with high-frequency values and is combined with a gossip-based aggregation of remaining low-frequency terms. To provide fault tolerance and increased availability within systems,

Barshan, Fathy, and Yousefi (2012) proposed a 3-tier hierarchical architecture based on P2P modeling for network management. Within their architecture, peers used in several roles give the network its increased availability. In support of recent movements toward cloud computing, a proposed cloud-based parallel data processing system, MapReduce, was described in the 2012 effort of Morozzo, Talia, and Trunfio. A 4-dimentional model to collect information about and obtain a peer status description was proposed by Mirtaheri et al. (2014) for both server and P2P server-less architectures. Most recently Xu et al. (2013) proposed a methodology for using P2P modeling to determine computer network traffic matrices, which served as the inspiration for this dissertation. More on this effort will be presented in the methodology chapter of this dissertation.

Transportation Endeavors

Within the peer-to-peer modeling dynamic, transportation research trends are similar to those seen in the social forces modeling. Where social forces research has focused mainly on pedestrians, peer-to-peer transportation modeling has concentrated in supply chain systems, the carsharing spectrum, the connected vehicle, and minimally into transit and human interactions with respect to safety. A 2011 study by Min looked at the P2P services of paratransit, which included the door-to-door and curb-to-curb services offered by the Massachusetts Bay Transport Authority in the Greater Boston area. Min examined the number of requested and cancelled trips, completed rides, on time completed trips in his effort. With respect to the human interaction P2P efforts, an anonymous P2P observation-feedback was used in the Clear Signal for Action (CSA) project, which is a proactive safety risk management method (Ranney et al. 2007). This project examined the ability of CSA to improve safety within the rail industry when operators function under constraining signals. Where the efforts by Ranney et al. focused

on rail safety, efforts by Winston and Jacobsohn (2009) and Tisdale (2013) used P2P to promote safe driving for teens and college aged individuals, respectively. Supply chain, carsharing, and connected vehicle efforts will be further described below.

Supply Chain

In 2003 a project sponsored by the Office of the Secretary of Defense in conjunction with the Department of Defense was undertaken by Boyson, Corsi, and Verbraeck to identify the characteristics and demonstrate the effectiveness of a portal-based architecture for management of defense supply chains. The authors noted that the exchange of information between companies was often hindered by the varying systems used, the number of peer-to-peer relationships with other companies, and the lack of openness of existing systems for the exchange of information. The resulting e-supply chain portal provided all parties involved the necessary tools for effective management and reduced significantly the response times for critical events. Gumzej and GajAiek (2011) investigated the quality of service provided by suppliers to customers within a supply chain and assessed the impacts of the measurements of quality of service which provided identification of weak spots within the supply chain. This effort spurred the 2012 effort of Gumzej, Sukjit, and Unger which considered overlay networks of P2P systems for data interchange in the global e-marketplace. The proposed overlay network allowed for searching and routing of information via a coordinate space to participating peers using a novel decentralized structure-building algorithm.

Carsharing

Research into peer-to-peer carsharing has been conducted since 2011. Peer-to-peer carsharing allows individual car owners the ability to convert their personal vehicles

into vehicles that can be rented for short-terms by other drivers. This early effort examined the potential renter demand for P2P carsharing based on the current methods use for assessing demand for traditional carsharing (Hampshire and Gaites 2011). One of the main differences the authors noted was the need to estimate the number of parked cars for the P2P carsharing effort. In 2012, Shaheen, Mallery, and Kingsley explored P2P carsharing with respect to business models, marketing opportunities, and barriers of services to determine the mode viability with respect to sustainable transportation. This effort was followed by the Dill, Howland, and McNeil (2013) examination using participants from Portland, Oregon. Participants had the number of trips taken and length of trip calculated via GPS and ignition data; this data was used to examine block groups that contained one or more P2P vehicles, traditional business-to-consumer (B2C) carsharing vehicles (i.e., Zipcar), and one-way pay-per-minute vehicles (i.e., Car2Go) finding that the P2P model yielded vehicles that served a greater number of block groups and a larger percentage of families in poor, non-white as well as foreign-born population compared to the other models. The authors found that the P2P models had the benefit of a lesser concentration of vehicles that allowed for larger geographic coverage and thus more potential users. Also in Portland, Chen, McNeil, and Dill (2014) conducted a P2P carsharing participation survey concerning Getaround, a P2P carsharing program within the area. The survey results indicated that renters tended to be a more heterogeneous group, with full-time employment and limited travel options (i.e., low numbers of house hold members with transit passes) and that schedule flexibility was an important factor for vehicle owner participation.

In the San Francisco Bay area, an intercept survey was conducted by Ballús-Armet et al. (2013) to determine the existing attitudes toward traditional carsharing, peer-to-peer carsharing, and the sharing economy, an economic model based on sharing assets

among groups of people rather than owning. The results indicated a low awareness of P2P carsharing which was most notable among participants without private vehicle access and that there was openness toward P2P carsharing and the sharing economy with many individuals agreeing that P2P seemed to be a convenient, affordable, and innovative mobility approach. A study by Rivasplata et al. (2013) explored the relationship between P2P carsharing and off-street parking in the San Francisco Bay area. The study found that each of the 441 site within the study provided 2.8 carsharing spaces with approximately half located in parking lots or garages. A global perspective study of P2P carsharing was conducted by Shaheen and Cohen (2013) which included 26 nations and their future carsharing developments from 2006 through 2015.

Connected Vehicle

As mentioned above, the anticipation of connected vehicle deployments has influenced research in recent years for the transportation community. However, the P2P community has been researching these vehicles since the turn of this century. The 2000 effort by Breisemeister, Schafers, and Hommel developed an approach for distributing warnings about hazards in road traffic for vehicle to vehicle communications through omni-directional antennas that allowed a sender to simultaneously transmit to multiple hosts. Bogenberger and Kosch (2002) presented a test bed and software for ad-hoc P2P communications for vehicles. Füßler et al. (2002) also examined ad-hoc networks for vehicles by comparing routing strategies for highway traffic. The authors note the advantages of ad-hoc routing are significant when communications spans more than two or three hops. If knowledge of geographic position of network nodes, better performance through geocast-routing algorithms can be achieved for the dissemination of information in multi-hop vehicle-to-vehicle (V2V) networks according to Kosch, Schwingenschlogl,

and Ai (2002). A decentralized vehicle-based traffic information system that would eliminate public investment was proposed by Ziliaskopoulos and Zhang (2003) for a freeway with various levels of market penetration and congestion. In 2004, Festag et al. examined an effort to a leverage V2V communication platform and developed suitable communication concepts based on ad-hoc network and available position information. The authors noted that field trials yielded promising first results. Work by Caizzone et al. (2005) proposed an enhancement to the GPSR routing protocol to improve the performance of point-to-point IP-based voice communications within a vehicular ad-hoc network that yields significant reductions for the end-to-end delay resulting in network scalability and quality improvements were shown with this method.

Yang and Recker (2006) examined inter-vehicle communications (IVC) to analyze information-sharing between vehicles via P2P communication within the network. The effort models a self-organizing, situated traffic information system that was built upon V2V information exchange testing the pre-trip route-choice and in-trip re-route behaviors of drivers with access to traffic information from the proposed information system. Using average travel time to compare different groups of vehicles, results show that IVC-capable vehicles required less time to complete their simulated trips than vehicles restricted to following their initial paths. Explorations into the reliability of IVC for different penetration rates, transmission ranges, and traffic scenarios were examined by Jin and Recker (2006). Yang and Recker (2008) continued their previous work and presented an analysis of system performance of proposed self-organizing, distributed traffic information based on real-time V2V information-sharing architecture with information propagation through the simulated traffic network via IVC. The research effort has the goal of determining the IVC market penetration rate that is desired for information propagation to provide useful information pertaining to the entire

network. The results of the simulation show a dramatic increase in the percentage of vehicles that are able to use P2P information exchange for market penetration rates greater than about 10%, which results in significant travel time savings for vehicles. Lee, Jo, and Kum (2011) focused on social networking services in SMART highways through the use of vehicle-to-infrastructure (V2I) and the multi-hop V2V communication environment developing an architecture where all services have a defined function and content scenario. In 2012, Rivas and Guerrero-Zapata examined ad-hoc networks for the connected vehicle network with respect to points of interest using *poiSim* software which is capable of simulating a large number of nodes.

CONCLUDING STATEMENTS

The above literature discussion has explored the relevant existing transportation planning and modeling approaches which provides the basis for further discussion on the proposed many-to-many modeling methodology of this dissertation. Additionally, trends in traditional data collection and discussions on general social media as well as location-based social media provide a foundation that will be further explored within this dissertation.

Chapter 3: LBSN Dataset Analysis

The purpose of this dissertation effort is to further explore the use of location-based social networking data (LBSN) for transportation demand planning through the investigation of many-to-many connections modeling. The literature review that precedes this section demonstrated the interest in the use of LBSN data from the research community with respect to human mobility, as well as the variety of areas that many-to-many modeling has been applied to within transportation related areas. However, many-to-many models have not been used for transportation demand modeling nor have they been used with social media applications making this dissertation novel in its approach. This chapter will present additional information on LBSN.

LOCATION-BASED SOCIAL NETWORKING

The literature review covered the research that has been done concerning general, as well as location-based social networking data, demonstrating its relevance for this effort. In this chapter further details on the data source used will be provided including the advantages and limitations of the data source.

Foursquare Data

Foursquare is a smartphone and tablet application or app that is used by individuals to connect with the places they visit. This connection is done by checking-in, i.e., indicating the user is at a certain location, which is shared with the venue (aka the location) as well as friends that also uses the app. Founded in 2008 by Dennis Crowley and Naveen Selvaduai, the app officially launched in 2009 at the annual South by Southwest Interactive event in Austin, TX and used "badges" to encourage check-ins. A user could share in real-time locations that were visited and become "Mayor" of a venue by checking-in to the venue more often than anyone else within a 60 day period. The

ability to become “Mayor” provided an incentive for individuals to check-in to venues for which venue business could then offer rewards such as coupons. Business could also use the Foursquare platform for promotion of news, events, and discounts and the app could offer users insight on new locations and activities within their locations (Wikipedia 2015).

In May 2014, Foursquare recognized the competing nature of how users (business and individuals) were using the app. This led to the current version of Foursquare and its sister app Swarm (What Can We Help You With? 2015). Foursquare now is a platform where users can rate venues and find local places based on recommendations from the app. These recommendations come from user input on desirable venues, user ratings from similar places, and input from friends and experts that the user trusts the most. The social networking component of the original Foursquare has moved to the new app Swarm. Swarm allows users to share their location with friends as well as see which friends are close by to the user’s current location. Swarm still employees the “Mayor” competition, but has changed the method to allow for “Mayoring” to be done within friend groups. Additionally, users can become the “Mayor” of categories (i.e., “Mayor” of going to parks the most) as well as venues. Additionally, the app allows users to add a photo, leave a comment, add a friend, and/or add stickers to describe a mode or category with their check-in. Categorical stickers are based upon the surroundings of the users (i.e., coffee cup while at a coffee shop). To further incentivize check-ins, these stickers are attained by users based on real-world accomplishments based on how often a user is checking-in to a venue, checking-in at a particular venue, and commenting within a check-in. Stickers, comments, and the ability to add a friend to a check-in allows for more insight into a user’s check-in than was available in the past. Figure 3.1 and 3.2 show the new interfaces of Foursquare and Swarm, respectively. All future references to

Foursquare will refer to the previous version of Foursquare that housed the combination of current Foursquare and Swarm capabilities.

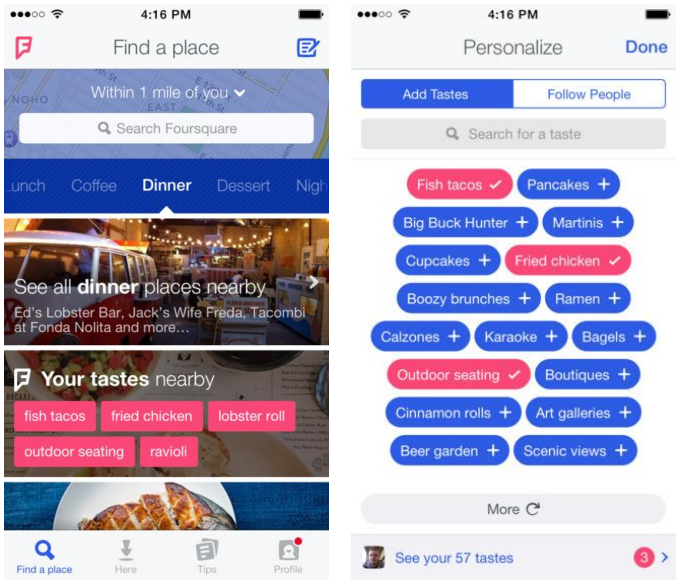


Figure 3.1: Foursquare Interface (Foursquare 8 2015).

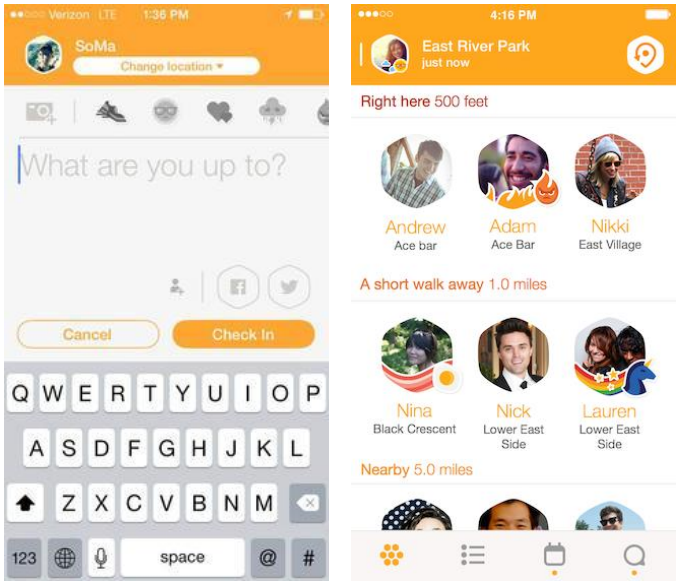


Figure 3.2: Swarm Interface (Swarm 2015).

According to the Foursquare website, there are over 55 million users worldwide that have contributed to the over seven billion check-ins (2015). Additionally, of the 65+ million venues throughout the world, more than two million businesses have claimed their venue locations. Claiming a venue allows businesses to access tools to update business information, create discounts or freebies for users when they check-in, and to attain visitor statistics.

Foursquare also has a developer component via its application programming interface (API) which can be used for Foursquare or Swarm applications. Currently there are over 85,000 developers that use Foursquare location data. Foursquare's API has four different functions:

1.) API Endpoints – provides a method for accessing a resource (i.e., venue, user), which can then be drilled into for information on an aspect (check-ins, likes, mayor-ships).
2.) Real-time API – provides a push service that provides real-time information from the user or venue perspective. The User Push API notifies when an authenticated user checks-in, while the Venue Push APO notifies venue managers when a certain action (i.e., check-ins, likes, tips) occurs at a venue.
3.) Venues Service – allows developers to search for places through the Venues Database to find information on tips, photos, and check-in counts.
4.) Merchant Platform – allows developers to create applications for venue owners to manage their presence on Foursquare via customer experiences.

Of these four functions, only the Venues Service option is free to users; all other versions require authorization. Venues Services allows for searches to be done within a certain location or through an entire city. The free data is available at a rate of 5,000 requests per

hour, which may not be enough for some applications of the dataset. However, rate limit increases can be requested. The limited amount of data available per hour makes this method for data collection impractical for the data collection used in this dissertation.

In addition to the use of the Foursquare API for data collection, data from Foursquare as well as other LBSN data sources can be attained using two other approaches: through a third party or through the use of a trolling algorithm. While the third party method is able to provided data via historical or "firehose," the instantaneous streaming of data from the app as it is received to the app, there are costs associated ranging from a few hundred dollars to a few thousand dollars (Gnip 2014) which make the source less accessible for this research project.

The final method of using a trolling algorithm included the identification of venues and the creation of computer programing code that would create a snapshot of the total number of check-ins for each venue within the study area. In addition to the number of check-ins, each venue's unique ID, name, categorical information, geographical information in the form of latitude and longitude, and the number of unique users was collected for each time period during the study duration. Since the computer program was coded to take snapshot at 45-50 minute intervals, the hourly check-in rate for each venue was calculated using the following formula:

$$C_{hr} = \left(\frac{x_2 - x_1}{t_2 - t_1}\right) * 60 \quad (Eqn. 3.1)$$

where C_{hr} represents the check-ins per hour, x_i represents the number of check-ins from the two time intervals, and t_i represents the time interval in minutes.

Due to its confirmed venue locations, popularity, comprehensive functionality, and free application programming interface (API), Foursquare was selected as the data source for this dissertation. The following sections will delve further into the Foursquare

venue and user demographics with respect to general trends as well as the attained Austin dataset used within subsequent chapters of this dissertation. The Austin dataset was collected over the three week period that encompassed June 11 to July 2, 2012.

Foursquare Venue Characteristics

Within Foursquare, users can check-in to existing or create new venues. Foursquare has identified four types of places: public places that can be checked-in to by any user, sub-places that are public or private places inside another place (i.e., shops within a mall), private places that will only show up for people who frequently check-in at that location (i.e., office break rooms, personal cars), and homes, which keep the address private from non-friends. When a venue is created and after determining the place type of the venue, a category as well as subcategories can be assigned at the discretion of the venue creator. Foursquare has ten predetermined categories which include:

1.) Arts & Entertainment – 30 defined subcategories including aquariums, art galleries, casinos, museums (subcategories of type), performing arts venues (subcategories exist), and zoos.

2.) College & University – 23 defined subcategories including collegiate buildings (subcategories exist), bookstores, classrooms, laboratories, libraries, stadiums (subcategories of type), trade schools, and universities.

3.) Events* - eight defined subcategories including conferences, conventions, festivals, music festivals, parades, and street fairs.

4.) Food – over 100 defined subcategories including American restaurants, bakeries, cafeterias, Chinese restaurants (subcategories by type), coffee shop, food trucks, juice bars, pizza places, and frozen yogurt places.

5.) Nightlife Spots – 22 defined subcategories including bars, breweries, lounges, nightclubs, and sports bars.

6.) Outdoors & Recreation – 47 defined subcategories including athletics and sports (subcategories by type), beaches, campgrounds, farms, gardens, harbors/marinas, lakes, national parks, nature preserves, pedestrian plazas, scenic lookouts, states and municipalities, and trails.

7.) Professional & Other Places – 29 defined subcategories including buildings, community centers, convention centers, distribution centers, factories, governmental buildings (subcategories by type), libraries, medical centers (subcategories by type), offices (subcategories by type), parking areas, schools (subcategories by type), spiritual centers (subcategories by type), and warehouses.

8.) Residences – five defined subcategories including assisted living places, homes, housing developments, residential buildings/apartments/condos, and trailer parks.

9.) Shops & Services – over 100 defined subcategories including automotive shops, banks, car dealerships, clothing stores (subcategories by type), discount stores, EV charging stations, food and drink shops (subcategories by type), gyms/fitness centers (subcategories by type), malls, outlet stores, pawn shops, shipping stores, and storage facilities.

10.) Travel & Transport – 30 defined subcategories including airports (subcategories by type), bike rentals/shares, boats or ferries, border

crossings, bus stations, bus stops, hotels (subcategories by type), intersections, light rails, RV parks, rental car locations, rest areas, roads, streets, subways, taxis, and train stations.

At the time of data collection the Events category did not exist and will not be included in within further discussions on venues within the dissertation. The remaining nine were included within the dataset attained with the naming convention kept consistent with the exception of the "Outdoors & Recreation" category which was called "Great Outdoors" within the dataset.

Since categories are user assigned and not mandatory, venues existed within the dataset without categorical assignment. For these venues two methods were employed to determine the category for the analysis. The first used a key word search on venue name to assign the appropriate category. The remaining venues were then explored via their subcategories. Those that were still unidentified, were categorized as "Unclassified" and were not included within the dataset for the travel demand analysis component of this dissertation.

Exploration into the Austin dataset was done to understand the coverage and demographics that existed within. Using ArcGIS, a visual analysis was performed initially to determine the spatial coverage of the venues (blue) and residences (green) included within the study area using the latitude and longitude data collected for each (Figure 3.3). While this initial figure shows reasonable coverage, Figure 3.4 was created to provide a better pictorial representation of the density of all venues per traffic analysis zone (TAZ) and further demonstrated the spatial coverage and the concentration of check-in venues within the central business district (CBD) located in the center of the study area. Moreover, the graphic illustrates the existence of venues with check-ins in almost every TAZ within the study area with only the three highlighted TAZ without any

venues, further demonstrating the spatial coverage available by this data set. Further exploration of the type of venues in each TAZ with respect to land use will be in a subsequent section of this chapter, which will also address the three TAZs without venues.

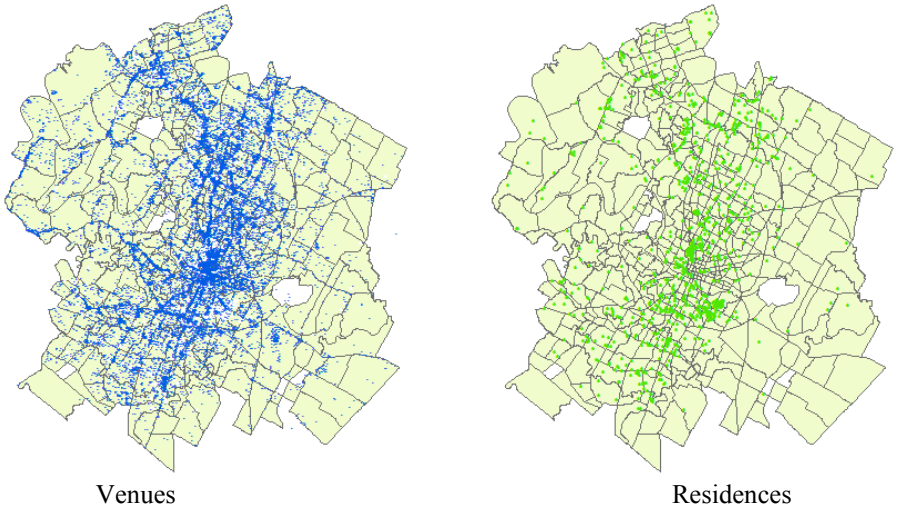


Figure 3.3: Foursquare Venue and Residence Spatial Coverage

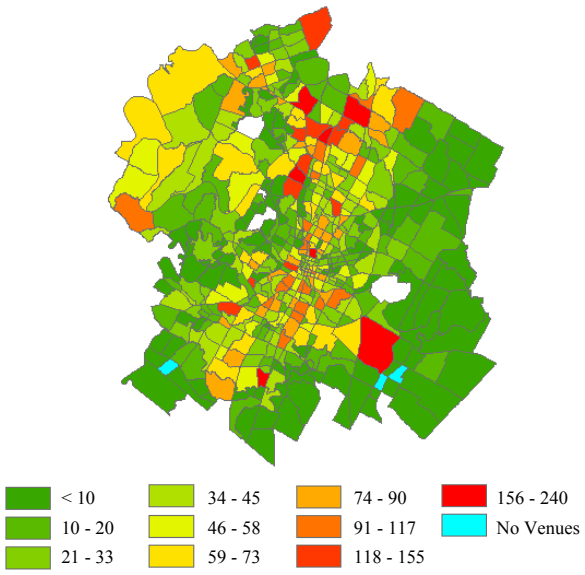


Figure 3.4: Foursquare Check-in Venues Density

An analysis was also done to determine how the venues within the dataset were categorically broken down (Figure 3.5) with respect to weekday, weekend, and total check-ins. Some additional insights can be seen from this representation of the data. Professional and Other Places are more commonly affiliated with weekday check-ins indicating that these check-ins are likely for work purposes. Leisure related venues, Shops & Services and Nightlife, are more frequently checked into on weekends when individuals with traditional working schedules (Monday through Friday) have more leisure time. Finally, residential locations are more likely to be checked into during the week, which may relate to home activated of a significant duration occurring on the weekends (i.e., staying home all day).

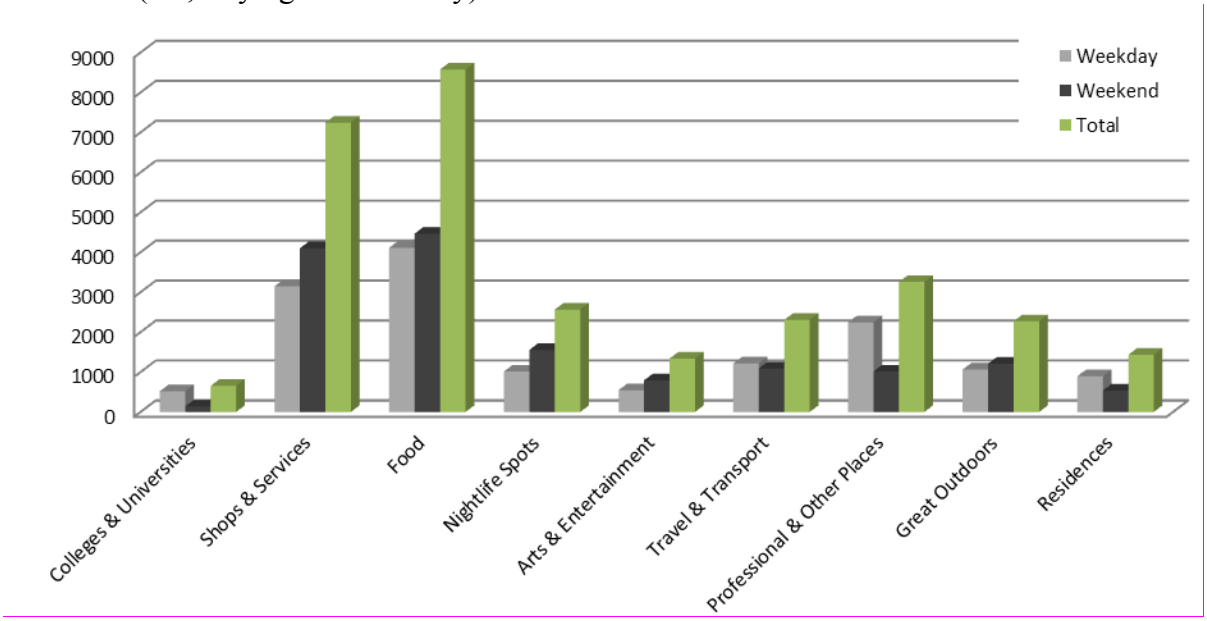


Figure 3.5: Venue Weekday and Weekend Categorical Breakdown

To further understand the data that was attained from the data source, an analysis was performed to determine the number of venues and check-ins as well as the average number of check-ins collected for each of the Foursquare categories. Table 3.1 and

Figure 3.6 provide general details on the data collected. While the number of Shops & Service venues is the largest, this data category also is one of the more commonly checked into categories and has a similar percentage with respect to these check-ins. The most commonly checked-in venues are Food category venues, which, with the Shops & Services category, account for 51.3% of all check-ins within the data set. This check-in statistic will be examined further with respect to time to determine whether the data produces an unrealistic skew or if these check-ins can be attributed to pre- and post-work and lunchtime activities. Additionally, the low representation of residences will be further examined to determine its statistical impact.

Category	# of Venues	% of Venues	# of Check-ins	% of Check-ins	Avg. Check-ins
Colleges & Universities	719	3.8%	367,866	5.5%	512
Shops & Services	5187	27.1%	1,389,636	20.9%	268
Food	2809	14.7%	2,021,897	30.4%	720
Nightlife Spots	547	2.9%	669,712	10.1%	1224
Arts & Entertainment	592	3.1%	324,249	4.9%	548
Travel & Transport	792	4.1%	479,305	7.2%	605
Professional & Other Places	4679	24.4%	832,999	12.5%	178
Great Outdoors	1596	8.3%	278,065	4.2%	174
Residences	711	3.7%	182,825	2.7%	257
Unclassified	1538	8.0%	102,692	1.5%	67
TOTAL	19170		6,649,246		347

Table 3.1: Foursquare Category Venue and Check-in Statistics.

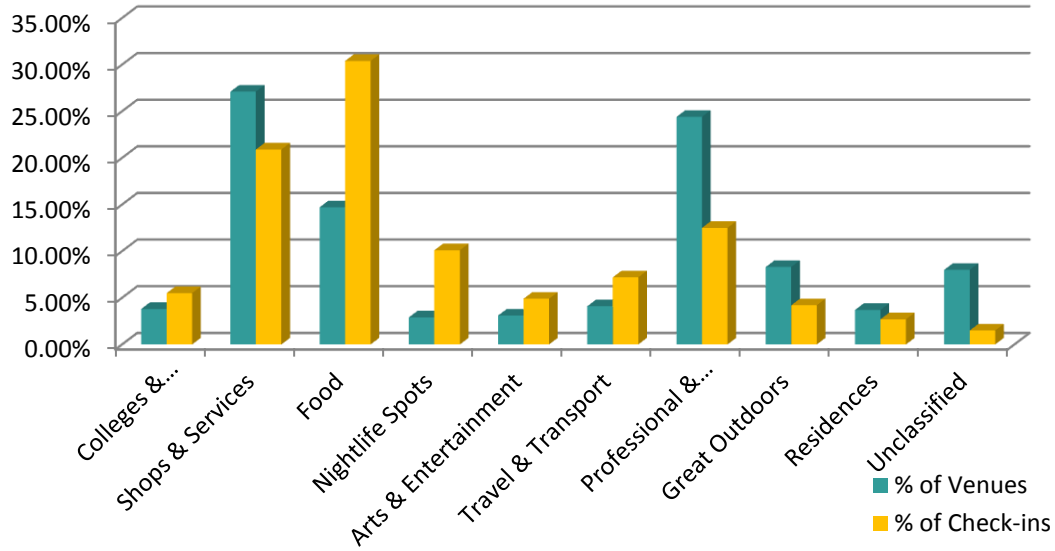


Figure 3.6: Venue and Check-in Statistics

To better understand the check-in and venue relationship, the average number check-ins per venue per category was also calculated. As shown in Table 3.1, the largest average number of check-ins is found within the Nightlife Spots category and the least is found from the Unclassified category. This low average and the low percentage of check-ins for the Unclassified category led to its removal from the travel demand analysis without major compromise to the data set. With respect to average check-ins, the Transportation category was surprisingly large, but can be accounted for by the limited number of bus and rail routes within Austin, which are the common venues checked into for the category, and the number of users of the public transit system within Austin.

Further analysis of the dataset was performed to get a better understanding of any existing categorical trends with respect to day of the week. Due to the size of the dataset, over 30 million data points, SPSS was used to created crosstabs for conduction of this analysis. Table 3.2 provides the number of categories by day of week. From this table one can see that during the weekdays, Tuesday is the most commonly check-in on day (noted

with green background coloring) for all categories with Food and Professional & Other categories receiving the most (noted with red background coloring). Wednesday has less check-in venues being checked-in than the other weekdays. For the weekend, there is consistency throughout the weekend shown with the exception of Great Outdoors and Professional having higher check-ins on Saturdays and Nightlife Spot and Residence having the largest number of check-ins on Sunday (noted with red background coloring). For the Nightlife Spot and Residence categories the Sunday Values were the highest for the entire dataset (noted with red background coloring) and are likely attributed to Saturday night activities. Figures 3.7 show pictorially each categories day of week trends.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Arts & Entertainment	24264	31900	23927	24262	24263	24264	24264
College & University	32040	41966	31595	32040	32039	32040	32040
Food	187776	247594	185168	187773	187775	187776	187776
Nightlife Spot	30816	40473	30388	30815	30816	30816	186120
Great Outdoors	67800	89375	66881	67824	67803	67800	30816
Professional & Other	186120	245551	183535	186117	186120	186120	32328
Residence	32328	42667	31879	32327	32327	32328	67800
Shop & Service	240840	317962	237494	240838	240839	240840	240840
Travel & Transport	32832	43198	32376	32831	32832	32832	32832
Total	834816	1100686	823243	834827	834814	834816	834816

Table 3.2: Venue Categories Checked-in to by Day of Week - Weekday.

Figure 3.7: Day of Week Breakdown by Category

The data set was explored to determine if there were trends in time of day for check-ins occurring for during the week with a focus on weekdays since there are

common trends found within them (i.e., traditional working hours). Table 3.3 provides a percentage breakdown of check-ins by hour by weekday for all of the category types. Within Monday, the check-ins begin to have a noticeable increase starting during the 6 a.m. hour, which continues until the 8 a.m. hour. During the a.m. peak hours for traffic (6 a.m. to 9 a.m.) there are fairly consistent increases in the amount of check-ins across the days of the week. After this time period, the number of check-ins drops off slightly until the 12 p.m. lunch hour. During this time period, there is a noticeable increase that is consistent throughout the week with a total of 7.75% of all weekday check-ins occurring during this hour. The highest percentage of check-ins are during the 1 p.m. hour on Mondays, which may be attributed to individuals going to lunch or running errands and then returning to the workplace. This trend in not seen within the rest of the week days, and warrants further investigation. The next increase in check-in activity is found during the p.m. traffic peak (5 p.m. to 7 p.m.) when individuals are able to leave their workplaces and participate in other activities. After 8 p.m., there is a trend of lesser check-ins which continues until the a.m. peak period of the following day. It is of interest to note that Friday evening does not show an increase in check-in activity, despite the following day being a non-work day for many individuals. One other interesting check-in trend is the Tuesday 2 a.m. increase in check-ins, which will be further explored by examining trends in check-ins by hour and by category. Figure 3.8 provides a visual of the check-in trends by weekday that illustrates the increasing and decreasing percentages of check-ins throughout the hours of the day.

Hour of Day	Monday	Tuesday	Wednesday	Thursday	Friday	Total
12 a.m.	1.19%	0.95%	1.27%	1.45%	1.76%	1.31%
1 a.m.	0.71%	0.54%	0.73%	0.83%	0.96%	0.75%
2 a.m.	0.34%	3.89%	0.82%	0.42%	0.44%	1.35%
3 a.m.	0.12%	0.36%	0.31%	0.19%	0.32%	0.27%
4 a.m.	0.27%	0.26%	0.37%	0.23%	0.29%	0.28%
5 a.m.	0.87%	0.78%	0.82%	0.91%	0.66%	0.80%
6 a.m.	2.17%	1.87%	1.95%	1.68%	1.58%	1.84%
7 a.m.	3.96%	3.95%	3.92%	3.91%	3.55%	3.86%
8 a.m.	5.56%	5.36%	5.84%	5.85%	4.83%	5.47%
9 a.m.	4.91%	5.01%	5.24%	5.16%	4.72%	5.00%
10 a.m.	4.36%	4.43%	4.27%	4.26%	4.04%	4.27%
11 a.m.	5.29%	5.32%	5.33%	5.43%	5.58%	5.39%
12 p.m.	7.54%	7.69%	7.84%	7.93%	7.77%	7.75%
1 p.m.	10.00%	6.73%	7.64%	6.51%	6.69%	7.42%
2 p.m.	5.67%	5.22%	5.70%	5.33%	5.77%	5.52%
3 p.m.	5.11%	4.72%	4.62%	4.72%	5.30%	4.89%
4 p.m.	5.36%	5.05%	5.01%	4.92%	5.21%	5.10%
5 p.m.	6.48%	6.26%	6.11%	6.56%	6.58%	6.39%
6 p.m.	7.42%	7.79%	7.74%	7.79%	6.96%	7.55%
7 p.m.	7.11%	7.47%	7.49%	7.45%	7.16%	7.34%
8 p.m.	6.17%	6.27%	6.04%	6.59%	6.86%	6.39%
9 p.m.	4.21%	4.77%	5.11%	5.38%	5.36%	4.97%
10 p.m.	3.28%	3.07%	3.40%	3.58%	4.40%	3.54%
11 p.m.	1.92%	2.24%	2.41%	2.92%	3.22%	2.54%

Table 3.3: Check-ins by Hour by Weekday.

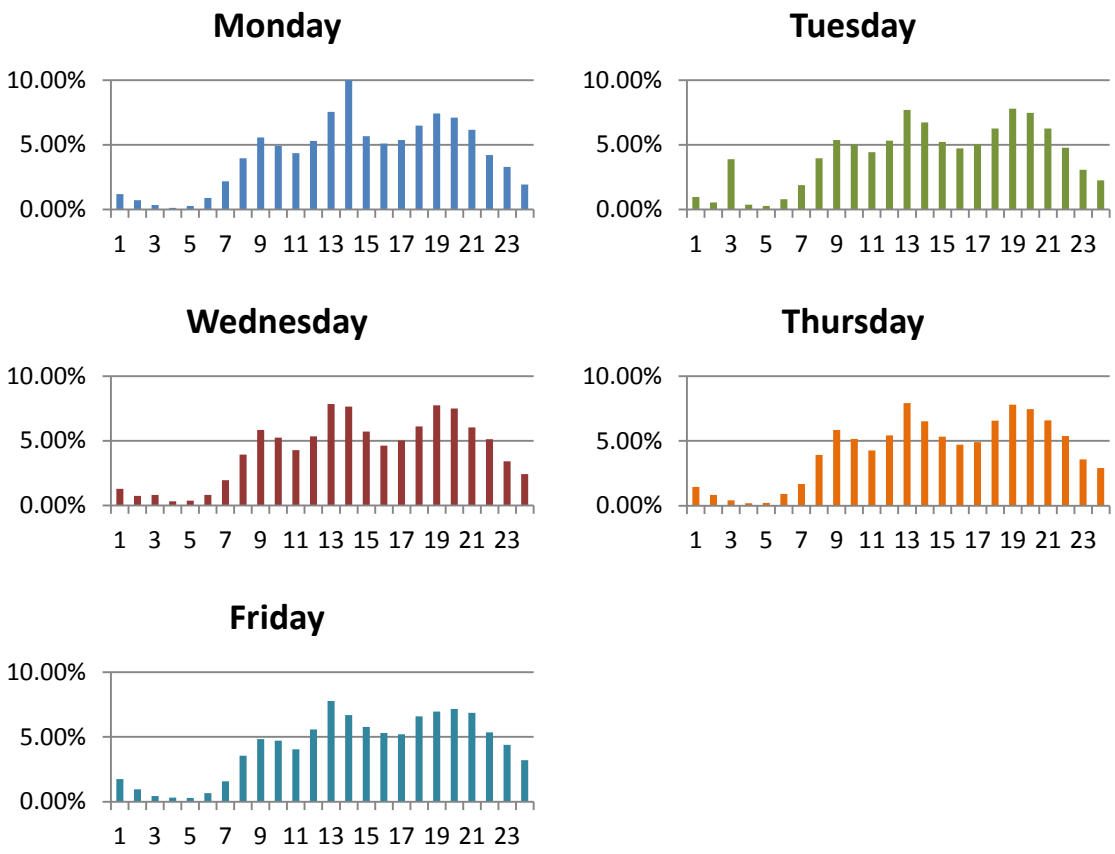


Figure 3.8: Weekday Check-ins by Hour

As was mentioned above, analysis into how users check-in throughout the day at different venue categories needs to be done to be able to provide any insight into the type of trips that users are making. Table 3.4 presents the percentages of check-ins per hour for each of the venue categories and uses a color gradation scheme of green (low) to red (high) to demonstrate trends. For the table, one can clearly see Professional & Other Places check-ins tend to be between 8 a.m. and 10 a.m., indicating the likelihood of these being individuals arriving at work for the day. This is also confirmed by the percentage of check-ins falling throughout the day within this category. The previous statement concerning the increase in check-in activities during the lunch hour is confirmed from the

examination of the Food venue check-ins by the hour, where the largest percentages of check-ins occur during this time. With respect to this same category, an increase in check-ins is also found during the traditional dinner hours. The previous statement concerning errands being run during the lunch hour is also confirmed by the increase in check-in activities in the Shops & Services categories during that time period. With respect to the Residences category, there is little check-in activity during the day until the evening hours, when individuals return home from the workplace. As expected, recreational activities that fall into the Arts & Entertainment, Great Outdoors, and Shops & Services have the largest percentage of check-ins in the evening after 5 p.m. Additionally, the category of Nightlife Spots sees the majority of its check-ins in the evening starting at 8 p.m. and continuing until 2 a.m. when bars and nightclubs close. An interesting trend within the Colleges & Universities is the large percentage of checking during the 8 a.m. hour, indicating the likelihood of these check-ins being for the start of the work day. The final category of Travel & Transport also reveals two trends that fall in line with the working commuters' activities. For this category the largest number of check-ins occur during 8 a.m. and 5 and 6 p.m. These are likely user check-ins as they go to and from their workplaces.

Hour of Day	Arts & Entertainment	Colleges & Universities	Food	Nightlife Spots	Great Outdoors	Professional & Other Places	Residences	Shops & Services	Travel & Transport	Total
12 a.m.	2.57%	0.45%	1.11%	6.43%	1.19%	0.42%	3.48%	0.55%	2.46%	1.51%
1 a.m.	1.34%	0.20%	0.66%	3.88%	0.61%	0.24%	1.86%	0.22%	1.69%	0.86%
2 a.m.	1.70%	1.09%	1.98%	2.64%	1.30%	0.87%	2.59%	1.06%	2.17%	1.54%
3 a.m.	0.16%	0.13%	0.35%	0.25%	0.29%	0.20%	1.07%	0.17%	0.82%	0.31%
4 a.m.	0.07%	0.02%	0.28%	0.12%	0.15%	0.31%	0.61%	0.29%	1.36%	0.32%
5 a.m.	0.12%	0.20%	0.51%	0.11%	1.70%	0.59%	1.54%	1.11%	3.22%	0.91%
6 a.m.	0.33%	0.69%	1.48%	0.17%	4.49%	2.34%	3.36%	2.12%	4.48%	2.08%
7 a.m.	0.92%	5.67%	4.19%	0.36%	6.27%	7.79%	4.25%	2.85%	5.67%	4.31%
8 a.m.	1.67%	12.18%	5.73%	0.54%	5.59%	13.16%	4.21%	3.73%	6.73%	6.14%
9 a.m.	2.12%	9.31%	5.24%	0.67%	4.94%	11.83%	3.50%	4.16%	4.45%	5.62%
10 a.m.	2.99%	7.72%	2.63%	0.62%	4.15%	8.37%	3.16%	4.47%	3.87%	4.37%
11 a.m.	3.63%	8.39%	5.42%	1.21%	4.25%	6.96%	3.10%	5.66%	4.30%	5.19%
12 p.m.	3.77%	8.55%	10.81%	2.51%	4.43%	7.34%	3.58%	6.89%	4.31%	6.95%
1 p.m.	4.45%	9.28%	9.47%	2.56%	5.11%	7.63%	3.64%	7.75%	4.48%	7.05%
2 p.m.	5.16%	6.29%	5.51%	2.07%	4.38%	6.19%	3.36%	6.64%	4.33%	5.41%
3 p.m.	4.52%	4.82%	4.14%	2.19%	4.55%	5.21%	3.44%	6.81%	4.51%	4.93%
4 p.m.	4.52%	4.40%	3.86%	3.43%	5.30%	4.65%	4.92%	7.41%	5.84%	5.21%
5 p.m.	4.90%	6.00%	5.00%	5.85%	6.35%	4.65%	6.70%	8.98%	6.68%	6.37%
6 p.m.	9.51%	5.78%	7.15%	8.42%	8.67%	3.76%	7.97%	8.82%	6.73%	7.26%
7 p.m.	11.73%	3.12%	8.23%	9.86%	8.65%	2.78%	7.24%	7.41%	5.08%	6.93%
8 p.m.	8.76%	2.37%	7.12%	11.09%	7.68%	1.67%	6.80%	5.94%	4.71%	5.96%
9 p.m.	10.50%	1.70%	5.12%	11.63%	4.64%	1.20%	7.10%	3.66%	4.27%	4.68%
10 p.m.	8.78%	1.05%	2.72%	11.79%	3.35%	1.03%	6.48%	2.10%	3.93%	3.48%
11 p.m.	5.77%	0.58%	1.30%	11.59%	1.94%	0.84%	6.05%	1.18%	3.93%	2.62%

Table 3.4: Weekday Venue Categories Checked-in to by Hour with Emphasis on Category Trends.

Table 3.5 provides another visual for the data from Table 3.4 using the same color gradation along the hours to find trend in check-ins within the categories. Beginning in the early morning at the 6 a.m. and 7 a.m. hours there is a trend of checking-in at Great Outdoors venue categories. This is followed by trend of checking into Professional & Other Places as well as College & Universities, which may indicate individuals wanting to perform outdoor activities (i.e., walking, running) before starting their work day.

Moving through the day another trend is found during the 12 p.m. hour, where again there are significant check-ins at Food venues. During the 1 p.m. hour there are trends toward venues within the Colleges & Universities and Food categories being checked-into. This may relate to an extended lunch hour (Food) and individuals returning to campuses for class or work (Colleges & Universities). Afternoon activities, if not at work, were likely to be in the Shops and Services category between 2 p.m. and 5 p.m. From 6 p.m. to 8 p.m. there are trends of checking-into Arts & Entertainment venues, which is expected since movies and theater performances often begin during the 7 p.m. time period. At the 8 p.m. hour the largest category for check-ins is in the Nightlife Spots category, which continues throughout the evening until the bars and restaurants close at 2 a.m. It is interesting to note that there is another high percentage of venues checked-into during 9 p.m. for the Arts & Entertainment category which relates to typical second showings at movie theaters and at other performance venues. Additionally, at the 2 a.m. hour there is a visible trend of increased check-ins for Residences that may relate to individuals returning home from their evening out.

Hour of Day	Arts & Entertainment	Colleges & Universities	Food	Nightlife Spots	Great Outdoors	Professional & Other Places	Residences	Shops & Services	Travel & Transport
12 a.m.	2.57%	0.45%	1.11%	6.43%	1.19%	0.42%	3.48%	0.55%	2.46%
1 a.m.	1.34%	0.20%	0.66%	3.88%	0.61%	0.24%	1.86%	0.22%	1.69%
2 a.m.	1.70%	1.09%	1.98%	2.64%	1.30%	0.87%	2.59%	1.06%	2.17%
3 a.m.	0.16%	0.13%	0.35%	0.25%	0.29%	0.20%	1.07%	0.17%	0.82%
4 a.m.	0.07%	0.02%	0.28%	0.12%	0.15%	0.31%	0.61%	0.29%	1.36%
5 a.m.	0.12%	0.20%	0.51%	0.11%	1.70%	0.59%	1.54%	1.11%	3.22%
6 a.m.	0.33%	0.69%	1.48%	0.17%	4.49%	2.34%	3.36%	2.12%	4.48%
7 a.m.	0.92%	5.67%	4.19%	0.36%	6.27%	7.79%	4.25%	2.85%	5.67%
8 a.m.	1.67%	12.18%	5.73%	0.54%	5.59%	13.16%	4.21%	3.73%	6.73%
9 a.m.	2.12%	9.31%	5.24%	0.67%	4.94%	11.83%	3.50%	4.16%	4.45%
10 a.m.	2.99%	7.72%	2.63%	0.62%	4.15%	8.37%	3.16%	4.47%	3.87%
11 a.m.	3.63%	8.39%	5.42%	1.21%	4.25%	6.96%	3.10%	5.66%	4.30%
12 p.m.	3.77%	8.55%	10.81%	2.51%	4.43%	7.34%	3.58%	6.89%	4.31%
1 p.m.	4.45%	9.28%	9.47%	2.56%	5.11%	7.63%	3.64%	7.75%	4.48%
2 p.m.	5.16%	6.29%	5.51%	2.07%	4.38%	6.19%	3.36%	6.64%	4.33%
3 p.m.	4.52%	4.82%	4.14%	2.19%	4.55%	5.21%	3.44%	6.81%	4.51%
4 p.m.	4.52%	4.40%	3.86%	3.43%	5.30%	4.65%	4.92%	7.41%	5.84%
5 p.m.	4.90%	6.00%	5.00%	5.85%	6.35%	4.65%	6.70%	8.98%	6.68%
6 p.m.	9.51%	5.78%	7.15%	8.42%	8.67%	3.76%	7.97%	8.82%	6.73%
7 p.m.	11.73%	3.12%	8.23%	9.86%	8.65%	2.78%	7.24%	7.41%	5.08%
8 p.m.	8.76%	2.37%	7.12%	11.09%	7.68%	1.67%	6.80%	5.94%	4.71%
9 p.m.	10.50%	1.70%	5.12%	11.63%	4.64%	1.20%	7.10%	3.66%	4.27%
10 p.m.	8.78%	1.05%	2.72%	11.79%	3.35%	1.03%	6.48%	2.10%	3.93%
11 p.m.	5.77%	0.58%	1.30%	11.59%	1.94%	0.84%	6.05%	1.18%	3.93%

Table 3.5: Weekday Venue Categories Checked-in to by Hour with Emphasis on Hourly Trends.

Further exploration was done to examine the combination of day of the week, category, and hour to look for additional trends within the data set. Examination of the dataset was done using the time groups identified below for each day of the week (Monday through Friday) and used the categories listed after each time group, which were selected based on the trends seen in Tables 3.4 and 3.5 above:

1.) AM Peak (6 a.m. to 9 a.m.) – Colleges & Universities, Great Outdoors, Professional & Other Places, and Travel & Transport
2.) Mid-Morning (10 a.m. to 11 p.m.) – Colleges & Universities, Food, Professional & Other Places, and Shops & Services
3.) Lunch Hour (12 p.m. to 1 p.m.) – Colleges & Universities, Food, Professional & Other Places, and Shops & Services
4.) Mid-Afternoon (2 p.m. to 4 p.m.) – Colleges & Universities, Professional & Other Places, and Shops & Services
5.) PM Peak (5 p.m. to 7 p.m.) – Arts & Entertainment, Food, Great Outdoors, Nightlife Spots, Residences, Shops & Services, and Travel & Transport
6.) Evening (8 p.m. to 10 p.m.) – Arts & Entertainment, Food, Great Outdoors, Nightlife Spots, Residences, and Shops & Services
7.) Late Night (11 p.m. to 5 a.m.) – Nightlife Spots, and Residences

To do this, ArcGIS was used to map venue check-in data (venue GPS location and number of check-ins) to the study area of Austin, TX. The ArcGIS symbology features were then employed to provide meaningful visualization to each categorical venue map and to be able to visual examine special changes in check-in locations throughout the day. The GIS analysis also examined the number of check-ins per hour based on the size of the circles, not just the number of venues being checked-into in an hour. In addition, venues with the largest amounts of check-ins during the analysis periods were identified for each category.

Starting with the A.M. Peak for analysis, the first category that was examined was the Colleges and Universities. Figure 3.9 shows the trend of Wednesdays and Thursdays having greater check-in intensity in the University of Texas (UT) area than the other

days. For the Mid Morning time period, there is consistency throughout the time period and throughout the weekdays with lesser intensity found in the UT area than other time periods. For the Lunch and Mid Afternoon time periods, check-ins were similar to the Mid Morning time period with consistency seen throughout the week (see Appendix A for graphics). Figure 3.10 shows the changes in number of venues with various check-in amounts throughout the time periods. The majority of single check-ins occur during the A.M. Peak at over 900 unique venues, which are individuals checking-in for work or classes. Table 3.6 shows the venues with the most check-ins for each of the time periods all which occur within the UT area. It is interesting to note that few locations have multiple check-ins, and those that do are affiliated with UT.

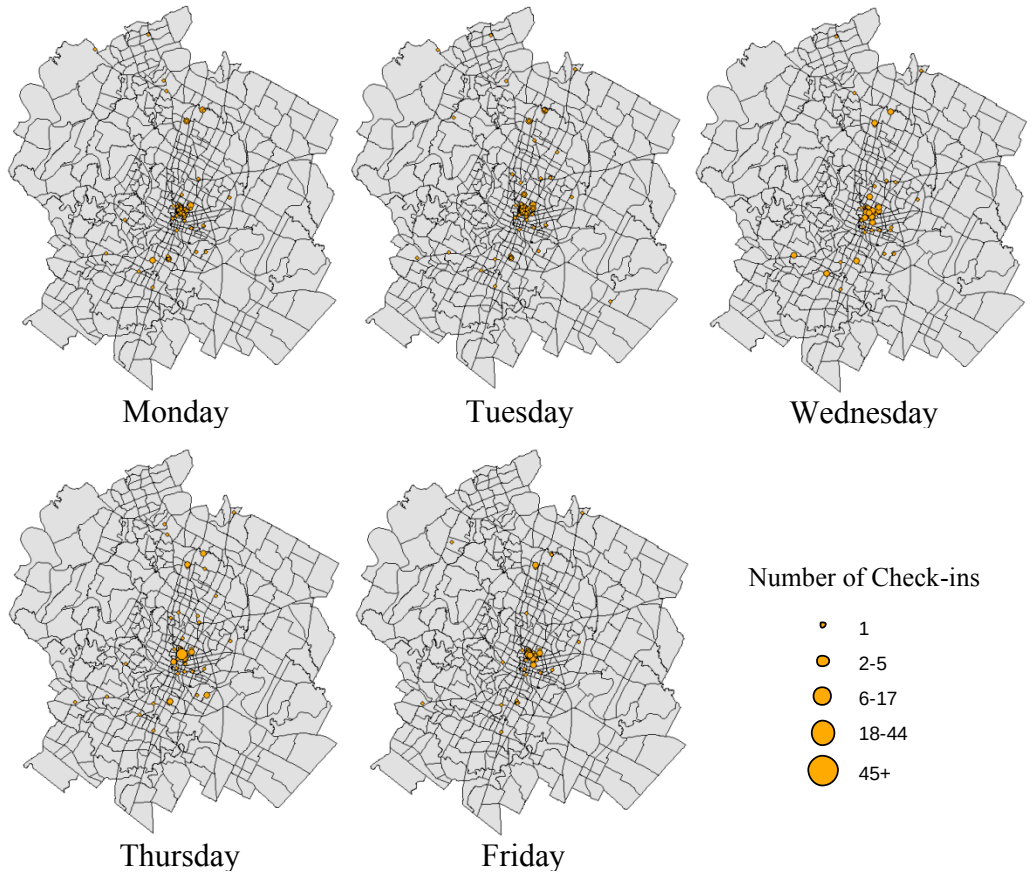


Figure 3.9: A.M. Peak Colleges & Universities Venue Check-ins

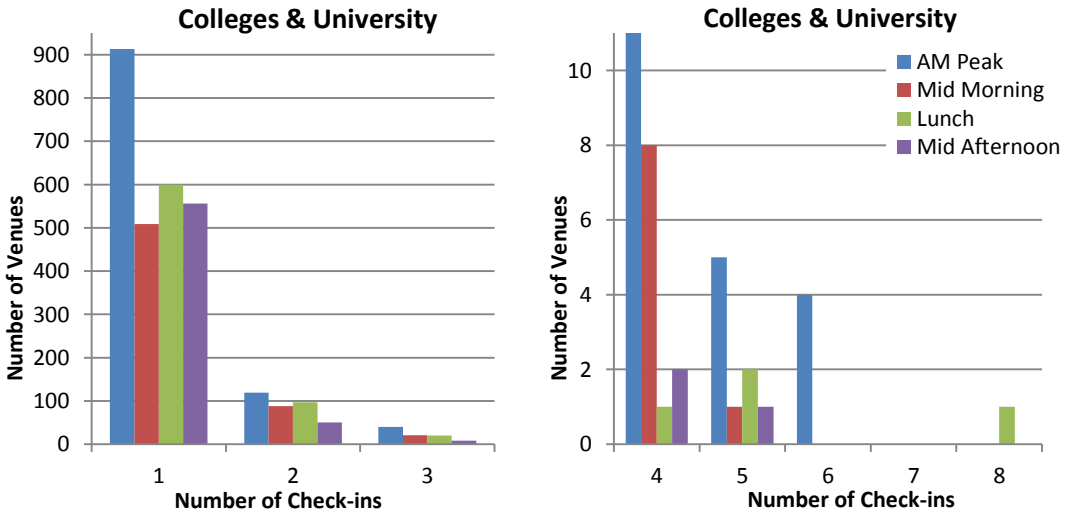


Figure 3.10: Colleges & Universities Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day(s) of Week
AM Peak	6	4	The University of Texas at Austin	Monday, Thursday
Mid Morning	5	1	Student Activity Center (SAC)	Monday
Lunch	8	1	Hogg Memorial Auditorium (HMA)	Monday
Mid Afternoon	5	1	University Teaching Center (UTC)	Thursday

Table 3.6: Colleges & Universities Venues with the Most Check-ins.

The next category identified for analysis was the Professional & Other Places. Figure 3.11 shows the A.M. Peak trends for the weekdays. Based on this graphic, there is good special coverage with respect to the category type with the expected concentration of venues and check-ins in the central business district (CBD), which is located in the middle of each graphic. The graphic also shows a trend for higher intensity of check-ins in the CBD on Wednesdays and Fridays. This pattern is seen in the other time periods of Mid Morning and Lunch (see Appendix A for graphics for other analyzed time periods). The Lunch time period was noted to have more intensity with respect to check-ins at venues throughout the analysis area and higher intensity was noted for the time period on Tuesdays as well. The Mid Afternoon time period showed lesser intensity throughout the week with the exception of Tuesdays, which showed similar intensity as the Lunch time period. Figure 3.12 shows the changes in number of venues with various check-in amounts throughout the analyzed time periods. Similar to the Colleges & Universities category, the majority of single check-ins occur during the A.M. Peak at over 4000 unique venues. It is of interest to note that there is a jump in the Mid Afternoon single venue check-ins for all of the weekdays, which may be workers returning to their place of employment after going out to lunch. Table 3.7 shows the venues with the most check-ins for each of the time periods all of which occur on Mondays.

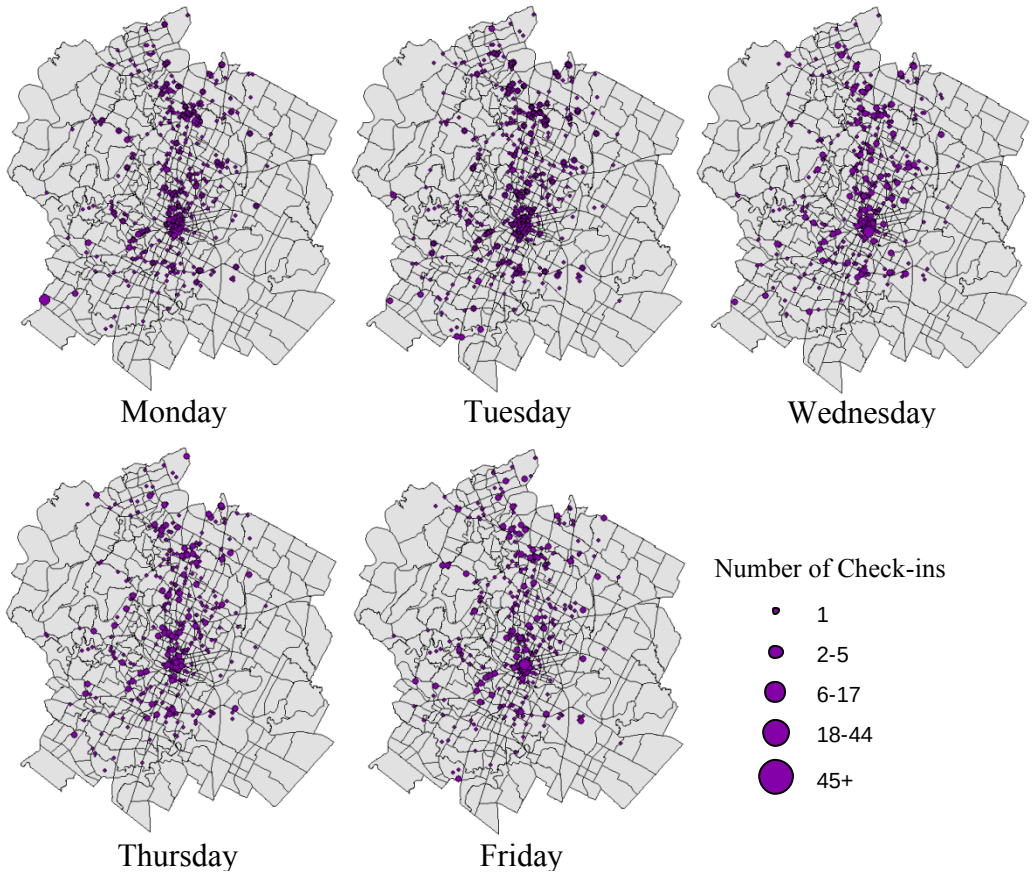


Figure 3.11: A.M. Peak Professional & Other Places Venue Check-ins

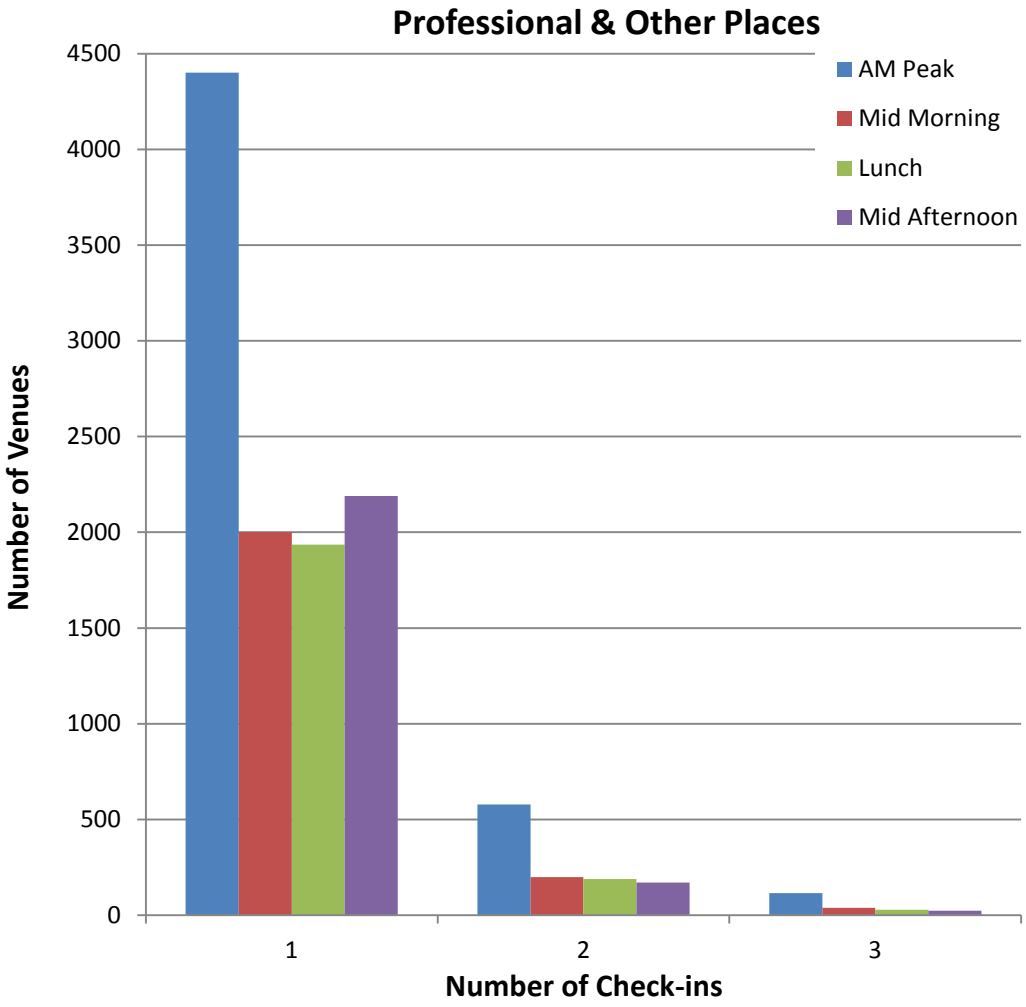


Figure 3.12: Professional & Other Places Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
AM Peak	10	1	Austin Convention Center	Monday
	10	1	Texas State Capitol	Monday
Mid Morning	29	1	Texas State Capitol	Monday
Lunch	77	1	Texas State Capitol	Monday
Mid Afternoon	44	1	Texas State Capitol	Monday

Table 3.7: Professional & Other Places Venues with the Most Check-ins.

The examination of the Great Outdoors category revealed a wide range of venues check-into throughout the study area. For the A.M. Peak, the check-ins were consistent throughout the weekdays with concentrations near the CBD where there are a lot of parks and a trail that circles the Lady Bird and Town Lake areas. Figure 3.13 shows the P.M. Peak for the Great Outdoors category. This graphic shows the more intense check-ins along the trail and park locations compared to the A.M. Peak, with a noticeable increase on Wednesdays. The Evening time period shows similar trends as the P.M. Peak time period with higher intensity shown of Fridays for the trail and park locations. Appendix A contains the graphics for the A.M. Peak and Evening time periods. Figure 3.14 shows the changes in number of venues with various check-in amounts throughout the analyzed time periods. The majority of single check-ins occur during the P.M. Peak with 1170 unique venues. It is of interest to note that the number of A.M. Peak single check-in venues is 1131. As was seen with the previous category, there are limited locations with multiple check-ins during any of the time periods analyzed for the Great Outdoors category. However, there are some single venue locations that have a significant number of checks and the venues with the most check-ins are shown in Table 3.8. From this table, Wednesdays see a significant number of check-ins in the Zilker Park area indicating the areas popularity with Foursquare users.

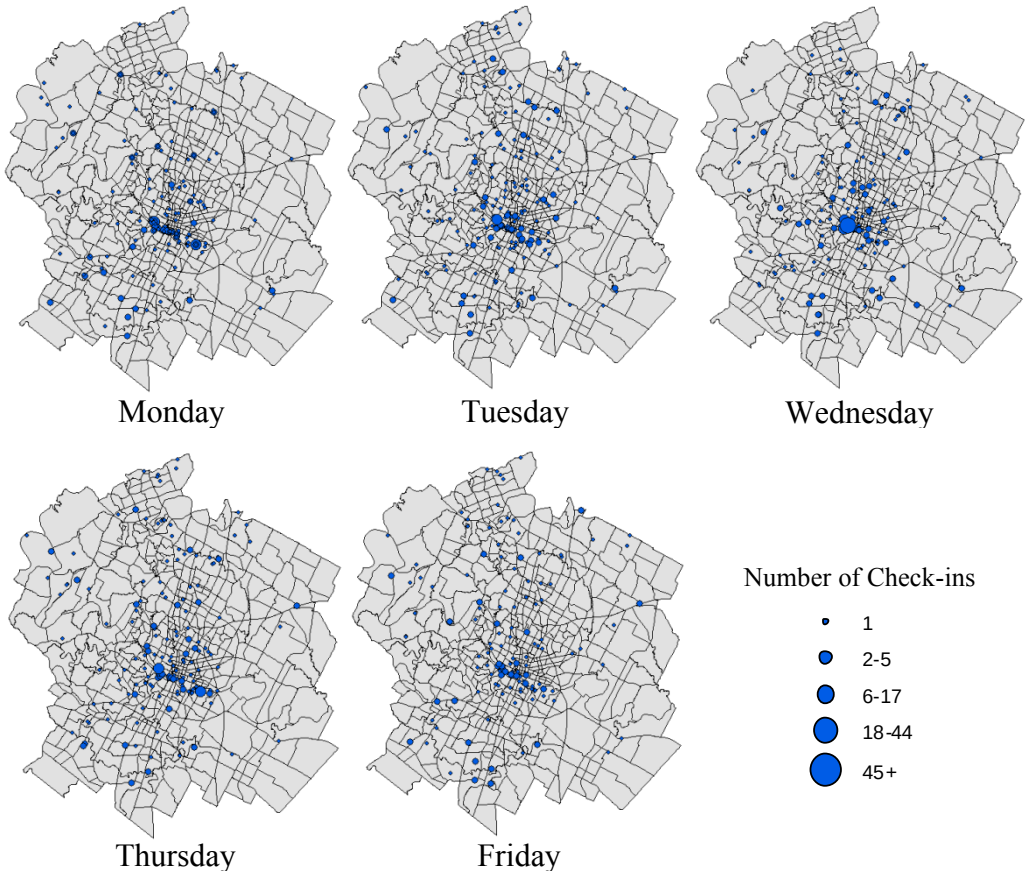


Figure 3.13: P.M. Peak Great Outdoors Venue Check-ins

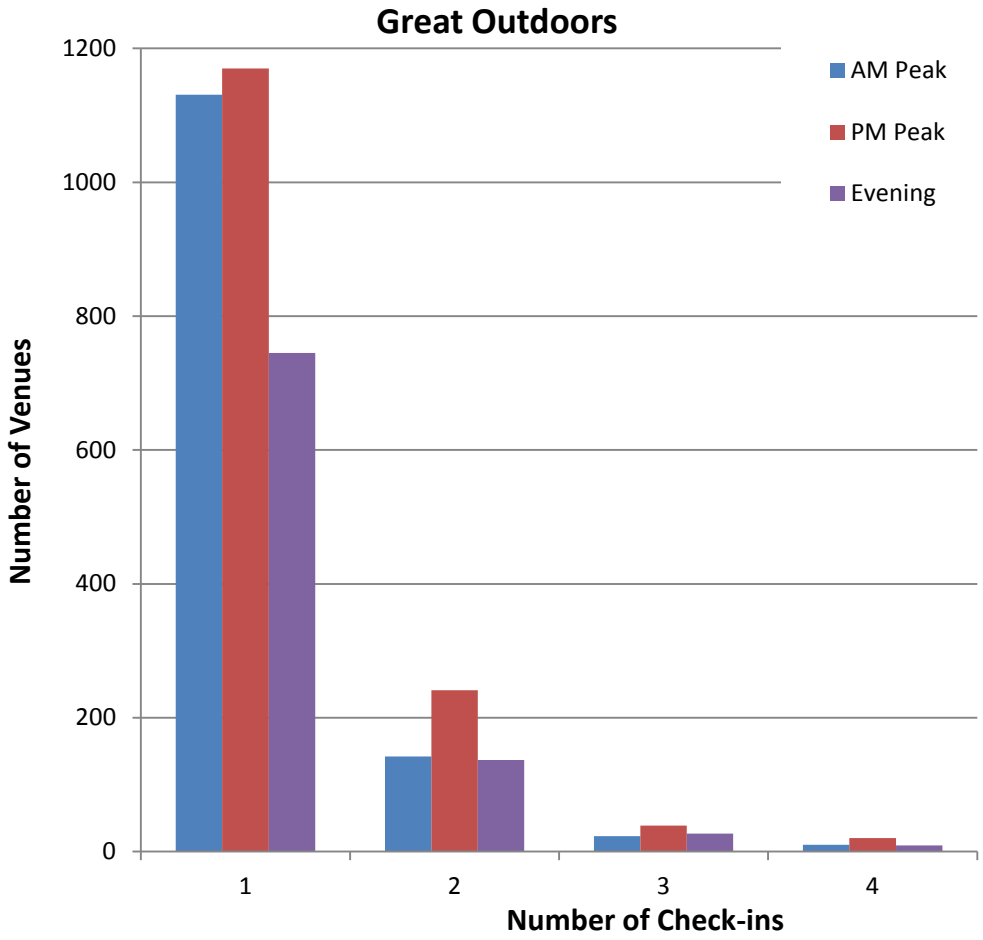


Figure 3.14: Great Outdoors Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
AM Peak	5	1	Lady Bird Lake Trail	Tuesday
	5	1	Lady Bird Lake Trail	Tuesday
PM Peak	67	1	Zilker Park	Wednesday
Evening	78	1	Zilker Park	Wednesday

Table 3.8: Great Outdoors Venues with the Most Check-ins.

The Travel & Transport category revealed trends in checking-in at the airport throughout the weekdays and time periods examined (AM and P.M. Peaks). For the P.M.

Peak, Figure 3.15 illustrates the airport, which is located in the lower right area of the study area, check-in trend as well as the increased intensity of check-ins on Fridays. Mondays also see an increase in the number of venues checked-into compared to the other days of the week for the time period. For the A.M. and P.M. Peaks there are consistent check-ins within the CBD area, although there are less check-in intensity for the A.M. Peak time period (Appendix A). Figure 3.16 shows the changes in number of venues with various check-in amounts between the A.M. and P.M. Peak periods. The majority of single check-ins occur during the A.M. Peak with 800 unique venues. It is of interest to note that the locations with two check-ins have a consistent number of venues during the two time periods. From Table 3.9 the largest number of check-ins are consistently found at the airport but on two different days: Tuesdays for the A.M. Peak and Thursdays for the P.M. Peak. This trend may be related to business travelers departing on Tuesdays and then returning on Thursdays.

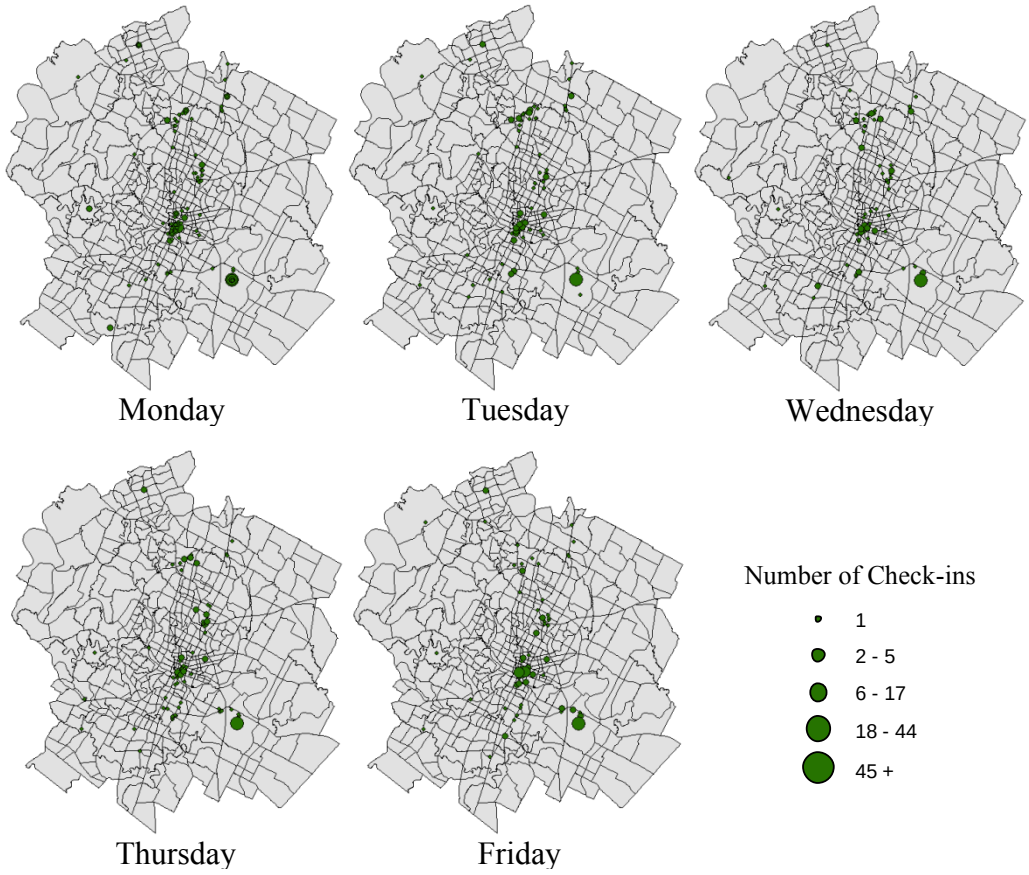


Figure 3.15: P.M. Peak Travel & Transport Venue Check-ins

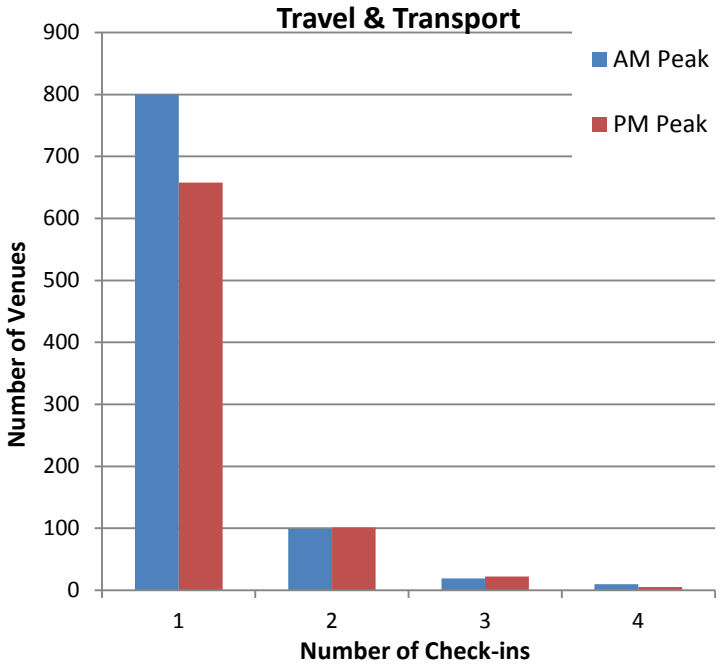


Figure 3.16: Travel & Transport Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
AM Peak	27	1	Austin Bergstrom International Airport	Tuesday
PM Peak	38	1	Austin Bergstrom International Airport	Thursday

Table 3.9: Travel & Transport Venues with the Most Check-ins.

The Food category venues were next examined for the Mid Morning, Lunch, P.M. Peak, and Evening time periods. The Lunch time period, shown in Figure 3.17, displays trends that are seen throughout the time periods. There are numerous venues found within the CBD and throughout the study area. Greater intensities are apparent on Tuesdays, Wednesdays, and Fridays. For the Mid Morning time period, the Lunch trends are seen, but to a lesser degree with respect to the number of venues checked-into. The P.M. Peak and Evening time periods are similar to the Lunch time period with some area having a greater number of check-ins. Appendix A provides the graphics for the Mid Morning,

P.M. Peak, and Evening time periods for the Food category. Figure 3.18 demonstrates how closely the number of single check-ins are for the Lunch and P.M. Peak time periods, 5645 for Lunch and 5771 for P.M. Peak. This close relationship is shown for venues with multiple check-ins as well. With respect to the venues with the most check-ins for each time period, Table 3.10 shows a skew within the dataset towards a single location, Chi\Lantro BBQ, for a single day of the week. The high count of check-ins, 3089, may be due to a promotion, a glitch in the system, or individuals checking-in multiple times.

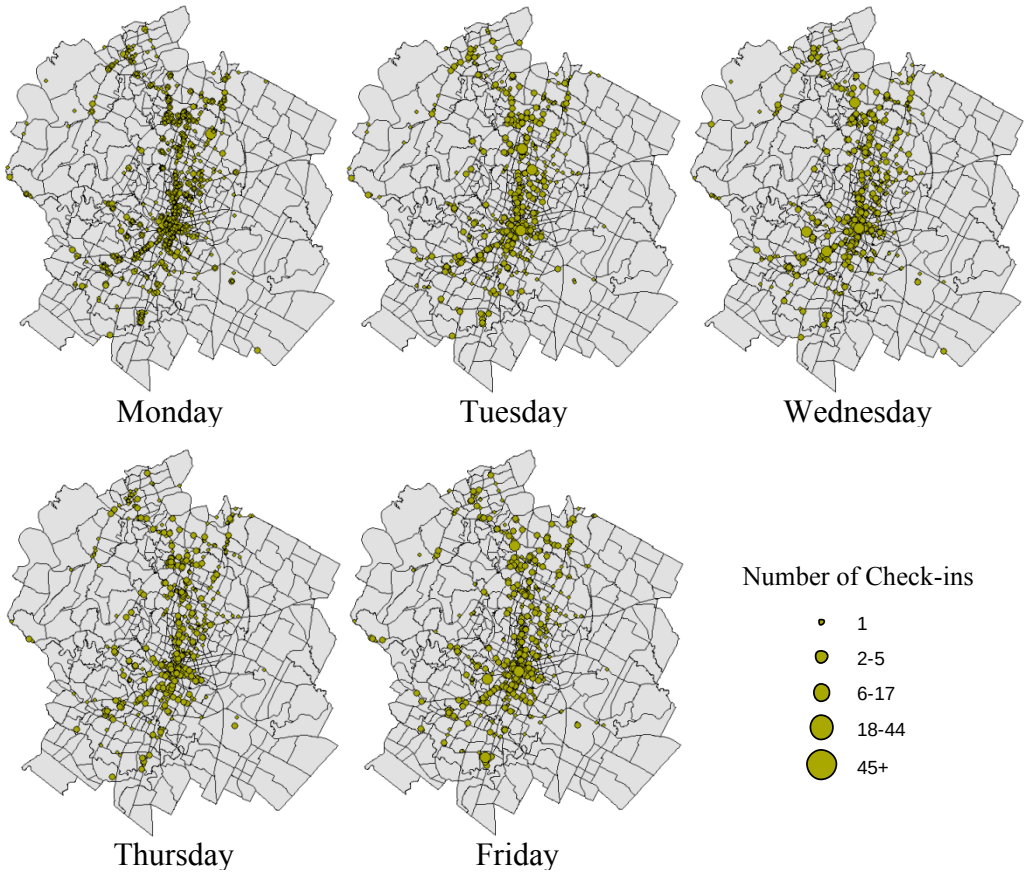


Figure 3.17: Lunch Food Venue Check-ins

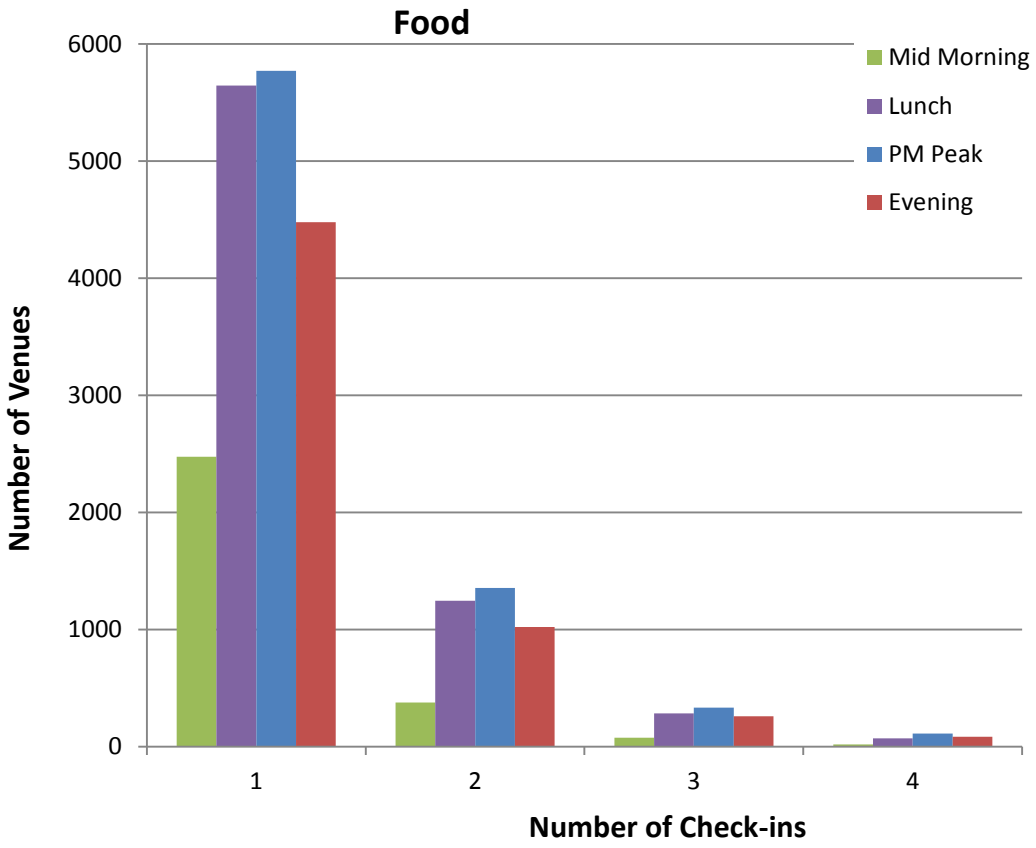


Figure 3.18: Food Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
Mid Morning	5	1	Franklin Barbecue	Friday
	5	1	Hopdoddy Burger Bar	Tuesday
	5	1	Moonshine Patio Bar & Grill	Friday
Lunch	191	1	Whole Foods Bakery	Monday
PM Peak	3089	1	Chi\Lantro BBQ	Monday
Evening	27	1	Stubb\s Bar-B-Q	Tuesday

Table 3.10: Food Venues with the Most Check-ins.

For the next chronological category, the Shops & Services were analyzed for the Mid Morning, Lunch, Mid Afternoon, P.M. Peak, and Evening time periods. The Lunch time period is shown in Figure 3.19 (all other graphics can be found in Appendix A) and

clearly depicts trends that exist within the weekdays. For this time period, higher intensity check-ins are seen on Tuesdays, Thursdays, and Fridays especially in the CBD area. Compared to the Mid Morning time period, there are more venues that are checked-into during the Lunch period, which may correlate to errands being run during this time period. For the Mid Morning time period, there are noticeable fewer venues checked-into with less intensity seen throughout the week. For this time period, Tuesdays see the least intensity of check-ins while the other days have similar intensity patterns. The Mid Afternoon has similar trends that were seen in the Lunch time period with the exception of the higher intensity on Wednesdays especially in the CBD area. For the P.M. Peak the check-in trends are similar to the Lunch and Mid Afternoon trends, while the Evening time period shows a lessening of check-ins across the board but still includes the higher intensity of check-ins on Tuesday, Wednesday, Thursday, and Friday in various areas. Examining the trends in number of venues checked-into based on differing check-in amounts, Figure 3.20 presents the increase in single check-ins that occurs throughout the day with a drop off during the evening, which follows the expectation of many individuals doing shopping activities after working hours. It is of interest to note that there are more multiple (two through six) check-ins to single venues during the P.M. Peak, which includes locations such as gyms and grocery stores. Areas with the largest number of check-ins include a variety of locations (Table 3.11) and are spread throughout the weekdays.

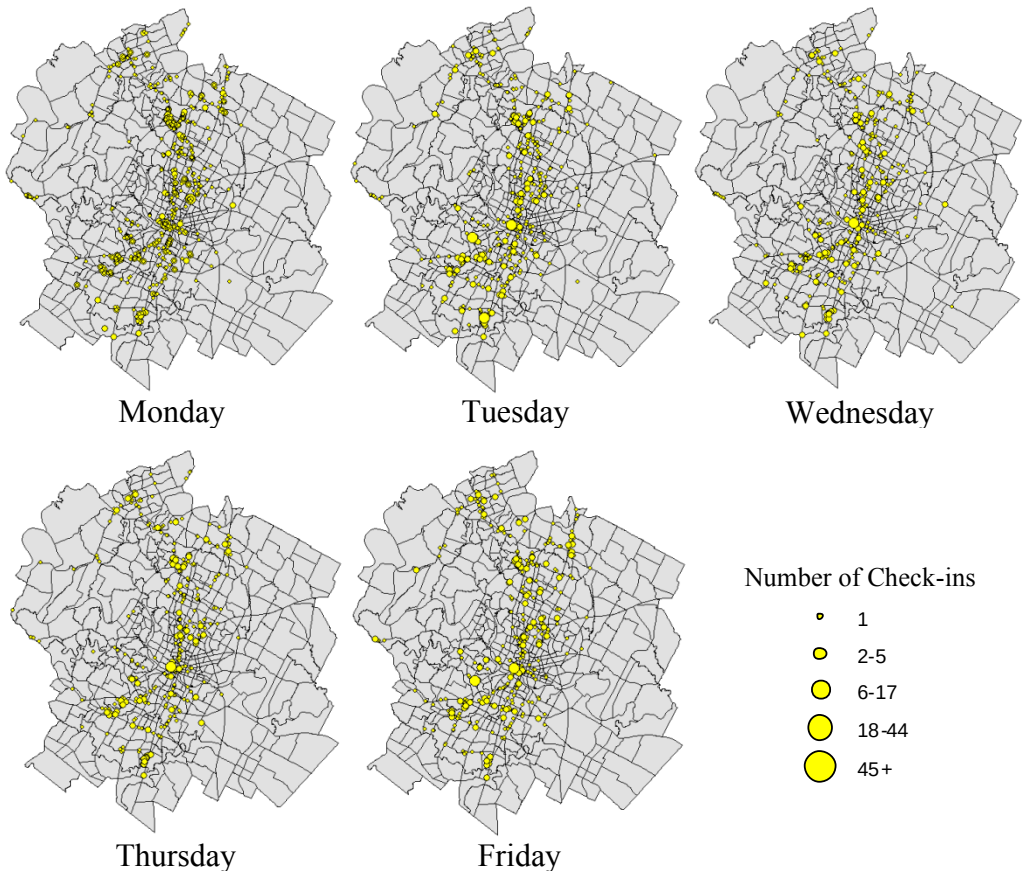


Figure 3.19: Lunch Shops & Services Venue Check-ins

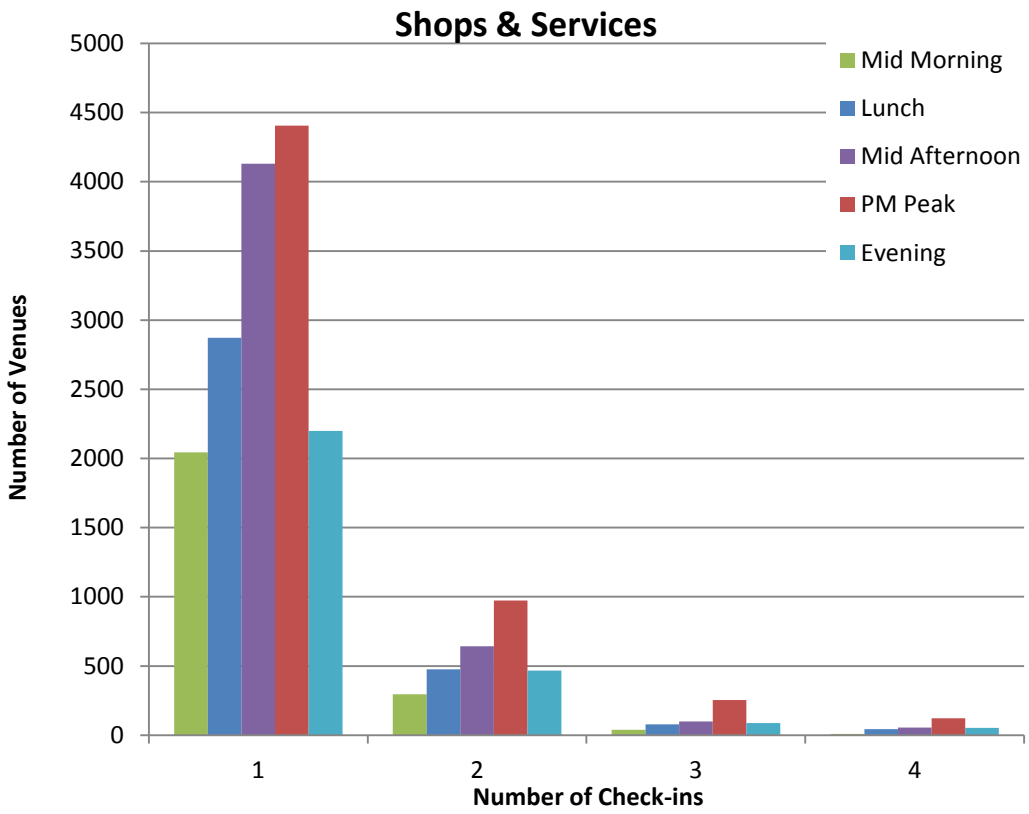


Figure 3.20: Shops & Services Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
Mid Morning	7	1	Tears of Joy Hot Sauce Shop	Monday
	7	1	Whole Foods Market	Thursday
Lunch	12	1	Whole Foods Market	Tuesday
	12	1	Whole Foods Market	Thursday
Mid Afternoon	22	1	Chevron	Wednesday
PM Peak	12	1	Barton Creek Square Mall	Friday
Evening	34	1	Waterloo Records	Tuesday

Table 3.11: Shops & Services Venues with the Most Check-ins.

Art & Entertainment venues are prominent in the P.M. Peak, and Evening time periods. Examining the P.M. Peak (Figure 3.21) time period first, the CBD has high

number of check-ins in both number of venues and intensity, understandably. Of interest to note, the Alamo Drafthouse, a local movie theater that is located in the lower left of the graphic, has high intensity during each day of the week except for Tuesdays. This trend is also visible in the Evening time period graphics (Appendix A) indicating Tuesdays are potentially not popular days for watching movies by Foursquare users. Figure 3.22 provides additional insight into check-in trend within the Arts & Entertainment category. There is an increase in single check-ins at venues from the P.M. Peak to the Evening time periods from to 513 to 563 venues. The venues with the largest number of check-ins within the Arts & Entertainment category include the previously mentioned Alamo Drafthouse for the P.M. Peak on Fridays and the Long Center for Performing Arts for the Evening time period on Wednesdays (Table 3. 12). The Long Center's check-ins likely relate to a specific performance or event being held at the venue.

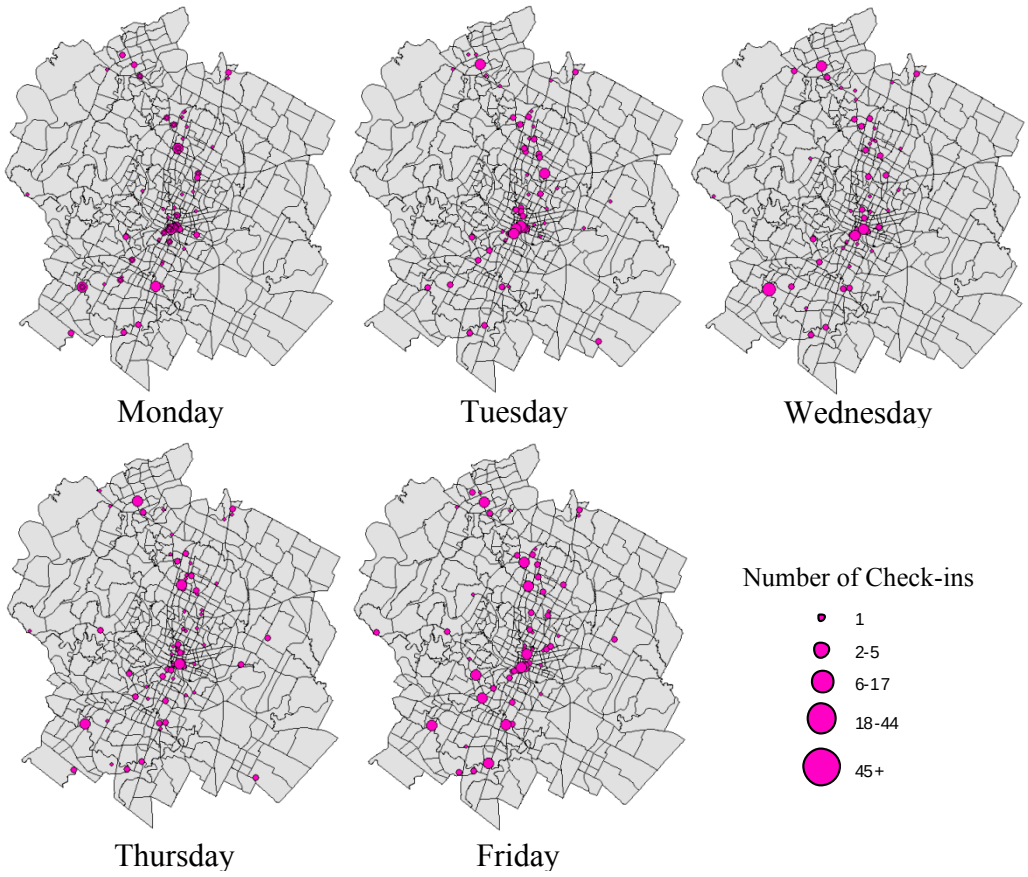


Figure 3.21: P.M. Peak Art & Entertainment Venue Check-ins

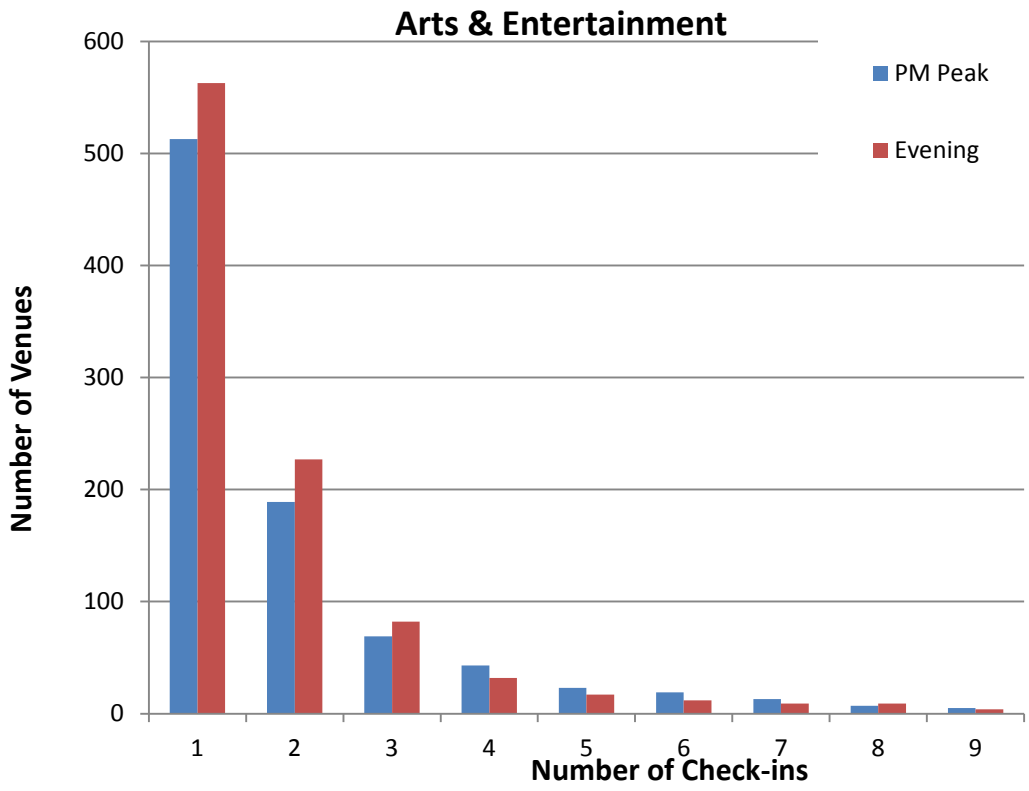


Figure 3.22: Art & Entertainment Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
PM Peak	18	1	Alamo Drafthouse Cinema @ Slaughter Lane	Friday
Evening	29	1	Long Center for the Performing Arts	Wednesday

Table 3.12: Art & Entertainment Venues with the Most Check-ins.

For the Nightlife Spots category, there are significant number of check-ins that occur within the CBD area where there are many bars and clubs located throughout the downtown. Figure 3.23 shows the trends in check-ins during the Evening time period, where Monday is has the least intensity of the weekdays. This trend is also seen in the P.M. Peak time period (Appendix A). For the Late Night time period, the trends are similar to the Evening time period just with a lesser number of venues. The largest

number of single check-in venues is found during the Evening time period with around 1500 venues (Figure 3.24, Table 3.13). This time period also has the most number of two check-ins per venue with 600 venues falling into this category.

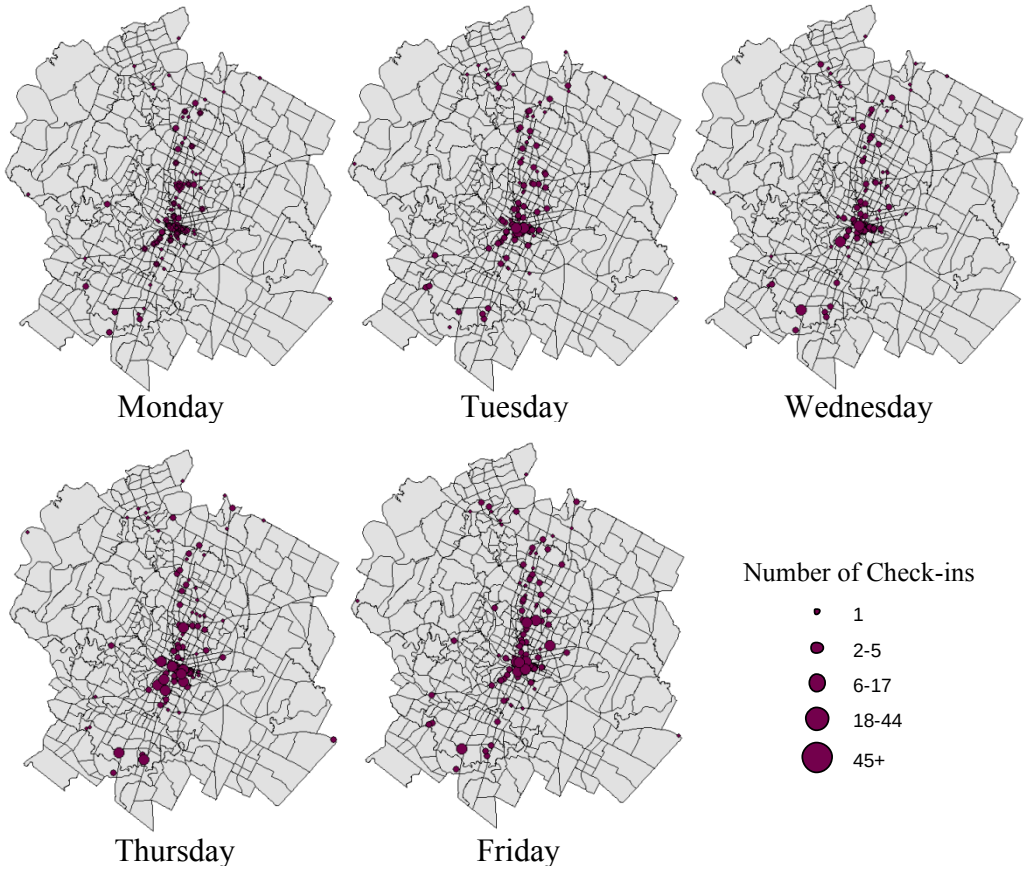


Figure 3.23: Evening Nightlife Spots Venue Check-ins

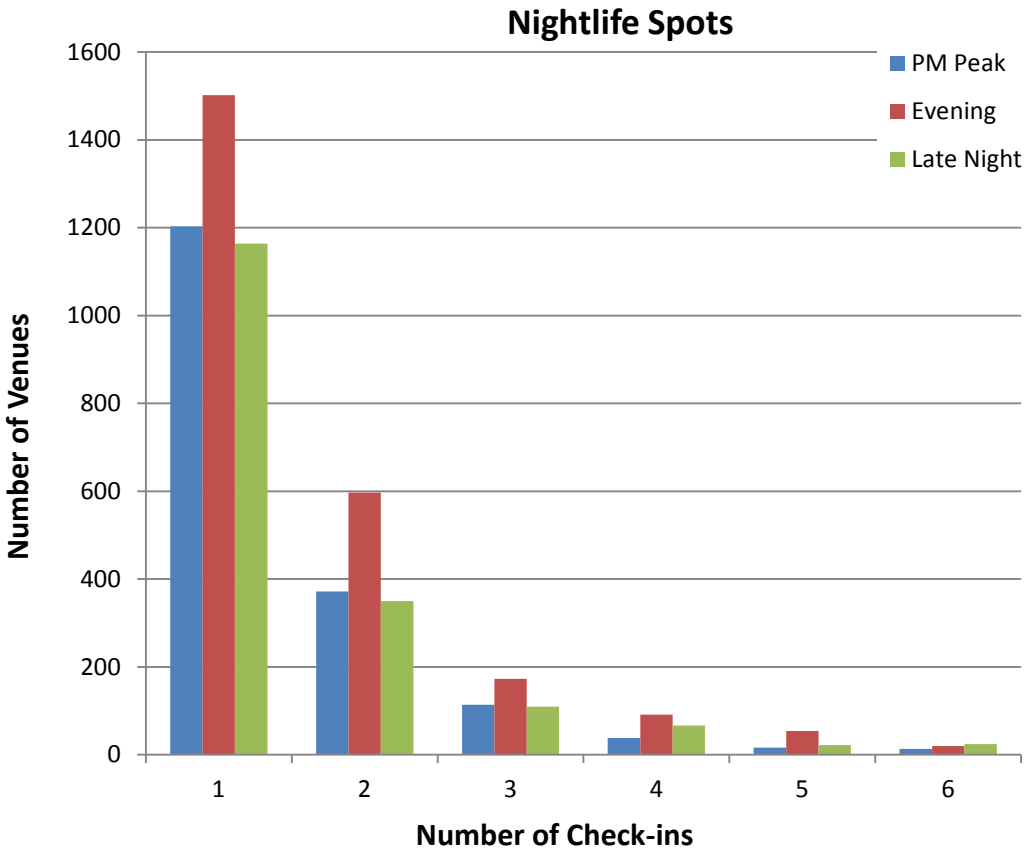


Figure 3.24: Nightlife Spots Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
PM Peak	13	1	The Palm Door	Tuesday
Evening	17	1	Maggie Mae's	Monday
Late Night	12	1	Barbarella	Wednesday

Table 3.13: Nightlife Spots Venues with the Most Check-ins.

The final category is the Residences category which was analyzed for the P.M. Peak, Evening, and Late Night time periods. Figure 3.25 provides a visual for the P.M. Peak time period, which has similar trends to the other time periods. From this graphic, the spatial distribution of residences throughout the study area can be seen. Additionally, the graphic shows some residences being checked-into consistently throughout the week,

which likely indicates that these check-ins are by individuals that reside in the location and are not visitors. For the Evening and Late Night time periods (Appendix A), check-ins follow similar trends to the P.M. Peak time period with a perceptible lessening of check-ins during the Late Night time period. The intensity of check-ins for residences only ranges from one to four check-ins per hour (Figure 3.26, Table 3.14) and likely relates to the number of individuals in the households that are over 13 and are Foursquare users.

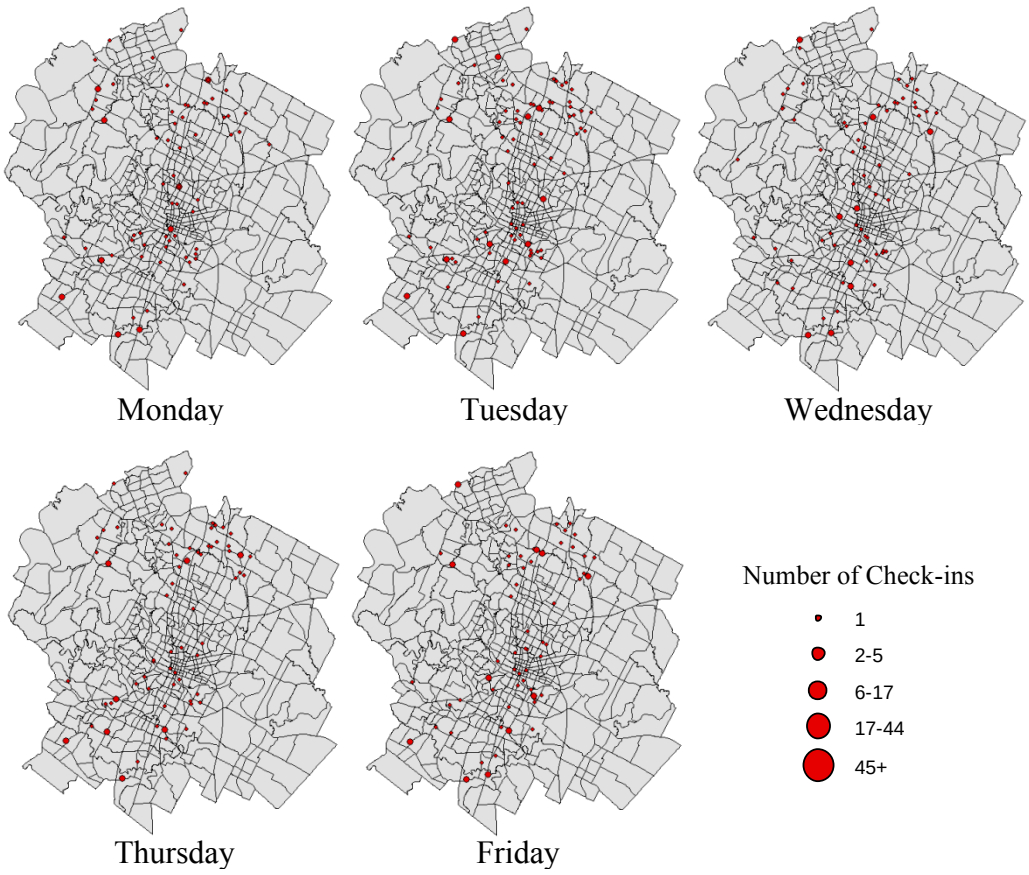


Figure 3.25: P.M. Peak Residence Venue Check-ins

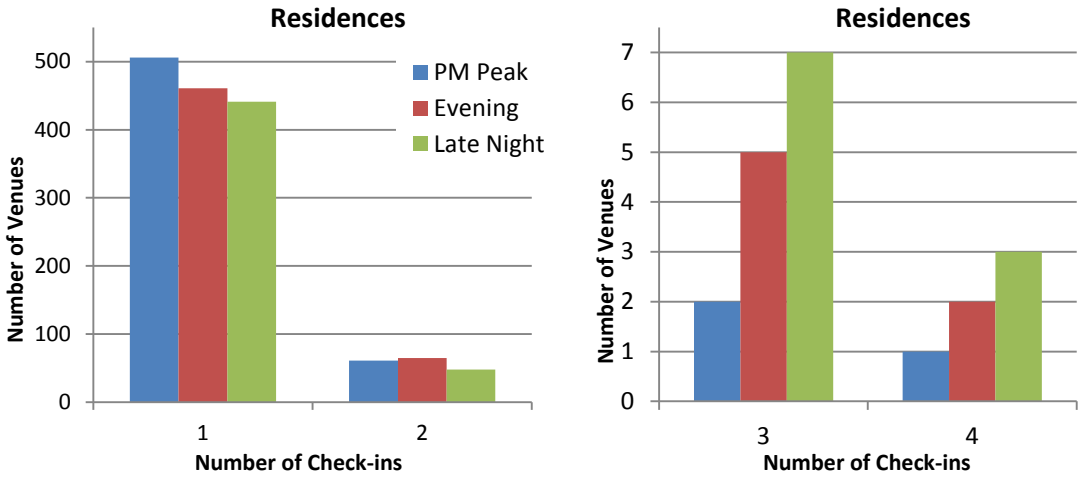


Figure 3.26: Residence Number of Venue by Check-in Amount

Time Period	No. of Check-ins	No. of Venues	Venue Name	Day of Week
PM Peak	4	1	Circle C Ranch	Tuesday
Evening	4	1	Berkshire South Congress	Tuesday
	4	1	Circle C Ranch	Wednesday
Late Night	4	1	Barton's Mill	Thursday
	4	1	Stassney Woods	Wednesday
	4	1	Onion Creek Lux Apts	Tuesday

Table 3.14: Residence Venues with the Most Check-ins.

Foursquare User Demographics

As noted above, Foursquare has over 55 million users throughout the world that have used the app at least once. Using comparative worldwide demographics for Foursquare (Chappell 2013), the case study area of Austin, TX (USCB, 2013, CLRSearch, 2012), and the US (Howden, 2011), the Foursquare data source was examined to determine if there were any notable limitations or biases. Figure 3.27 provides a visual examination of the comparison of age, gender, household income, and education. Based on this examination, the following characteristics were noted:

1.) The data source has a notable under representation from individuals under 17, which can be attributed to the minimum age of 13 restriction imposed by Foursquare. This is not necessarily a considerable limitation since these younger individuals are often passengers within the commuting public's vehicles (i.e., parents).
2.) Approximately 80% of the app users are between the age of 25 and 54, which indicates a skew toward this demographic that is not seen either in the Austin, TX nor US demographics. This may not be problematic, since the age range covers a majority of the working public, which is the focus of commuter's studies (Labor Force 2014).
3.) There is an over representation of females (65%) within the demographics of Foursquare, which may impact the value of the data.
4.) Over representation exists for household incomes of $25,000 to $74,999; this may partially be explained by the age demographic over representation within this category and with the "Some College" over representation that may limit income potential.
5.) While slight under representation exists for the $0 to $24,999, $100,000 to $149,999, and $150,000 or more categories, the latter two could be attributed to the similar under representation seen in the age demographics. Logically, this range of salary is traditionally experienced by individuals that are in the later stages of their career, who may be slow adopters to newer technologies and trends. The earlier category is more concerning, since it could speak to lower income individual's limited access to the technologies needed for the app (i.e., smartphones). However, recent research conducted by the Pew Research Internet Project

(Smith 2014) indicated that this barrier is being overcome in the US with approximately 50% of adults owning a smartphone, Table 3.15.

6.) Finally, the education demographic has a significant over representation within the "Some College" category. The statistics may relate to the income skew.

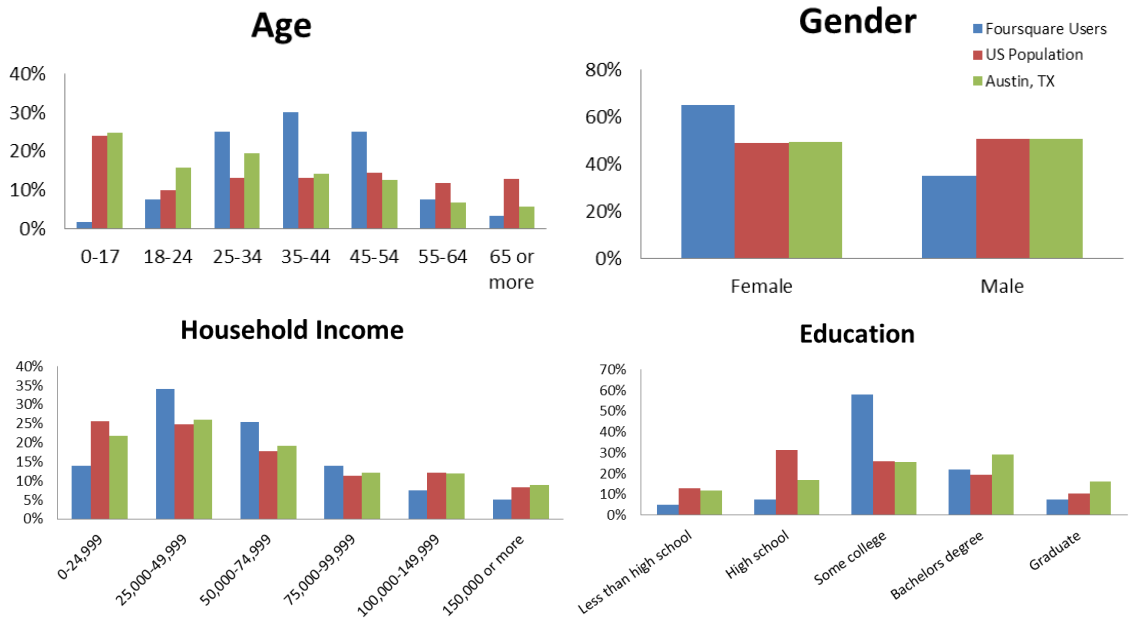


Figure 3.27: Comparative Demographics

Income Level (per year)	Ownership of Adults - Smartphone
Less than $30,000	47%
$30,000 - $49,999	53%
$50,000 - $74,999	61%
$75,000+	81%

Table 3.15: US Adult Smartphone Ownership by Income Level

Recently Brandon Gaille (2015) provided additional user facts concerning Foursquare including the popularity of the site in the US where the site is used numerically more often than any other country with 60% of all check-ins occurring in the US. This article noted that the highest per capita population usage was found in Venezuela, Singapore, Azerbaijan, and Belgium, respectively. Gaille noted that the average Foursquare users was likely male (60%), between the age of 18-29 (40%), have some to no college (83%), and make less than $50,000 per year (49%). These statistics vary from the data on Foursquare provided above, but could be attributed to the difference in years for the data, 2012 versus 2014. Other notable demographic statistics include:

1.) 13.2% adult Hispanics log onto Foursquare regularly.
2.) Only 2.8% Caucasians log into the site at least once per day.
3.) African-Americans are three times as likely to use Foursquare daily compared to Caucasians.
4.) Men are more likely to use Foursquare for checking-in at travel locations.
5.) Women are more likely to check-in at educational locations.
6.) 34% of the check-ins at beauty shops are by males.
7.) Approximately 4% of users will check-in to their own residence.

These trends were noted to possibly be from the site's worldwide influences, but also could be related to "an ethnic family-based element." This trend implies that the data source could be useful in targeting specific demographics. The study also asked how often Foursquare users checked –in at venues finding that 87% did so at least once a day.

Since the Austin Foursquare dataset does not have any user details, comparison to the Capital Area Metropolitan Planning (CAMPO) Travel Demand Model (TDM) population will require some assumptions. The first is that the individuals checking-in to

residential locations live there and are not visitors. The second is that each check-in relates to a single individual that acts as the head of household. Most of the residential locations included within the analysis are apartment or condo buildings, which house many individuals in one location, and complicates the analysis without these assumptions.

ArcGIS was again employed to create a relationship between the residential check-in data and the TAZ's demographical information from the *2005 to 2035 Plan Amendment 1110* and the *2010 TAZwDems* shapefiles. The *2005 to 2035 Plan Amendment 1110* file contains information on estimated population and employment for CAMPO. The data includes information on total TAZ population for a specified year (POPXX), number of households for a year (HHXX), and the average household size (HHSIZEXX). The *2010 TAZwDems* file contains information on TAZ population for 2010 (POP10), the median income for the TAZ (MedInc10), and the total employment for the TAZ (TotEmp10). Based on this joining, the following are some demographics on the individuals within the dataset:

Number of Unique Residential Venues – 170

Number of Check-ins – 3094

Average HHSize – 2.23

Mean Household Income (2010) - $52,539.06

The CAMPO dataset had the following related demographics:

Number of Households– 559,423

Average HHSize – 2.61

Median Household Income (2005) - $53,627

From the above details, the joined Foursquare dataset is slightly less in average house hold size and household income.

Table 3.16 provides a breakdown of income categories found from the joined Foursquare datasets and for the CAMPO dataset. While there is an over representation in the income categories of $20,000 to $35,000, $35,000 to $50,000, and $50,000 to $75,000, this may be due to the CAMPO data being reflective of the entire region and not just the study area. Additionally, the fact that most of the residential data is for apartments and condos and not houses could explain the under representation of the $75,000 + category. It is also a limitation of the dataset that the lower income category was significantly under represented. Finally, the Foursquare dataset is additionally limited by the lack of age, gender, and employment details, which would need to be addressed before the dataset could be used for planning purposes to ensure the data has biases that could be accounted for.

	Foursquare		CAMPO (2005)	
Income Groups	Number	Percentage	Number	Percentage
$0 to $20,000	139	4.49%	83785	14.98%
$20,000 to $35,000	726	23.46%	91857	16.42%
$$35,000 to $50,000	764	24.69%	89776	16.05%
$50,000 to $75,000	906	29.28%	120063	21.46%
$75,000 +	559	18.07%	175983	31.46%

Table 3.16: Foursquare Dataset Income Breakdown Comparison

Foursquare and Land Use

The final data analysis component examined the Foursquare dataset with respect to land uses for the study region. For this analysis, the 2010 Land Use shapefiles from the City of Austin were used to examine the land use composition of each TAZ. Within the 2010 Land Use files, there are 39 different codes used to describe each parcel within a TAZ. Each of these 39 codes was assigned to the Foursquare category when possible. This effort resulted in 22 codes that were assigned to the Commercial (Shops & Services)

Education (Colleges & Universities), Outdoor (Great Outdoors), Professional (Professional & Other Services), and Travel (Travel & Transport). For each of the 520 TAZs within the study area, the land uses were categorized into their relevant category and a total for each category was determined as well as a percentage. This effort was also done for the venues within each TAZ. Since the parcel level land use data has higher counts for each category, the percentages were compared between the two data sources to determine if there were significant areas of missing land uses from the Foursquare dataset. If the difference between percentages were found to be within 10%, the representation was considered acceptable. Table 3.17 shows the categorical breakdown with respect to the number of TAZs (and overall percentages) that had representative land uses from the Foursquare venues data. Based on the data in the table, only the Outdoor/Great Outdoors and Travel/Travel & Transport categories had a significant number of TAZs with representative land use venues. All other categories had less than 50% representation, indicating a potential limitation of the dataset. An examination to determine if any TAZs had all categories fall within the 10% cutoff was also conducted. Figure 3.28 provides a visual depiction of the location of the 51 (9.81%) TAZs that met this criteria and demonstrates that these TAZs are scattered throughout the study area. Finally, the Foursquare dataset has no mining categories represented, which is a land use category represented within the 2010 Land Use file. However, mining land use only existed in 18 of the 520 TAZs and was not considered critical for the datasets usage.

Land Use Category	No. of TAZs within 10%	% of TAZs
Commercial/Shops & Services	191	36.73%
Education/Colleges & Universities	185	35.58%
Outdoor/Great Outdoors	443	85.19%
Professional/Professional & Other Services	213	40.96%
Travel/Travel & Transport	426	81.92%

Table 3.17: Land Use Comparison Data

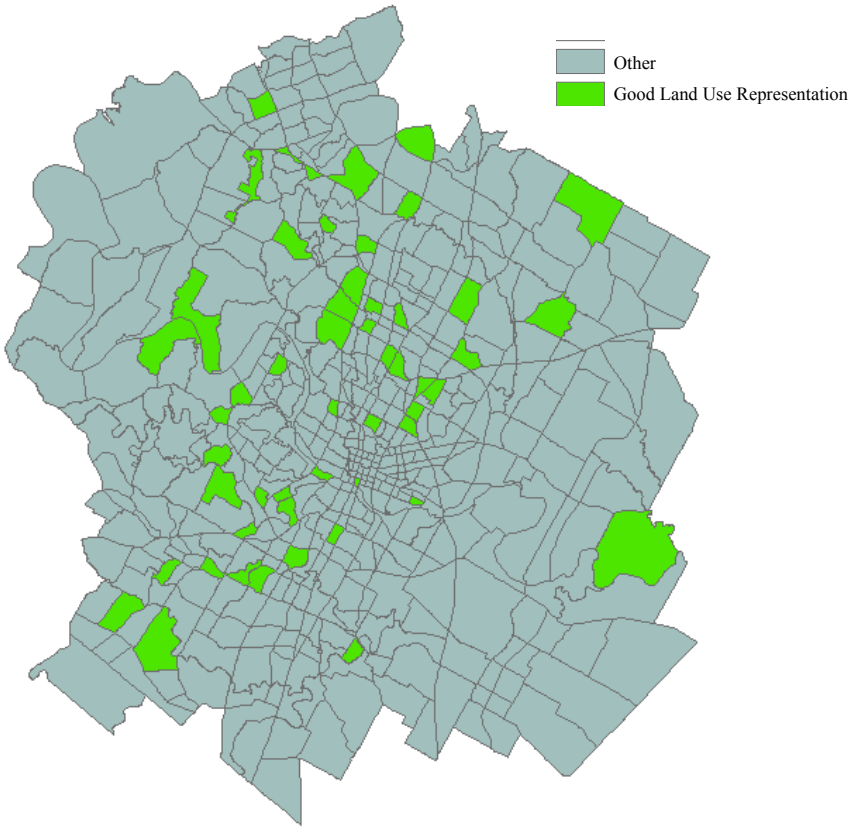


Figure 3.28: TAZs with Good Land Use Representation in Foursquare Data

CONCLUDING STATEMENTS

The above examination of the Foursquare dataset explored the types of check-ins with respect to time of day and day of week for each category type. This analysis presented the areas where the dataset showed strength and where there were limitations.

Additional analysis was performed to determine user demographics through the addition of a secondary data source. Finally, the data was compared to existing land use data to determine how representative the data sample was for the overall study area. Based on this final analysis, the sample fell short in the majority of categories and TAZs throughout the study area.

The need for external data sources indicates that the Foursquare data could only act as supplementary data to data collected by traditional household surveys or technology-based methods. Due to the data's relative richness with respect to venues, the data source should be able to be used for activity-based planning, especially if the data came from a data vendor and thus would have user identification affiliated with each check-in allowing for user tracking. Additional concluding remarks will be provided within the Conclusion chapter of this dissertation.

Chapter 4: Methodology

This chapter will present the methodology used to investigate how location-based social networking data can be used for transportation demand planning. The chapter builds upon the discussions included within the literature review and presents the novel approach of using many-to-many modeling, specifically the peer-to-peer model, for transportation planning. This chapter begins by presenting details on the mathematical components of the peer-to-peer model. It then presents details on the doubly-constrained gravity model, and the friction functions used within the exploration of both models. The final sections discuss the calibration and validation methods used for the models.

ORIGIN-DESTINATION MODELING

Within this section, a detailed explanation of how the location-based social networking (LBSN) data, discussed in depth in the previous chapter, will be used in conjunction with the two different proposed models. The first component will describe the trip generation modeling using LBSN data as well as the data preparation processes undertaken. The second component will present the two trip distribution modeling methodologies that will be examined: the novel peer-to-peer modeling and the doubly-constrained gravity model. This second component will include details on the multiple friction functions that will be used within the analysis when comparing the two models to each other and the existing CAMPO model.

Trip Generation Using Location-based Social Networking Data

Previous chapters discussed the location-based social networking and, specifically, Foursquare data from both a general and Austin, TX view point. In this section the methodology used to transform check-ins into productions and attractions for the creation of the origin-destination matrix estimation is described.

Foursquare Data Collection

As discussed in the previous chapters, Foursquare was selected for use due to its confirmed vendor locations and good spatiotemporal coverage. Data collection was begun by first identifying venues within the study area of Austin and then running a trolling algorithm that created a snapshot of the total number of check-ins and unique users for each venue. These snapshots were done at intervals of 45-50 minutes and were used to create an hourly rate through the following calculation:

$$C_{hr} = \left(\frac{x_2 - x_1}{t_2 - t_1}\right) * 60 \quad (Eqn. 4.1)$$

Where

- C_{hr} is the check-in rate per hour;
- x_i is the number of check-ins collected between the two time intervals;
- t_i is the time in minutes for each of the time intervals.

Additionally, the method collected each venues unique ID, name, category, and GPS coordinates. The trolling algorithm collected data 24 hours a day for a three week period, resulting in over 6 million check-ins.

Using the nine first tier categories for venue, categories assigned to venues within the dataset were confirmed and for those venues without an assigned category, a keyword search was performed assigning an appropriate category when possible. Table 4.1 provides a list of the first tier categories and the number of second tier categories within each that are currently available for venue identification; it should be noted that the Events category was not an option at the time of the data collection. Venues where no category could be assigned or confirmed were removed from the study. The number of venues that fell into this group totaled 1538 which corresponds to 8% of all of the venues within the study, but only represented 103,692 check-ins representing 1.5% of all checks.

After the removal of the non-categorized venues, the remaining venues were assigned to CAMPO TAZs through a GIS mapping technique that uses the shapefile for the CAMPO TAZs and the GPS coordinates for each venue.

First Tier	Number of Second Tier
Arts & Entertainment	30
College & University	13
Event*	8
Food	120
Nightlife Spot	22
Outdoors & Recreation	47
Professional & Other Places	29
Residence	5
Shop & Service	123
Travel & Transport	30

Table 4.1: Foursquare Categories for Classification.

*New category not included in original data collection.

Trip Generation Model Methodology

To determine the trip distributions, the previous check-in Foursquare data was separated into weekday and weekend check-ins and the Foursquare weekday check-in dataset was then aggregated to the TAZ level resulting in a total number of check-ins per TAZ per category. The categories used were the nine first tier categories and the "unknown" category resulting in a 10x1462 matrix called "weekday". It is important to note that the length of this matrix is based on the 1462 TAZs within the CAMPO region. Using this "weekday" matrix, MATLAB code was run to create a check-in matrix, called "checkins," that contained only the data for the 520 TAZs that are included within the study area. The code for this effort is provided in Appendix B.

After creating the "checkins" matrix, the data was further manipulated to create the productions and attractions for each TAZ. This was done within the coding by referencing each row of the weekend matrix to its corresponding category:

- Row 1 = Professional Locations
- Row 2 = Shops & Services
- Row 3 = Colleges & Universities
- Row 4 = Residences
- Row 5 = Travel & Transport
- Row 6 = Arts & Entertainment
- Row 7 = Food
- Row 8 = Nightlife Spots
- Row 9 = Outdoors & Recreation

The tenth row was comprised of the unknown venue check-ins and was not included within this study per the previous discussion. Productions were then calculated using the following formula:

$$O_i = \gamma * x_i \quad (Eqn. 4.2)$$

Where

- O_i is the productions from TAZ i
- x_i is the total check-ins within TAZ i, and is found using the following formula:

$$x_i = \sum_i (Professional + Shops + Universities + Residence + Travelspots + Entertainment + Food + Nightlife + Outdoor) \quad (Eqn. 4.3)$$

- γ is the adjustment factor used to suitably scale the trip productions to the Foursquare check-ins.

The attractions were calculated using the check-in data and the following formula:

$$D_j = (\varepsilon * x_i) + \frac{{x_i}^{\eta}}{\sum_i {x_i}^{\eta}} \sum_i (\gamma - \varepsilon) x_i \quad (Eqn. 4.4)$$

Where

- D_j is the attractions to TAZ j
- x_i is as defined above
- ε is the adjustment factor used to scale the trip attractions to the Foursquare check-ins
- γ is as defined above
- η is the weighting factor assigned to the total check-ins within the residual term that guarantees the total productions equal the total attractions via the following formula:

$$residuals = \frac{{x_i}^{\eta}}{\sum_i {x_i}^{\eta}} \quad (Eqn. 4.5)$$

For the Austin Foursquare data, the adjustments and weighting factors described above were found using a genetic optimization algorithm in the 2014 Jin et al. study and are provided in Table 4.2 below. The genetic optimization algorithm will be discussed in more detail in a subsequent section of this chapter.

Factor	Numerical Value
γ	1.14301
ε	0.66967
η	0.21198

Table 4.2: Trip Generation Factors.

Trip Distribution Using Location-based Social Networking Data

Past efforts have considered two variations of gravity models to determine the origin-destination (O-D) model from location-based social networking data. The 2011 study by F. Yang et al. (2014) and the 2013 study by Jin et al. used a singly- constrained version of the gravity model. These efforts provided proof of concept and were followed by the efforts of 2014 Jin et al. and thesis by this author (2013), which explored the doubly-constrained gravity model.

This section will use the previously described trip generation data to create origin-destination matrices using two different methodologies: the doubly-constrained gravity model and novel peer-to-peer model. First, the doubly-constrained model will be presented as will the two-regime friction function exploration. This friction function exploration is done to determine how sensitive the doubly-constrained model is to the varying functions and for comparison purposes with respect to the newly proposed peer-to-peer model and the existing CAMPO model. The newly proposed peer-to-peer model will then be presented and explored with respect to the friction functions.

Friction Functions

The first step in the trip distribution process is to determine and calculate the friction function for the study area. The friction function describes the travel impedance from the current TAZ to each destination TAZ and for travel within itself. To begin this process, centroid GPS coordinates for each TAZ were used in the calculations for the Manhattan distances between TAZs using the following equation:

$$d_{ij} = \left(\left|latitude_i - latitude_j\right| + \left|longitude_i - logitude_j\right|\right) * 100 \ (Eqn. 4.6)$$

Where

- d_{ij} is the distance between two TAZs in miles

- i and j represent the starting and ending TAZ respectively.

The resulting i x j matrix values do not account for the intrazonal travel and additional manipulation is needed. To create these values, an identity matrix, *eye*, was created of size i x j, which was then used in the following formula:

$$tripdist_{ij} = d_{ij} + (5 * eye) \quad (Eqn. 4.7)$$

The use of a value of five was selected to accommodate the travel within each TAZ.

The study by Jin et al. (2013) found that a two-regime friction function accounted for the differentiation between short and long distance trip trends as found within the CAMPO data. This study examined the linear, negative exponential, and gamma functions (Equations 4.8 through 4.10) for use with in the two-regime function (Equation 4.11) in the same was the Yang et al. study did.

Linear: $$F_{ij} = \delta + \theta(d_{ij}) \quad (Eqn. 4.8)$$

Negative exponential: $$F_{ij} = \delta e^{-\theta(d_{ij})} \quad (Eqn. 4.9)$$

Gamma: $$F_{ij} = \delta(d_{ij}^{-\theta})e^{-\lambda(d_{ij})} \quad (Eqn. 4.10)$$

Two-regime friction function:

$$F_{ij}(d_{ij}) = F_{ij}{}^{(s)}(d_{ij})I_{d_{ij} \leq T_d} + F_{ij}{}^{(l)}(d_{ij})I_{d_{ij} > T_d} \quad (Eqn. 4.11)$$

Where

- δ is the positive scaling factor controlling the overall range of function values
- θ is the positive or negative constant value which affects the distribution of shorter trips

- λ is the parameter of friction relating to the efficiency of the transportation system between two locations, is always negative, and can impacts the distribution of longer trips
- d_{ij} is as defined above
- $I_{d_{ij} \leq T_d}$ and $I_{d_{ij} > T_d}$ are indicator functions for a logic clause that gives a value of 1 when true and 0 otherwise
- the superscripts *s* and *l* indicate short-distance and long-distance trip regime, respectively
- T_d is the threshold to determine the regime.

While, only one combination formula was used within the singly- and doubly-constrained comparison study by Jin et al. (2014), this effort examines all possible combinations with respect to the doubly-constrained model in an effort to better understand where each model excels and where there are deficiencies. Equations 4.12 through 4.20 show the two-regime friction function formula combinations that result in the nine variations that are examined within this dissertation.

Linear-Linear: $$F_{ij}(d_{ij}) = \begin{cases} \delta + \theta(d_{ij}) & d_{ij} \leq \phi \\ \delta_1 + \theta_1(d_{ij}) & d_{ij} > \phi \end{cases} \quad (Eqn.\ 4.12)$$

Linear-Negative Exponential: $$F_{ij}(d_{ij}) = \begin{cases} \delta + \theta(d_{ij}) & d_{ij} \leq \phi \\ \delta_1 e^{-\theta_1(d_{ij})} & d_{ij} > \phi \end{cases} \quad (Eqn.\ 4.13)$$

Linear-Gamma: $$F_{ij}(d_{ij}) = \begin{cases} \delta + \theta(d_{ij}) & d_{ij} \leq \phi \\ \delta_1 \left(d_{ij}^{-\theta_1}\right) e^{-\lambda_1(d_{ij})} & d_{ij} > \phi \end{cases} \quad (Eqn.\ 4.14)$$

Negative Exponential-Linear: $$F_{ij}(d_{ij}) = \begin{cases} \delta e^{-\theta(d_{ij})} & d_{ij} \leq \phi \\ \delta_1 + \theta_1(d_{ij}) & d_{ij} > \phi \end{cases} \quad (Eqn.\ 4.15)$$

Negative Exponential-Negative Exponential: $$F_{ij}(d_{ij}) = \begin{cases} \delta e^{-\theta(d_{ij})} & d_{ij} \leq \phi \\ \delta_1 e^{-\theta_1(d_{ij})} & d_{ij} > \phi \end{cases} \quad (Eqn.\ 4.16)$$

Negative Exponential-Gamma: $$F_{ij}(d_{ij}) = \begin{cases} \delta e^{-\theta(d_{ij})} & d_{ij} \leq \phi \\ \delta_1 \left(d_{ij}^{-\theta_1}\right) e^{-\lambda_1(d_{ij})} & d_{ij} > \phi \end{cases} \quad (Eqn. 4.17)$$

Gamma-Linear: $$F_{ij}(d_{ij}) = \begin{cases} \delta(d_{ij}^{-\theta}) e^{-\lambda(d_{ij})} & d_{ij} \leq \phi \\ \delta_1 + \theta_1(d_{ij}) & d_{ij} > \phi \end{cases} \quad (Eqn. 4.18)$$

Gamma-Negative Exponential: $$F_{ij}(d_{ij}) = \begin{cases} \delta(d_{ij}^{-\theta}) e^{-\lambda(d_{ij})} & d_{ij} \leq \phi \\ \delta_1 e^{-\theta_1(d_{ij})} & d_{ij} > \phi \end{cases} \quad (Eqn. 4.19)$$

Gamma-Gamma: $$F_{ij}(d_{ij}) = \begin{cases} \delta(d_{ij}^{-\theta}) e^{-\lambda(d_{ij})} & d_{ij} \leq \phi \\ \delta_1 \left(d_{ij}^{-\theta_1}\right) e^{-\lambda_1(d_{ij})} & d_{ij} > \phi \end{cases} \quad (Eqn. 4.20)$$

Where

- ϕ represents the cut off value that differentiates a short distance trip from a long distance trip and will be found using a genetic algorithm

As with all of the factors used within this dissertation, the friction function equation factors, $\delta, \delta_1, \theta, \theta_1, \lambda,$ and λ_1, will also be found using a genetic algorithm. A sample of the two-regime friction function MATLAB coding can be found in Appendix B.

Doubly-Constrained Gravity Travel Demand Model

As noted above, previous efforts in the area of location-based social networking and transportation planning have considered the singly- and doubly-constrained gravity models for origin-destination matrix estimation. With respect to the doubly-constrained gravity model, previous work examined the doubly-constrained gravity model with only one version of the two-regime friction functions (Jin et al. 2014, Cebelak 2013). Within this dissertation, the previous work will be expanded to examine the doubly-constrained gravity model with respect to the nine variations of friction functions as identified in the previous work of Jin et al. 2013. An origin-destination matrix will be created and used for comparison using each of these friction functions.

As discussed in the literature review chapter, the doubly-constrained gravity model is based off of the Newtonian gravitational law and formula, and builds off of the singly-constrained gravity model. Equation 4.21 through 4.23 provides the mathematical formulations for the doubly-constrained model, which incorporates balancing factors for the productions and attractions defined by β_i and α_j respectively.

Doubly-Constrained Modeling:

$$T_{ij} = \beta_i * O_i * \alpha_j * D_j * f(c_{ij}) \quad (Eqn. 4.21)$$

$$\beta_i = \frac{1}{\sum_j \alpha_j * A_j * f(c_{ij})} \quad (Eqn. 4.22)$$

$$\alpha_j = \frac{1}{\sum_i \beta_i * P_i * f(c_{ij})} \quad (Eqn. 4.23)$$

Where

- O_i is as defined above
- D_j is as defined above,
- β_i is the balancing factor for the productions, O_i
- α_j is the balancing factor for the attractions, D_j
- $f(c_{ij})$ is the friction function used as described in the previous section

Balancing factors β_i and α_j are found using an iterative process that uses the following steps:

1. Using values of one for β_i and α_j, and initial T_{ij} matrix is found.
2. An initial current difference is set to zero, an initial previous difference is set to one, and the step count is set to zero.

3. The absolute difference between the previous and current differences is calculated and while this value is greater than the set threshold and the step count is less than or equal to 20, steps 4 through 10 are repeated.
4. The total productions and attractions are calculated from the initial T_{ij} matrix.
5. The current β_i and α_j are renamed as previous β_i and previous α_j, respectively.
6. A new β_i and α_j is calculated using the created total production and attraction values created in step 4 and the Equations 4.22 and 4.23 from above.
7. Using the β_i and α_j created in step 5, a new T_{ij} matrix is created using Equation 4.21.
8. The current difference is then set to the previous distance and a new current difference is calculated using the following formula:

$$current\ difference = max\begin{pmatrix} max|\ \alpha_j - prev\ \alpha_j| \\ max|\ \beta_i - prev\ \beta_i| \end{pmatrix} \quad (Eqn.\ 4.24)$$

9. The step count is increased by one.
10. Return to step 3 until the threshold is met and a final T_{ij} matrix is created.

After the final T_{ij} matrix for each of the study friction functions is created additional model calibration is still needed and will be discussed in the section that follows the peer-to-peer modeling discussion. Sample code for the doubly-constrained model can be found in Appendix B.

Peer-to-Peer Travel Demand Model

As mentioned in the literature review section, peer-to-peer (P2P) modeling can be categorized as structured, unstructured, as well as hybrid networks. Structured overlay networks have a protocol that enables the any node to search the network with efficiency for a resource, commonly through an implemented distributed hash table (DHT). DHT assigns ownership of each resource to a particular peer that can then search for the resources within the network using the hash table (Ranjan, Harwood, and Buyya 2008). The hash table is comprised of hash functions that are used to compute an index into an array of buckets from which the correct value can be found (Hash Table 2014), Figure 4.1. According to Naor and Wieder (2007) this layout structure gives the network its efficiency, but requires that every node must maintain a list of neighbors that satisfy specific criteria thus making them less robust for networks with numerous nodes entering and leaving the network (Li, Liu, and Vasilakos 2009).

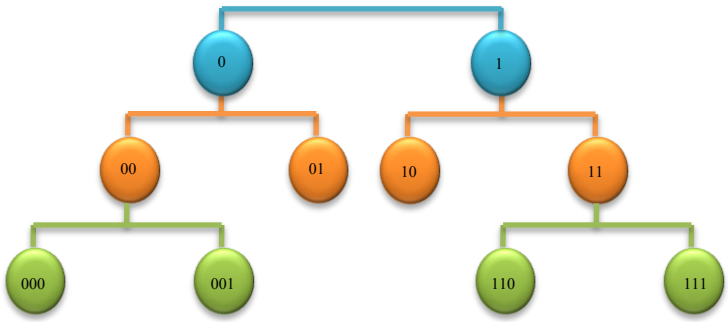


Figure 4.1: Example of a DHT Overlay

While structured networks have a specific layout/structure for their networks, unstructured networks do not enforce any particular structure on the overlay network but rather from forming random connections between nodes (Filali et al. 2011). These networks are easy to build and allow for optimization to occur locally for different

regions of the overlay (Chervenak and Bharathi 2008). Within this structure all peer have the same role leading to the more robust nature of the network with respect to numerous nodes entering and leaving the network (Jin and Chan 2010). One of the limitations to the unstructured networks is the need to query throughout the entire network for a particular resource leading to high traffic within the network and may not lead to the resolution of all search queries.

For hybrid networks, a combination of P2P and client-server models exist often with a central server that aids peers in finding one another (Darlagiannis 2005). These models often have better performance than traditional pure structured or unstructured networks due to the need for a centralized system for certain functions (i.e., searches) (Yang and Garcia-Molina 2001).

The formulation of the P2P demand estimation model uses the study by Xu et al. (2014) as its basis. In the study by Xu et al., the authors define individual users of the computer network as peers with the traffic generated due to uploading and downloading in a structured network. The authors define different types of peers based on user behavior within their study:

- Seeds – peers that only upload data,
- Free-riders – peers that only download data, and
- Leechers – peers that upload and download data, and have a preference toward those peers that have uploaded to them previously.

For this study, we consider venues as peers, uploading rates as attraction rates, downloading rates as production rates, and leechers as the user behavior type.

P2P systems can have tremendous numbers of peers leading to the need for aggregation. The aggregated P2P traffic matrices model formulation within the Xu et al. study is given in Equation 4.25:

$$X_{ij}(t) = K\frac{\mu_i(t)\mu_j(t)}{(d_{ij})^s}U_i(t)T_j(t) \quad (Eqn.4.25)$$

Where

- X_{ij} is the computer traffic between peer node *i* and peer node *j* during time interval *t*
- K is constant used to adjust the estimation scale
- $\mu_n(t)$ is the population ratio with *n* representing the peer node
- d_{ij} is the network distance between peer node *i* and peer node *j*
- *s* is the weighting factor assigned to the network distance
- $U_i(t)$ is the total uploading volume of the P2P traffic within zone *i*
- $T_i(t)$ is the total downloading volume of P2P traffic within zone *j*
- $\mu_n(t)$ is the population ratio and can be found using the following equation:

$$\mu_i(t) = \frac{|h_i^k|}{N} \quad (Eqn.4.26)$$

Where

- h_i^k is population of peers in the aggregated cluster *k*
- *N* is the total number of peers within the system

In terms of transportation planning, the above equations can be rewritten into terms of origin-destination matrix variables as follows:

Peer-to-Peer Modeling:

$$T_{ij}(t) = K\frac{\mu_i(t)\mu_j(t)}{(d_{ij})^s}D_i(t)O_j(t) \quad (Eqn.4.27)$$

Where

- $T_{ij}(t)$ is equivalent to the $X_{ij}(t)$ term representing the movement between aggregated peer node *i* (in this analysis TAZ *i*) and aggregated peer node *j* (TAZ *j*) during time *t*.
- $D_i(t)$ is equivalent to the $U_i(t)$ representing the attractions to TAZ *i* and will be calculated as shown in Equation 4.4
- $O_j(t)$ is equivalent to $T_j(t)$ representing the productions from TAZ *j* and will be calculate as shown in Equation 4.2
- d_{ij} is the two-regime friction function as described in the previous section
- K is the estimation scale constant used to calibrate the model and is found through the genetic algorithm
- s is the distance weighting factor, which is applied to the friction function and is found through the genetic algorithm
- $\mu_n(t)$ is the population ratio that creates a relationship between the number of venues within each TAZ to the total number of venues within the study area

While this analysis will not account for time, it is worth noting that this model does have the capabilities to account for the time feature making a dynamic origin-destination model vary computationally comprehensible.

Compared to the doubly-constrained gravity model, the P2P model methodology is significantly less cumbersome. The steps undertaken for this model are as follows:

1. The venue populations ratios are created using the following formula:

$$\mu_n(t) = \frac{TAZ_n\ venues}{\sum_1^n TAZ_n\ venues} \quad (Eqn. 4.28)$$

2. A venue population ratio matrix, "venues," is created to account for the $\mu_i(t) * \mu_j(t)$ component of the P2P equation for all TAZ to TAZ trips and uses the following equation for intra-zonal trips:

$$\mu_{ii}(t) = \left(\frac{TAZ_i\ venues}{\sum_1^n TAZ_n\ venues}\right) * \left(\frac{TAZ_i\ venues - 1}{\sum_1^n TAZ_n\ venues}\right) \quad (Eqn. 4.29)$$

3. This venue matrix is multiplied by the K factor to create a new venue matrix
4. The two-regime friction function is raised to the s distance weight factor to create a new friction matrix.
5. The venue matrix from step 3 is then divided by the friction function from step 4
6. An initial T_{ij} matrix is then found by multiplying the resulting step 5 matrix by the productions and attractions calculated within the trip generation process.

After the initial T_{ij} matrix from step 6 for each of the study friction functions is created additional model calibration is still needed and will be discussed in the section that follows. Sample code for the P2P model can be found in Appendix B.

MODEL CALIBRATION

The previous sections described the methodology employed to create the trip generations and distributions often referencing the use of a genetic algorithm for the determination of various factors used within the calculations. This section will describe the genetic algorithm that was engaged for the calibration of each model.

Genetic Algorithms

Genetic algorithms are an optimization strategy that mimics biological evolution principles through the repeated modification of a population of individual points using rules modeled on gene combinations used in reproduction (MATLAB 2015). In general, the algorithm selects random "individuals" from a current population of candidate solutions to be used as "parents" for the next generations "children." Between two generations "individuals" are allowed to "mutate" via the addition of a random vector from a Gaussian distribution. The process is iterative and evaluates the fitness of every individual within the population after each mutation. This process of random "individual" selection and "mutation" is repeated and the population "evolves" toward an optimal solution, which is "attained" either by exceeding the fitness threshold or by maximizing the number of generations.

This optimization strategy was selected for the improved chances of finding a global solution due to the algorithm's random nature. One of the limitations of the algorithm is its computational inefficiencies; however, the algorithms are extremely flexible (Charypar and Nagel 2005), thus their selection for use within this dissertation. Table 4.3 provides a comparison between the classic and genetic algorithms noting the differences between them.

Genetic Algorithm	**Classic Algorithm**
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 4.3: Algorithm Comparison (MATLAB, 2013).

MATLAB has a built-in tool for optimization that includes a genetic algorithm. This genetic algorithm has the following syntax:

$$[x\ fval] = ga(@fitnessfun, nvars, options)$$

Where

- $@fitnessfun$ references the fitness function to be evaluated
- $nvars$ is the number of independent variables for the fitness function
- $options$ contains the options that are customizable for the genetic algorithm
- x is point at which the final value is attained
- $fval$ is the final value of the fitness function

Further details with respect to the particulars of this analysis will be provided after the brief general discussion on how the algorithm works with in MATLAB.

The steps that MATLAB's algorithm undertakes are:

1. An initial random population is created
2. Each member of the population is scored via its fitness value
3. The raw fitness scores are scaled to a more usable range of values
4. "Parent" members are selected based on fitness
5. Identification of "elite" individuals is done based on low fitness values and these individuals are passed on to the next population
6. "Children" are produced from the "Parents" from step 4 via a mutation of a single parent or by crossover, which is the combining of the vector entries of a pair of "parents"
7. A new sequence of populations is created by replacement of the current population with the "children," "elite," and "parents"

For the selection of "parents," the default option, which is used for this dissertation, is the stochastic uniform. This method lays out a line where each parent corresponds to a section of the line that is proportional to its scaled value. The algorithm uses a uniform random number less than the step size to begin the process that moves along this line in equally sized steps allocating a parent from each section that is landed on. For the crossover process, a default crossover function randomly selects an entry, or, in terms of biology, a "gene" from the same location from one of the "parents" and assigns it to the same location for the "child." For the mutation process, a random vector from a Gaussian distribution is added to the "parent." For a genetic algorithm to be effective, both crossover and mutations are needed. The crossover function extracts the "best genes" from the "parents" to potentially create "superior children," while the mutation function adds diversity to the population and increases the likelihood that better fitness values will be attained from the new population.

Within the MATLAB genetic algorithm, there are a number of options that can be defined by the user. Figure 4.2 shows the default values used within the code. Within this analysis, the number of generations and the termination tolerance value were adjusted from the default values. This was done to assist in the algorithm's ability to converge in a reasonable amount of time without compromising the resulting optimized values. The number of generations was set to 100 and the termination tolerance was set to 1.000e-03. The code used to make this change can be found in Appendix B.

```
options =

        PopulationType: 'doubleVector'
          PopInitRange: []
        PopulationSize: '50 when numberOfVariables <= 5, else 200'
            EliteCount: '0.05*PopulationSize'
     CrossoverFraction: 0.8000
        ParetoFraction: []
    MigrationDirection: 'forward'
     MigrationInterval: 20
     MigrationFraction: 0.2000
           Generations: '100*numberOfVariables'
             TimeLimit: Inf
          FitnessLimit: -Inf
         StallGenLimit: 50
             StallTest: 'averageChange'
        StallTimeLimit: Inf
                TolFun: 1.0000e-06
                TolCon: 1.0000e-03
     InitialPopulation: []
         InitialScores: []
    NonlinConAlgorithm: 'auglag'
        InitialPenalty: 10
         PenaltyFactor: 100
          PlotInterval: 1
           CreationFcn: @gacreationuniform
     FitnessScalingFcn: @fitscalingrank
          SelectionFcn: @selectionstochunif
          CrossoverFcn: @crossoverscattered
           MutationFcn: {[@mutationgaussian]  [1]  [1]}
    DistanceMeasureFcn: []
             HybridFcn: []
               Display: 'final'
              PlotFcns: []
            OutputFcns: []
            Vectorized: 'off'
           UseParallel: 0
```

Figure 4.2: MATLAB Genetic Algorithm Default Options

For this analysis, the genetic algorithm syntax used was:

$$[params, fav, exitflag, output] = ga(@(x)eva(x, checkins, 'CR', totalOD, tripdist, n, alg, venues), nVars, [\ \ ], [\ \ ], [\ \ ], [\ \ ], lowerBdS, upperBds)$$

Where

- $@(x)eva(x, checkins,' CR', totalOD, tripdist, n, alg, venues)$ is the fitness function which references a function that will be discussed further below. The requirement for the fitness function is that it should be able to accept a row vector of length $nvars$ and return a scalar value.
- $nVars$ is the number of variables to be analyzed and uses positive integers
- [ ], [ ], [ ], [ ] are indicators that no linear inequalities exist for the following characteristics, respectively:
    - A is a matrix for linear inequalities with the constraint of the form $A * x \leq b$
    - b is a vector for linear inequalities with the constraint of the form $A * x \leq b$
    - Aeq is a matrix for linear inequalities with the constraint of the form $Aeq * x \leq beq$
    - beq is a matrix for linear inequalities with the constraint of the form $Aeq * x \leq beq$
- $lowerBdS$ is the vector of lower bounds that the genetic algorithm enforces to ensure the iterations stay above. If no lower bound exists, this value can be set to -Infinity
- $upperBds$ is the vector of upper bounds that the genetic algorithm enforces to ensure the iterations stay above. If no upper bound exists, this value can be set to Infinity
- $[params, fav, exitflag, output]$ are the results from the genetic algorithm and are described as follows:

- $params$ is a vector of best points, or solutions, as located by the genetic algorithm during its iterations based on the conditions set. This vector's length is customizable by the user and is defined by x in the above equation.
- fav returns the fitness function evaluated for the $params$.
- $exitflag$ returns an integer that relates to the reason why the algorithm was terminated. These values and meanings are presented in Table 4.4.
- $output$ returns details about the algorithm's performance and contains the following fields:
    - $problemtype$- a string that describes the type of problem as one of the following: unconstrained, bound constraints, linear constraints, nonlinear constraints, or integer constraints.
    - $rngstate$- the state of the random number generator just prior to the start of the algorithm. These values can be used for reproduction of the outputs from the algorithm.
    - $generations$ – the number of generations computed within the algorithm's run.
    - $funccount$ – the number of evaluations of the fitness function.
    - $message$ – the reason for algorithm termination.
    - $maxconstraint$ – the maximum constraint violation, if any exist.

Typical results for a run of the genetic algorithm are as follows:

```
params =

   1.0e+03 *

    0.0005     0.0046     0.0100     0.0000     0.0099     0.0146
          0.0095    0.0000    5.1171    0.4525    0.0001

fav =

   -0.4505

exitflag =

     1

output =

      problemtype: 'boundconstraints'

      rngstate: [1x1 struct]

      generations: 61

      funccount: 12400

      message: 'Optimization terminated: average change in the
          fitness value less than options.TolFun.'

      maxconstraint: 0
```

Exit Flag	Meaning
1	Average cumulative change in value of the fitness function $overStallGenLimit$ generations is less than $TolFun$, and the constraint violation is less than $TolCon$.
2	Fitness limit reached and the constraint violation is less than $TolCon$.
3	Value of the fitness function did not change in $StallGenLimit$ generations and the constraint violation is less than $TolCon$.
4	Magnitude of step smaller than machine precision and the constraint violation is less than $TolCon$.
5	Minimum fitness limit reached and the constraint violation is less than $TolCon$.
0	Maximum number of generations exceeded.
-1	Optimization terminated by an output function or plot function.
-2	No feasible point found.
-4	Stall time limit exceeded.
-5	Time limit exceeded.

Table 4.4: Exit Flags and Meanings from MATLAB Genetic Algorithm (MATLAB, 2015).

For the fitness function used in our analysis, code was written that had flexibility to use many different evaluation methods which the user can define. The code uses the genetic algorithms outputs for x as well as the previously defined variables of "checkins," and "venues." The other variables referenced, "totalOD," "n," and "alg," are defined as follows:

- "totalOD" refers to the OD matrix of the comparison model that encompasses all of the trip types to be analyzed.
- "n" is a scalar value that is defined by the length of the "totalOD" matrix.

- "alg" refers to the algorithm model type (i.e., doubly-constrained gravity, peer-to-peer) being analyzed.

Within the function, the customizable component is found within the apostrophes. For the purpose of this dissertation a coincidence ratio (CR) is used. The CR determines how "closely" the proposed model matches the comparison model using the following equation which measures the area that "coincides" between the matrices that are used within the comparison (Martin 1998).

$$CR = \frac{\sum_i min\left(p_i^M, p_i^O\right)}{\sum_i max\left(p_i^M, p_i^O\right)} \quad (Eqn. 4.30)$$

Where

- p_i^M represents the percentage of trips within the interval *i* in the predicted trips from the check-in data
- p_i^O represents the percentage of trips within the interval *i* in the survey trips from the comparison dataset.

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 genetic algorithm is to satisfy the fitness function by finding the values for the variables that result in a model with a CR as close to one as possible by minimizing the mean absolute error (MAE). The code for this model calibration method can be found in Appendix B.

Model Scaling

Once each model calculates a final T_{ij} matrix the following equations: are used to additionally calibrate the model:

$$TotalOD_{ij} = \sum MPO\ Trips \quad (Eqn. 4.31)$$

$$MPO\_Sum = \sum_j \sum_i TotalOD_{ij} \quad (Eqn. 4.32)$$

$$T\_model = \frac{T_{ij}}{\left(\sum_j \sum_i T_{ij}\right) * MPO_{Sum}} \quad (Eqn. 4.33)$$

Where

- *MPO Trips* are all of the trip purposes for each TAZ that are desired for inclusion within the study. These may be home based work, home based non-work, or any other purpose defined by the MPO.
- T_{ij} is the origin-destination matrix calculated from the previous steps.

This process is done to scale the calculated T_model to the MPO's origin-destination values for better comparison purposes. After this scaling is done, additional checks are performed on the T_model matrix to adjust for any bias from high frequency values and for extreme values using the following equations:

High Frequency Value Check:

1. Using the calculated model's T_{ij} any values that are larger than or equal to the lower bound threshold as found by the genetic algorithm are included in the adjustment.
2. Values are then adjusted using an adjustment factor, *adjMid*, which is found via the genetic algorithm, by means of the following formula:

$$T\_Model = adjMid * T\_Model_{highIdx} \quad (Eqn. 4.34)$$

Extreme Value Check:

1. Using the calculated model's T_{ij} any values that are larger than the upper bound threshold, found by the genetic algorithm, are selected for inclusion.

2. The total difference between the model's original T_{ij} and the T_Model from the high frequency check is calculated using the following formula:

$$Total\_diff = \sum_j \sum_i \left(org_{T_{Model}} - T_{Model}\right) \quad (Eqn. 4.35)$$

3. The total difference is then redistributed to the matrix using the following formula:

$$T\_Model = T\_Model + \left(Total_{diff} * \left(T_{Model}^{\left(\eta / \sum_j \sum_i (T_{Model}{}^{\eta})\right)}\right)\right) (Eqn. 4.36)$$

MODEL VALIDATION

To be able to determine how each of the proposed models performs, an origin-destination matrix will be created from the proposed models and a comparison to the local origin-destination matrix will be performed to determine how "closely" the proposed model matches. This "closeness" will be analyzed using the coincidence ratio described above. In addition to the coincidence ratio, the mean error (ME), the mean absolute error (MAE), the frequency ratio (FR), and the swap ratio will be used determine the validity of each model's origin-destination matrix within the analysis as compared to the MPO model.

The ME is a measure that indicates if the model is biased in a positive or negative manner with respect to the following calculation:

$$ME = \frac{\sum_{i=1}^{N} ModelTrips_i - MPOTrips_i}{N} \quad (Eqn. 4.37)$$

Where

- N is the total number of origin-destination pairs
- $MPOTrips_i$ is the number of trips for the origin-destination pair from the comparison dataset, which is a 1xN matrix
- $ModelTrips_i$ is the number of trips for each origin-destination pair from model matrix, which is a 1xN matrix

The MAE is a method used to determine how close a prediction comes to actual outcomes through the examination of the average magnitude of errors within the prediction. The result is a linear value ranging from zero to infinity that relates to the error that is expected from the prediction. One benefit from this method is the limited sensitivity to occasional very large error. The calculation for MAE is done using the following equation:

$$MAE = \frac{\sum_{i=1}^{N} |MPOTrips_i - ModelTrips_i|}{N} \quad (Eqn. 4.38)$$

The frequency ratio (FR) compares the relative frequency of each trip value between the comparison data and the model data for each origin-destination pair. The relative frequencies are found by categorizing the number of trips into bins of 50 trip intervals and then turning these frequencies into a ratio to the total amount of trips from the dataset. The frequency ratio is then found using the following formula:

$$FR = \frac{\sum_{i=1}^{m} min\left(Tru_{RelFreq}, Model_{RelFreq}\right)}{\sum_{i=1}^{m} max\left(Tru_{RelFreq}, Model_{RelFreq}\right)} \quad (Eqn. 4.39)$$

Where

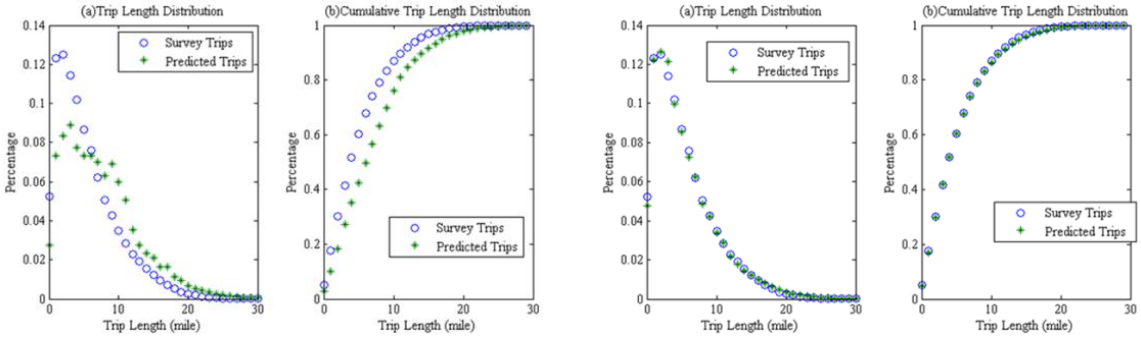
- $Tru_{RelFreq}$ is the relative frequency for the comparison dataset trip value and is a 1 x m matrix

- $Model_{RelFreq}$ is the relative frequency for the model dataset trip value and is a 1 x m matrix
- m is the number of bins used within the histogram calculation

With respect to the swap ratio, an absolute valued vector is found with respect to the two comparison points for each value within the compared matrices. The mean value of these vectors is the resulting swap ratio value, which gives a relative distance between the predicted trips and the comparison dataset's trips (Equation 4.40). For the swap ratio, new $ModelTrips_i$ and $MPOTrips_i$ matrices are created that remove any zeros within the matrix to ensure a value is attained for the ratio. Code used for the calculation of ME, MAE, FR, and swap ratio can be found in Appendix B.

$$swap\ ratio = \frac{\sum_{i=1}^{N} \left| \left(\tan^{-1} \frac{ModelTrips_i}{MPOTrips_i} \right) * \left(\frac{180}{\pi} \right) \right|}{N} \quad (Eqn. 4.40)$$

In addition to the above calculations, the examination of trip length distribution and cumulative trip length distribution are done to compare how each model performs with respect to the comparison data. Trip lengths for the models as well as the comparison dataset are calculated by adding up the total number of trips that occur within each interval of 100 between zero and 3,000 miles. A calculation is then performed to turn these totals into percent values. The cumulative percentages of trips for each interval are than calculated for the model and comparative datasets. These calculations result in a graphical depiction for each model that shows where there is consistency with respect to general curvature, and where over and under estimation exists with respect to trip lengths. The graphic will demonstrate where further adjustments to the models may be needed. Figure 4.3 provides a sample of this graphic. The MATLAB code for this is analysis is provided in Appendix B.



(a) Singly-Constrained Model Trip Length Frequency Results

(b) Doubly-Constrained Model Trip Length Frequency Results

Figure 4.3: Sample Trip Length Distribution Comparison (Jin et al. 2014)

For each model the productions and attractions for each TAZ will be compared to the productions and attractions for the comparison dataset. This is done by mapping the production to the TAZ using ArcGIS's join feature. For consistency purposes, the gradient color scheme used to differentiate the range of productions and attraction values will be kept constant for all of the models. This graphical representation, see Figure 4.4 for sample, will visually demonstrate where there are inconsistencies in how the proposed methods calculate productions and attractions, as well as where the method excels. Consequently, this will reveal where there may be an overabundance or lack of check-in data.

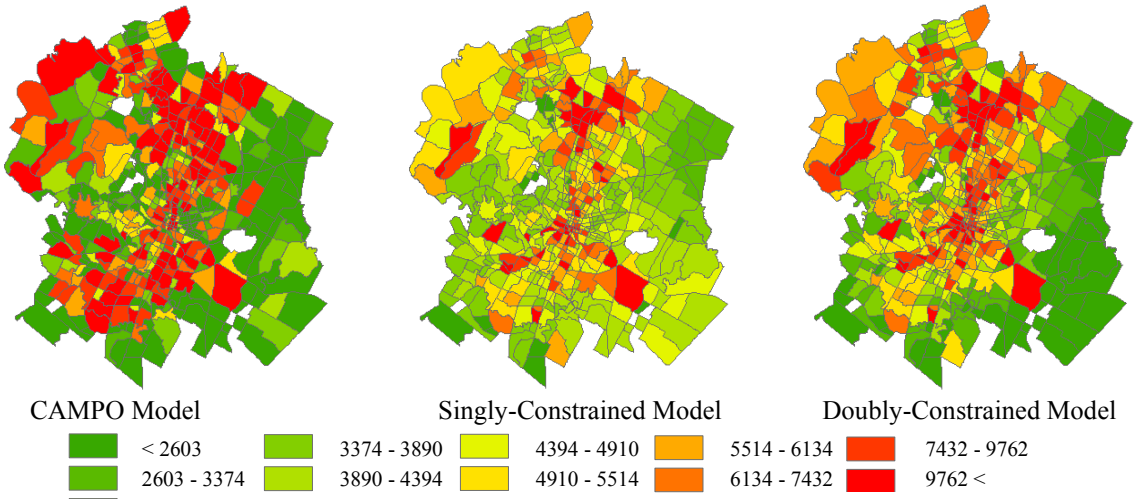


Figure 4.4: Sample Production Comparisons (Jin et al. 2014)

The final method used to validate each model's capabilities uses an origin-destination flow pattern intensity graphic. This graphic, Figure 4.5 provides a sample, is created using the following formula for each model and for the comparison dataset:

$$I_{ij} = log_{10}\left(\frac{T_{ij}}{\sum_i \sum_j T_{ij}}\right) \quad (Eqn. 4.41)$$

Where

- I_{ij} is the intensity of travel to TAZ
- T_{ij} are the number of trips per TAZ from either the model or the comparison dataset, depending on which is being analyzed

Higher origin-destination flows are shown via the darker coloring, while lesser flows are shown in the lighter coloring. The graphics allow for a visual analysis on how closely the variations in colors and the striations from the model graphic match the comparison data graphic. The more similar these colors and striations are, the better the fit of the model to the comparison data. In addition to the model and comparison intensity graphic, an intensity MAE matrix is created that visually shows TAZ by TAZ error

magnitude. A log of the histogram trip frequency values are also provided for the study area to demonstrate where there is over or underestimation with respect to amount of travel for various distances pictorially. Appendix B provides a sample of the code used for the creation of these analysis graphics.

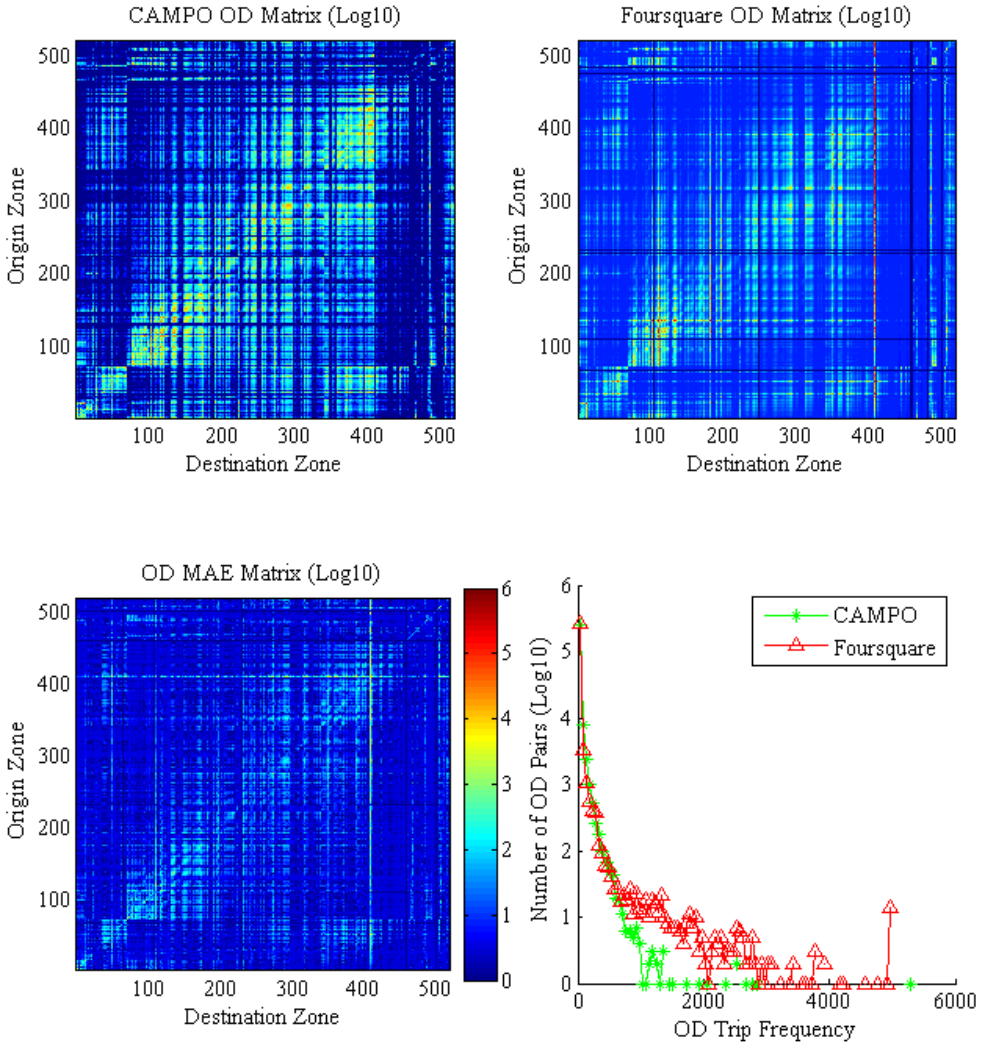


Figure 4.5: Sample OD Flow Pattern, MAE, Trip Frequency Intensity Graphic (Jin et al. 2014)

CONCLUDING STATEMENTS

This chapter presents the two methodologies used for an in-depth analysis of the use of location-based social networking data for the creation of origin-destination matrix. The two models presented, doubly-constrained gravity and peer-to-peer, used a two-regime friction function that was comprised of three different equations for a total of nine different functions. Each model was optimized using a genetic algorithm. Chapter 5 will present a case study using the methodologies described in this chapter and present the results of the analyses performed.

Chapter 5: Case Study

In order to determine the effectiveness of the proposed Peer-to-Peer modeling approach in conjunction with the use of location-based social networking, a case study using Austin, TX as a study area was performed. In this chapter, the study area will be described and the existing local model by the metropolitan planning organization (MPO) will be explained. The results of the methodologies used in Chapter 4 will be presented and a comparison to the CAMPO model will be done. Finally, the resulting peer-to-peer, doubly-constrained gravity, and CAMPO models will be presented with discussion on the strengths and weaknesses of each model.

STUDY AREA

In alignment with previous research efforts into the use of location-based social networking data for transportation planning (Jin et al. 2013, Cebelak 2013, Cebelak 2014, Jin et al. 2014), this dissertation uses Austin, TX as the location for analysis. Austin functions as the capital of the state of Texas and is part of the Austin-Round Rock Metropolitan Statistical Area (MSA), which is comprised of the five counties of Williamson, Travis, Hays, Bastrop, and Caldwell. According to the City of Austin website, as of April 1, 2015 the Austin-Round Rock MSA has a population of 1,990,593 and a land area of 4,285.70 mi^2 and the City of Austin has a population of 900,701 and a land area of 322.48 mi^2, the majority of which resides within Travis County (see Figure 5.1). In addition to its role as the capital, Austin is home to the University of Texas at Austin, the location for many Fortune 500 companies' headquarters and offices, examples of which include Dell, Whole Foods Market, and Advanced Micro Devices Inc. (CNN Money 2013), and is known as "The Live Music Capital of the World" playing host to more than 250 music venues and festivals each year which bring over 19 million visitors to the city annually (Austin Chamber 2013).

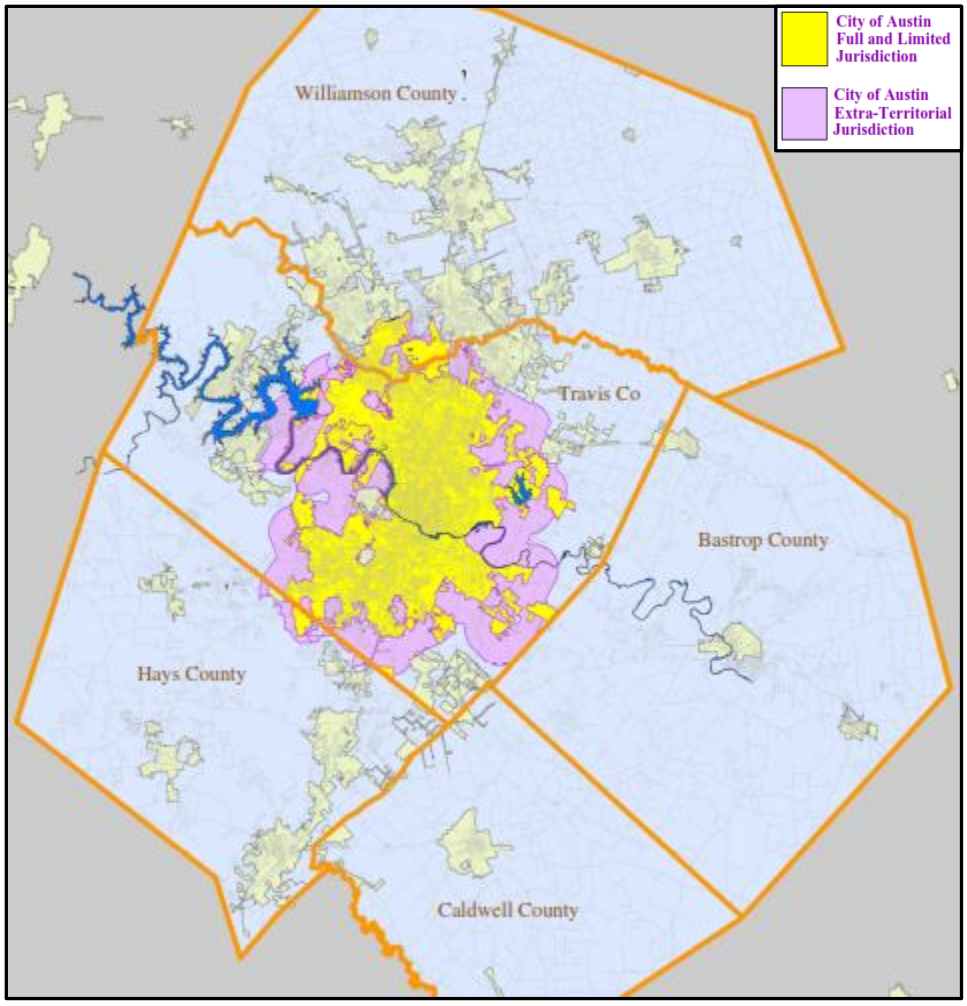


Figure 5.1: 2013 Austin-Round Rock MSA Map (City of Austin 2015)

Capital Area Metropolitan Planning Organization (CAMPO)

The metropolitan planning organization that the City of Austin is a part of is the Capital Area Metropolitan Planning Organization (CAMPO), which includes the counties within the Austin-Round Rock MSA as well as Burnet County. CAMPO is the coordinating body for the regional transportation planning efforts with the counties, the cities, the Capital Metropolitan Transportation Authority, the Capital Area Rural Transportation System, the Lone Star Rail, the Central Texas Regional Mobility Authority, and the Texas Department of Transportation (CAMPO 2015). As part of its

responsibilities, CAMPO produces the Long-Range Transportation Plan, the most recent of which is the 2005 version. This latest version of the Long-Range plan uses 2005 base year data, which was recalibrated and validated according to the *CAMPO Urban Transportation Study: 2005 Base Year Travel Demand Model Calibration and Validation for Updating the 2035 Long Range Plan* document.

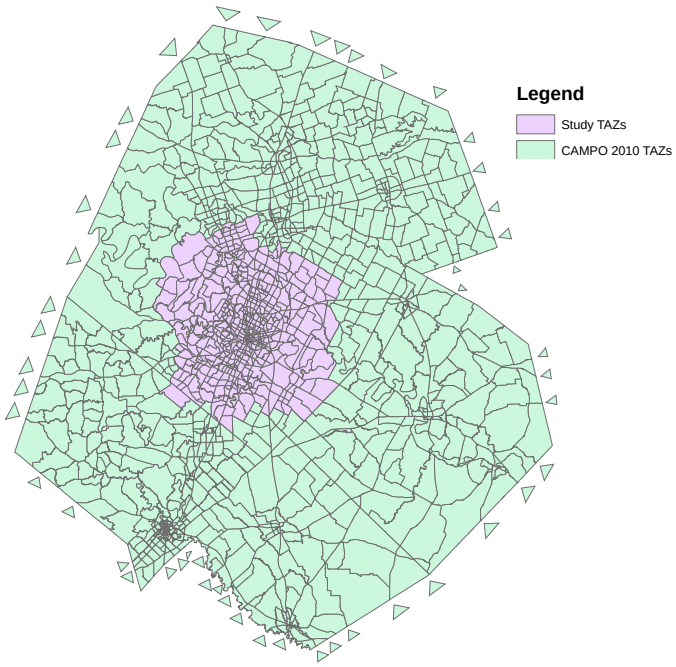
For the 2005 CAMPO Travel Demand Model (TDM), a total of 1,413 traffic analysis zones (TAZs) make up the regional plan. Of the 1,413 regional TAZs, the 520 that exist within the City of Austin area will be included in the case study analysis (Figure 5.2). The model defines a total of 17 trip purposes that include four external trip purposes and the commercial truck/taxi vehicle trips, which will not be included within the analysis. The remaining 12 person trip purposes are as follows:

1. Home Based Work Direct (HBW-Direct)
2. Home Based Work Strategic (HBW-Strategic)
3. Home Based Work Complex (HBW-Complex)
4. Home Based Non-work Retail (HBNW-R)
5. Home Based Non-work Other (HBNW-O)
6. Home Based Non-work Primary Education (HBNW-E1)
7. Home Based Non-work University/College (HBNW-E2)
8. Home Based Non-work UT-Austin Education (HBNW- UT)
9. HBNW/NHB (Non-work) Airport (NW-Airport)
10. Non-home Based Work-related (NHB-W)
11. Non-home Based Other (NHB-O)
12. Non-home Based External Commuter/Visitor Vehicle Trips (NHB-Exlo)

CAMPO defines Home Based Work trips into three different categories to provide additional insight for mode choice decisions. However, the analysis presented

within this Chapter will not differential between these categories and will group these trips into a single Home Based Work (HBW) category. Additionally, the two non-University of Texas (UT) educational categories are not included within the analysis. The rational for doing this is that the data collected occurred during the summer (June 11 through July 2) and there is little commuting traffic to these locations. Since UT is one of ten companies that employ over 6,000 (Austin Chamber 2015) the Home Based Non-work UT trips will be included in the study. These changes result in the following defined trip purposes:

1. Home Based Work (HBW)
2. Home Based Non-work Retail (HBNW-R)
3. Home Based Non-work Other (HBNW-O)
4. Home Based Non-work UT-Austin Education (HBNW- UT)
5. HBNW/NHB (Non-work) Airport (NW-Airport)
6. Non-home Based Work-related (NHB-W)
7. Non-home Based Other (NHB-O)
8. Non-home Based External (NHB-E)



9.

Figure 5.2: City of Austin TAZs (City of Austin 2015)

The data used for the TDM came from travel surveys that were conducted during 2005 and 2006. This data included individual surveys that were made up of 1,500 household samples that were selected based on household income and size (CAMPO 2010). Each household was asked to complete a travel diary, which was used to develop household trip production rates, trip length frequency distributions, and other salient data. From this data trip production and attraction rates were created based on household characteristics (i.e., income level, workers within the household) and the trip types described above. CAMPO utilizes a scaling effort for attractions to ensure agreement with the productions. Additionally, TAZs are assigned area types based on TAZ population and employment densities. The categories for area type include the following: central business district (CBD), urban intense (UrbInt), urban residential (UrbRes), suburban residential (SubUrbRes), and rural (Rural).

For trip distributions, CAMPO uses the atomistic trip distribution model, which is a triply-constrained gravity model and was described in detail within the literature review chapter of this dissertation. This method allocates intrazonal trips by utilizing radius data for each TAZ and trip length frequency distributions from the travel surveys. Trip length frequency model calculations use a gamma function that is fit to the average trip length in minutes, maximum allowable network separation in minutes, and trip purpose identifies. Bias factors can be applied within the modeling as needed and are used to address school district boundaries than may exist outside the CAMPO region where more trips were going out of the boundary than was appropriate. The model used for trip distributions includes a speed feedback loop that uses the method of successive averages using either the 24 hour highway trip table, the total misplaced flow of the 24 hour highway trip table, or statistics of the 24 hour assigned link flow table.

ANALYSIS OF PROPOSED METHODOLOGY

To be able to determine how the proposed models perform, the 2005 CAMPO Urban Transportation Study's TDM will be used and will be reduced to include only the 520 City of Austin TAZs and the respective 2005 Person Trip Table for the origin-destination matrix. To do this the CAMPO data was manipulated by creating a text file that contained only the TAZs to be included within the analysis. This file and the 2005 Person Trip Table text file was used within MATLAB to assign the trips to each of the eight trip categories defined above resulting in a 520x520 matrix. The code used for this effort can be found in the master's thesis by this author (Cebelak 2013).

It is important to note here that while the CAMPO model is used for comparison in this dissertation, it does not imply that the model is "ground truth." The CAMPO model uses survey data, which studies have shown have significant under-reporting (Bricka 2010, Srinivasan et al. 2006), concerns about data quality and completeness

(Bricka 2010, Srinivasan et al. 2005), and trust limitations (i.e., only valid if survey variables and stated preferences have not changed) (Devillaine, Munizaga, and Trépanier 2012). However, since this model is the accepted model used for the metropolitan areas planning, it has been deemed acceptable for comparison purposes. Additionally, it should be noted that the data used within the CAMPO model comes from a 2005 survey, while the Foursquare dataset comes from 2012. This is especially important since the growth rate of the city of Austin's population has been increasing since the 2005 survey with a growth of 16.4% between the two study years (Table 5.1).

Year	Population	Annualized Growth Rate
2000	656,562	-
2001	669,693	2.00%
2002	680,899	1.70%
2003	687,708	1.00%
2004	692,102	0.60%
2005	**700,407**	**1.20%**
2006	718,912	2.60%
2007	735,088	2.30%
2008	750,525	2.10%
2009	774,037	3.10%
2010	790,390	2.10%
2011	812,025	2.70%
2012	**824,205**	**1.50%**
2013	842,750	2.30%
2014	865,504	2.70%

Table 5.1: City of Austin's Growth Rates Since 2000 (Demographic Data 2015).

Each of the validation methods discussed in Chapter 4 will be used within this section to analyze the results from each of the 18 models. These include the coincidence ratio (CR), the mean error (ME), the mean absolute error (MAE), the frequency ratio

(FR), the swap ratio, the trip length distributions, the productions and attraction rates, as well as the intensity analysis efforts.

RESULTS AND DISCUSSION

This section will present the results for each of the 18 models included in the study: nine doubly-constrained gravity models and nine peer-to-peer models. The base models, doubly-constrained and peer-to-peer, will be compared independently and will identify "best" performing models using the criteria identified above and described in detail in Chapter 4. These "best" performing models will then be compared to one another to further examine the base models strengths and weaknesses.

Doubly-Constrained Gravity Model Results

Using the methodology described in the previous chapter, each friction function model was run using a genetic algorithm. The resulting nine doubly-constrained models will be discussed in this section with respect to the criteria from the previous chapter. For comparison against the peer-to-peer model, the "best" doubly-constrained model or models will be used. The qualification for "best" model(s) will be based on the model(s) that performs highest with respect to the criteria from the previous chapter.

Coincidence Ratio Analysis

The previous chapter defined the methodology used for the creation of the coincidence ratio (CR), which compares the doubly-constrained model to the CAMPO model, and determines how "closely" the doubly-constrained model trip distributions comes to the CAMPO trip distributions. Table 5.2 provides the CR values attained for each of the nine models. The table shows that the models are not sensitive to which version of friction function (linear, negative exponential, or gamma) is used for the short trips, but that they are sensitive to the friction function used for long trips. The negative

exponential and gamma functions perform significantly better than the linear for the long trips. The best performing two-regime friction functions are as follows:

1.) Linear - Negative exponential (0.9576)

2.) Gamma - Negative exponential (0.9449)

3.) Negative exponential - Negative exponential (0.9283)

Doubly-Constrained

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	0.4390	**0.9576**	0.7413
	Neg. Exp.	0.4389	0.9283	0.8132
	Gamma	0.4961	0.9449	0.8089

Table 5.2: Resulting Coincidence Ratio for Doubly-Constrained Gravity Models

Mean Error Analysis

The mean error was calculated for each of the nine models and the results are shown in Table 5.3. While the values range between positive and negative, all are very small in magnitude indicating little bias exists with respect to the origin-destination matrix creation. The friction-functions that would be considered "best" performers would be the following:

1.) Gamma - Linear (-1.0225E-14)

2.) Negative exponential - Negative exponential (-1.9623E-14)

3.) Linear - Linear (-2.0396E-14)

Doubly-Constrained

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	-2.0396E-14	2.8361E-14	-9.8484E-14
	Neg. Exp.	6.0113E-14	-1.9623E-14	-1.4100E-13
	Gamma	**-1.0225E-14**	-6.8777E-14	-6.5548E-14

Table 5.3: Resulting Mean Error for Doubly-Constrained Gravity Models

Mean Absolute Error Analysis

With respect to the MAE, the doubly-constrained methods were fairly consistent in the error calculations. Once again the models with negative exponential calculations for the long trips performed the best (Table 5.4), with the following ranked order:

1.) Linear - Negative exponential (9.9869)

2.) Gamma - Negative exponential (10.1379)

3.) Negative exponential - Negative exponential (10.5308)

Doubly-Constrained

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	13.1033	**9.9869**	12.4889
	Neg. Exp.	13.0792	10.5308	10.8799
	Gamma	13.0517	10.1379	10.7182

Table 5.4: Resulting Mean Absolute Error for Doubly-Constrained Gravity Models

Frequency Ratio Analysis

A comparison for each model examined the frequency of trips created with respect to the CAMPO model, which ranged from 0 to 6000. Trip frequencies were grouped into intervals of 50 for comparison and the results of the FR analysis are shown in Table 5.5. The models that had the closest values to one for their FR are as follows:

1.) Gamma - Gamma (0.9619)

2.) Gamma - Negative exponential (0.9588)

3.) Negative exponential - Gamma (0.9587)

Doubly-Constrained

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	0.9362	0.9565	0.9371
	Neg. Exp.	0.9277	0.9287	0.9587
	Gamma	0.9477	0.9588	**0.9619**

Table 5.5: Resulting Frequency Ratio for Doubly-Constrained Gravity Models

Swap Ratio Analysis

For the swap ratio analysis, the comparison resulted in similar trends as the MAE analysis. The values with the lowest swap ratio are seen when the negative exponential is used for the long trip component of the two-regime friction function (Table 5.6). The best performing models are as follows:

1.) Negative exponential - Negative exponential (26.5310)

2.) Linear - Negative exponential (26.7881)

3.) Gamma - Negative exponential (27.4977)

Doubly-Constrained

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	29.6414	26.7881	29.3637
	Neg. Exp.	29.7019	**26.5310**	28.5341
	Gamma	28.6112	27.4977	28.3214

Table 5.6: Resulting Swap Ratio for Doubly-Constrained Gravity Models

Trip Length Distribution Analysis

A graphical representation of trip length distributions was created in MATLAB for each of the nine models (Figures 5.3 to 5.11). Examining the graphics reveals that the models with the linear friction function used for the long trips have similar significant over and under estimation in trip length estimation (Figures 5.3, 5.6, and 5.9). When the gamma friction function was used for the long trips, the over and under estimation was less significant than was seen with the linear friction function (Figures 5.5, 5.8, and 5.11). The models with the closest trip length distributions are the models that use the negative exponential for the long trip component two-regime friction function (Figures 5.4, 5.7, and 5.10) with the following ranking order based on a visual analysis:

1.) Gamma - Negative exponential

2.) Linear - Negative exponential

3.) Negative exponential - Negative exponential

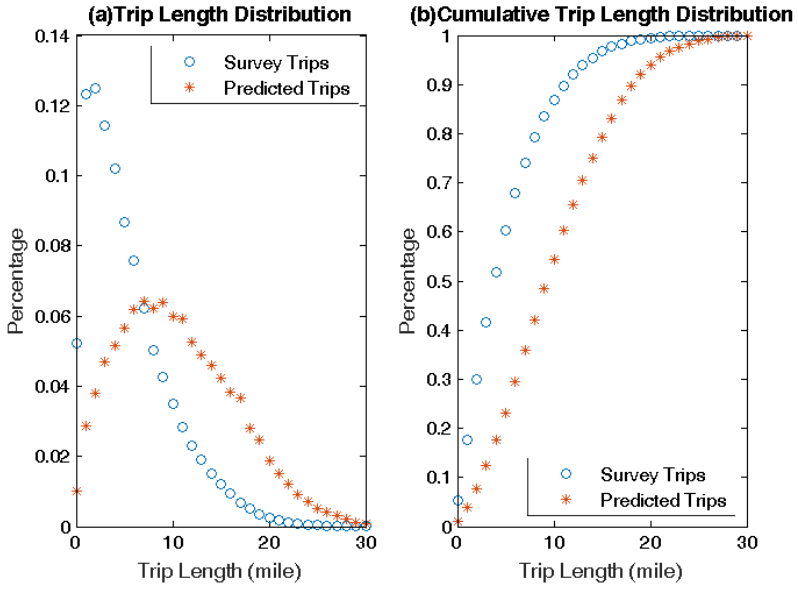


Figure 5.3: Trip Length Distributions for Linear-Linear Doubly-Constrained Gravity Model

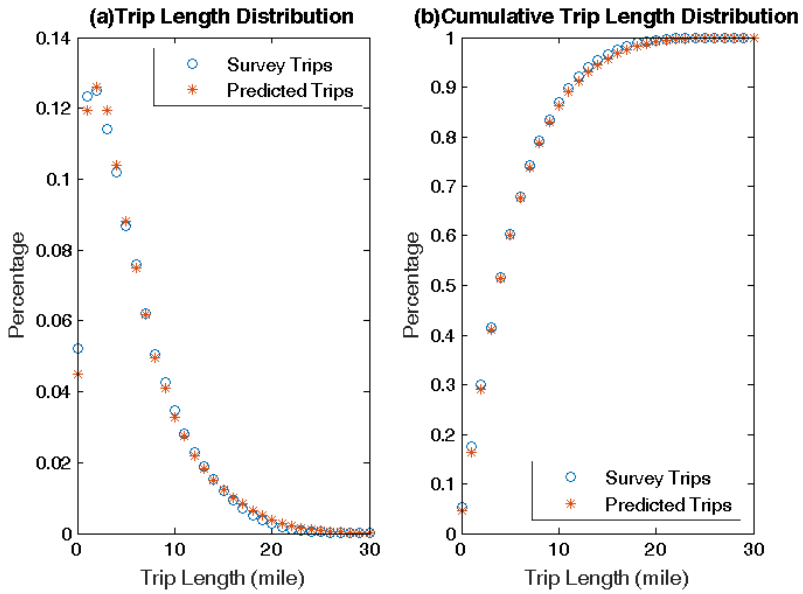


Figure 5.4: Trip Length Distributions for Linear-Negative Exponential Doubly-Constrained Gravity Model

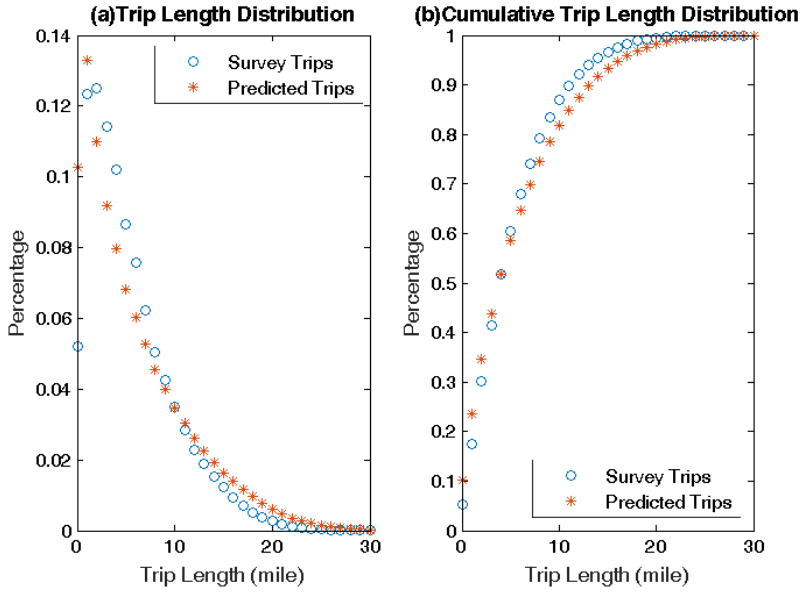


Figure 5.5: Trip Length Distributions for Linear-Gamma Doubly-Constrained Gravity Model

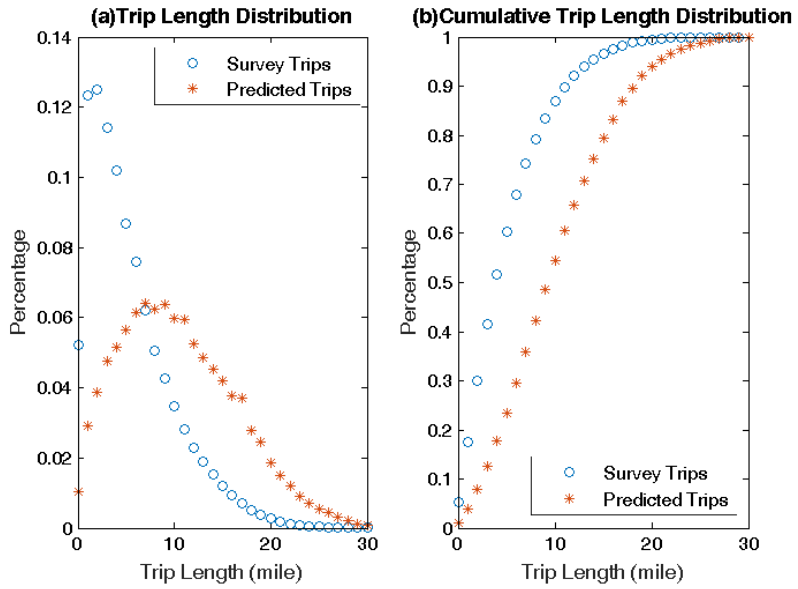


Figure 5.6: Trip Length Distributions for Negative Exponential-Linear Doubly-Constrained Gravity Model

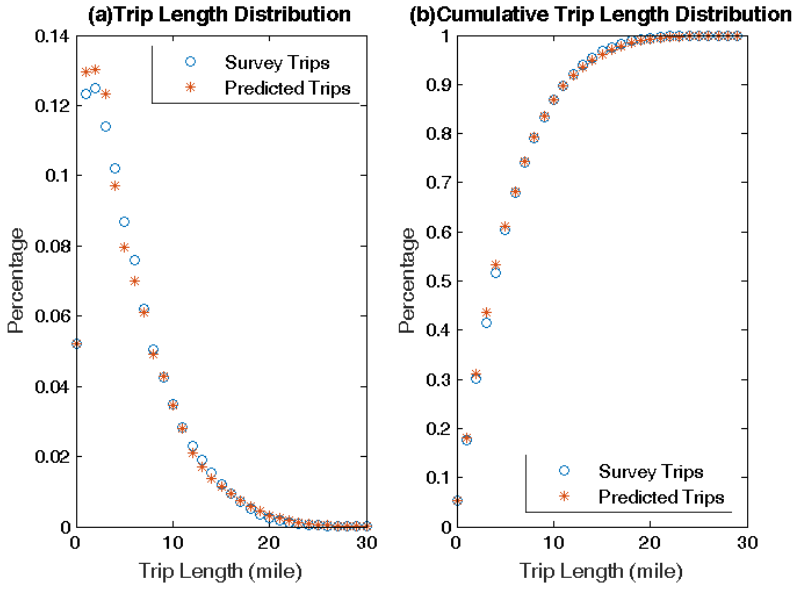


Figure 5.7: Trip Length Distributions for Negative Exponential-Negative Exponential Doubly-Constrained Gravity Model

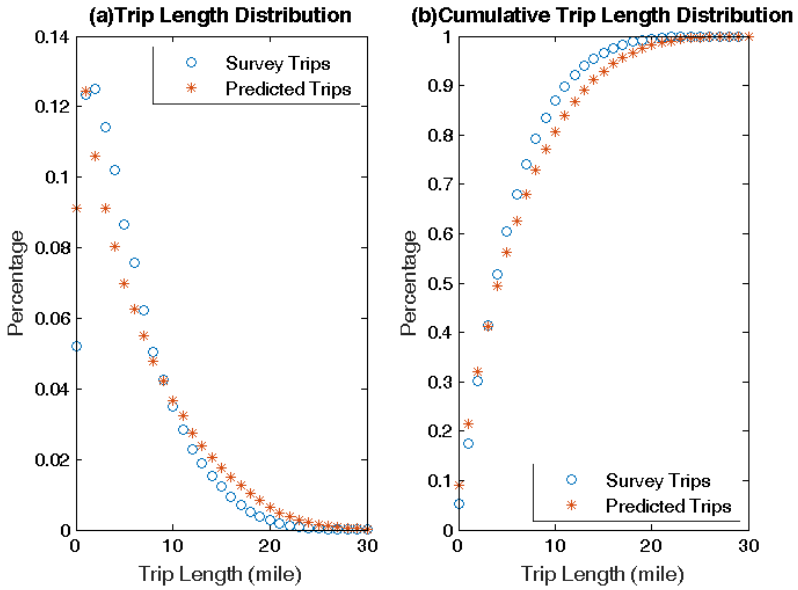


Figure 5.8: Trip Length Distributions for Negative Exponential-Gamma Doubly-Constrained Gravity Model

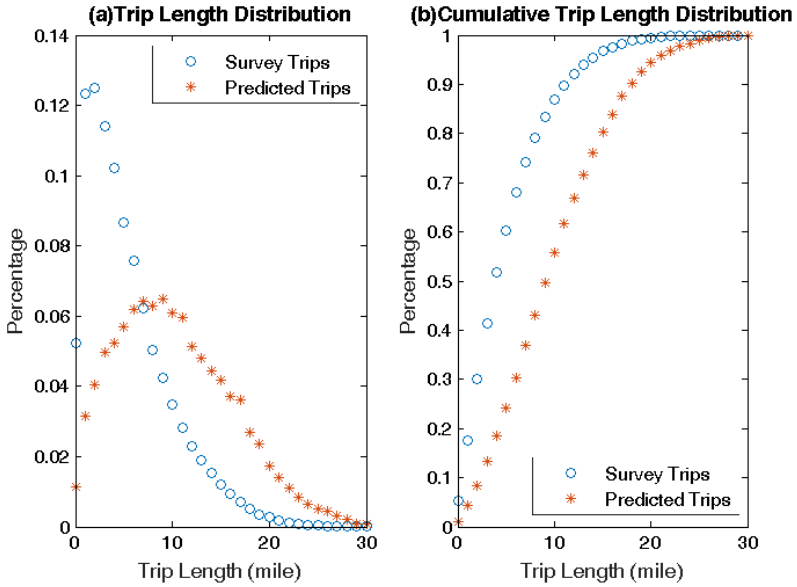


Figure 5.9: Trip Length Distributions for Gamma-Linear Doubly-Constrained Gravity Model

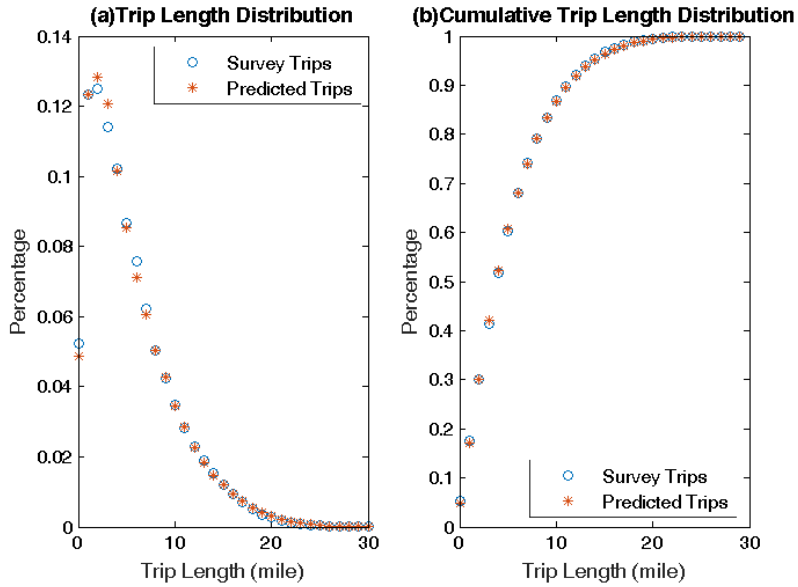


Figure 5.10: Trip Length Distributions for Gamma-Negative Exponential Doubly-Constrained Gravity Model

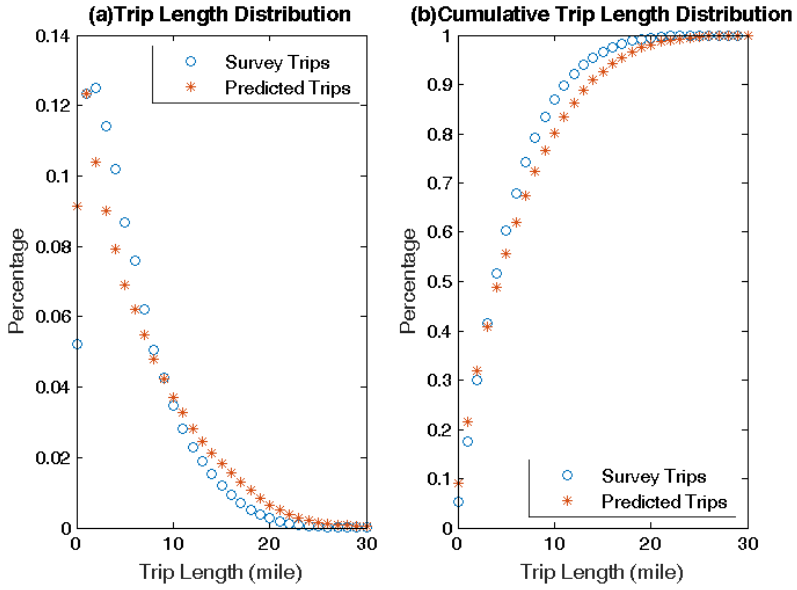


Figure 5.11: Trip Length Distributions for Gamma-Gamma Doubly-Constrained Gravity Model

Production and Attraction Analysis

Digging into the trip generations with respect to productions and attractions reveals additional insight into each models capabilities in matching the CAMPO model. Figures were created in ArcGIS that illustrated the number of trip productions and attractions for each TAZ as created by each model. Each figure created uses the same color gradation break down to aid in the ability of comparison. This color gradation was limited to ten color variations to assist in the ease of interpretation and provides emphasis on lower trip production and attraction value differentiation.

Figure 5.12 shows how the productions generated by the CAMPO model. Figure 5.13 provides the productions generated by each of the proposed doubly-constrained gravity models. To better analyze how each model's productions rate with respect to the comparison CAMPO model, the color denotation for each TAZ in each model was compared to the corresponding TAZ in the CAMPO model. If the color for the model TAZ is the same as the color for the CAMPO TAZ, a value of "Y" is given. If the color in the model TAZ is within one shade darker or lighter, a value of "C" is given. If the color in the model TAZ does not meet any of these criteria, a value of "N" is given. To determine which model(s) have the most TAZ with the same shade, "Y," or close shade, "C," the total number of "Y" and "C" matches were calculated and then reported as a percent of the total number of TAZs. Table 5.7 provides a breakdown of these statistics for the productions for each doubly-constrained model. Examining the models, it was determined that the models that had the most TAZs with the same categorization, "Y," of productions were as follows:

1.) Negative exponential - Negative exponential
2.) Linear - Negative exponential
3.) Gamma - Linear

With respect to the models with the most similar categorization, “Y” and “C,” the models that ranked the highest were:

1.) Gamma - Linear

2.) Negative exponential - Negative exponential

3.) Linear - Negative exponential

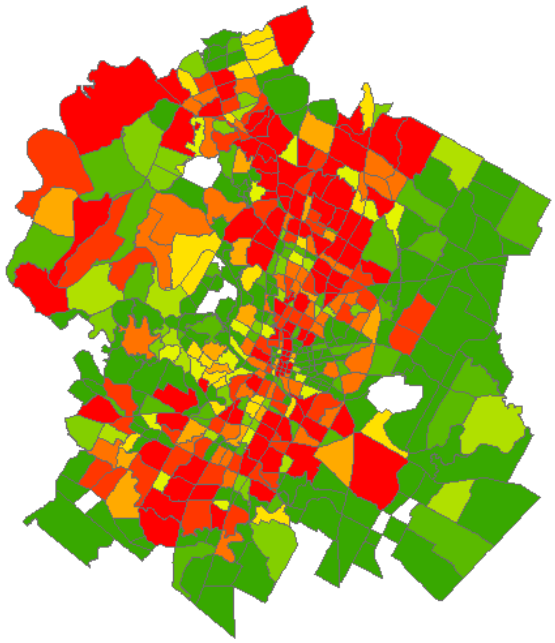


Figure 5.12: Trip Productions for the CAMPO Model

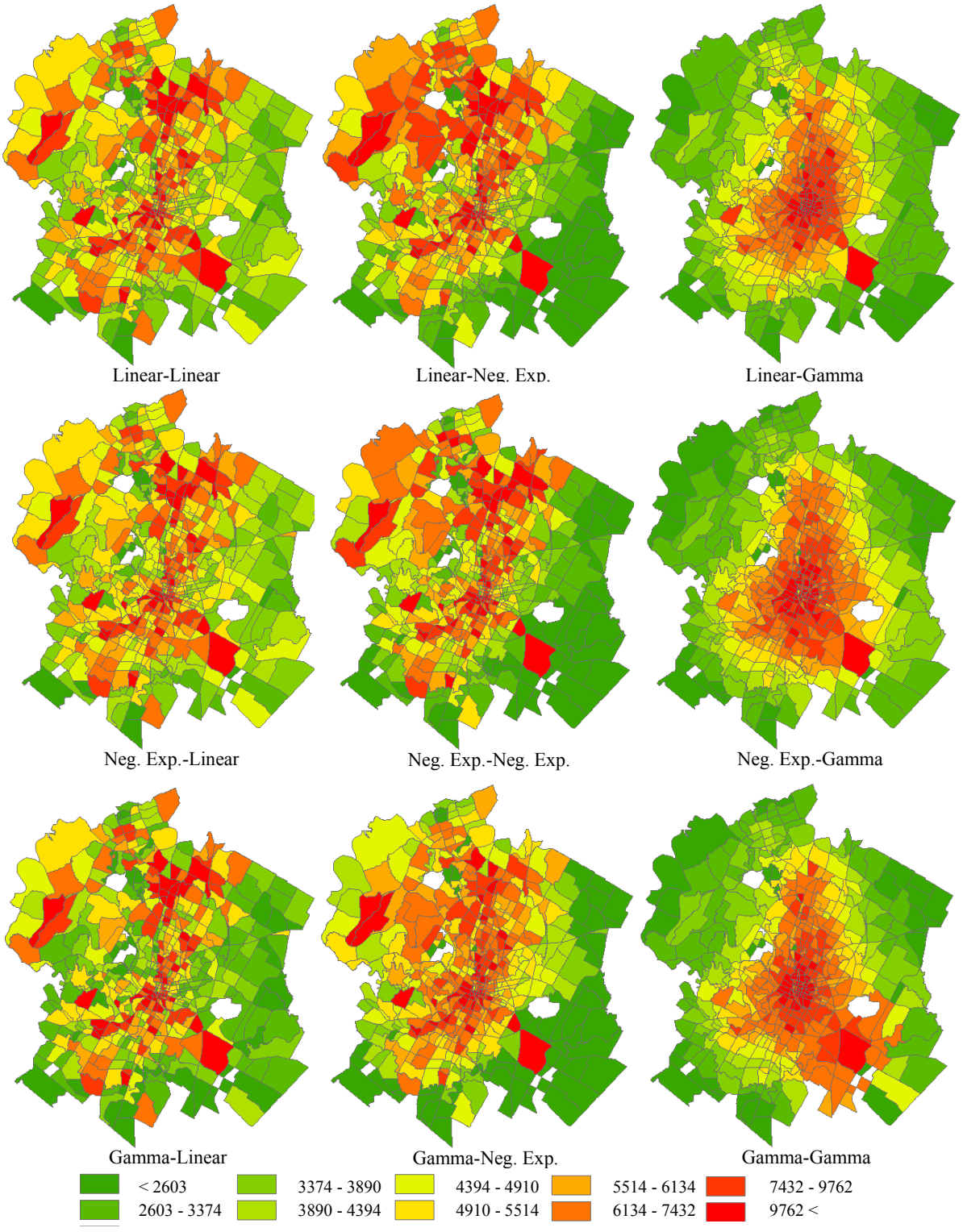


Figure 5.13: Trip Productions for the Proposed Doubly-Constrained Gravity Models

	Linear-Linear		**Linear-Neg. Exp.**		Linear-Gamma	
	#	%	#	%	#	%
Y	76	14.62	104	**20.00**	82	15.77
C	112	21.54	120	23.08	100	19.23
N	332	63.85	296	56.92	338	65.00
Y+C	188	36.15	224	**43.08**	182	35.00
	Neg. Exp.-Linear		**Neg. Exp.-Neg. Exp.**		Neg. Exp.-Gamma	
	#	%	#	%	#	%
Y	78	15.00	110	**21.15**	78	15.00
C	103	19.81	116	22.31	99	19.04
N	339	65.19	294	56.54	343	65.96
Y+C	181	34.81	226	**43.46**	177	34.04
	Gamma-Linear		Gamma-Neg. Exp.		Gamma-Gamma	
	#	%	#	%	#	%
Y	95	**18.27**	91	17.50	82	15.77
C	153	29.42	106	20.38	93	17.88
N	272	52.31	323	62.12	345	66.35
Y+C	248	**47.69**	197	37.88	175	33.65

Table 5.7: TAZ Production Rate Graphical Similarity Statistics Doubly-Constrained Gravity Models

Following the steps described previously, the trip attractions were examined. Figure 5.14 shows how the attractions generated by the CAMPO model are distributed. Figure 5.15 provides the attractions generated by each of the proposed doubly-constrained gravity models. Table 5.8 provides a breakdown of these statistics for the productions for each doubly-constrained model. Examining the nine doubly-constrained gravity models, it was determined that the models with the most TAZs with the same categorization, "Y," of attractions were as follows:

1.) Gamma - Linear

2.) Negative exponential - Negative exponential

3.) Linear - Negative exponential

With respect to the models with the most similar categorization, “Y” and “C,” the models that ranked the highest were:

1.) Gamma - Linear

2.) Linear - Linear

3.) Negative exponential - Negative exponential

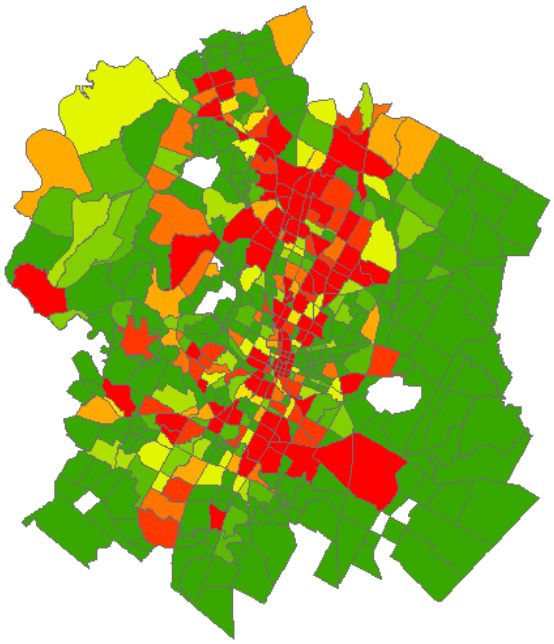


Figure 5.14: Trip Attractions for the CAMPO Model

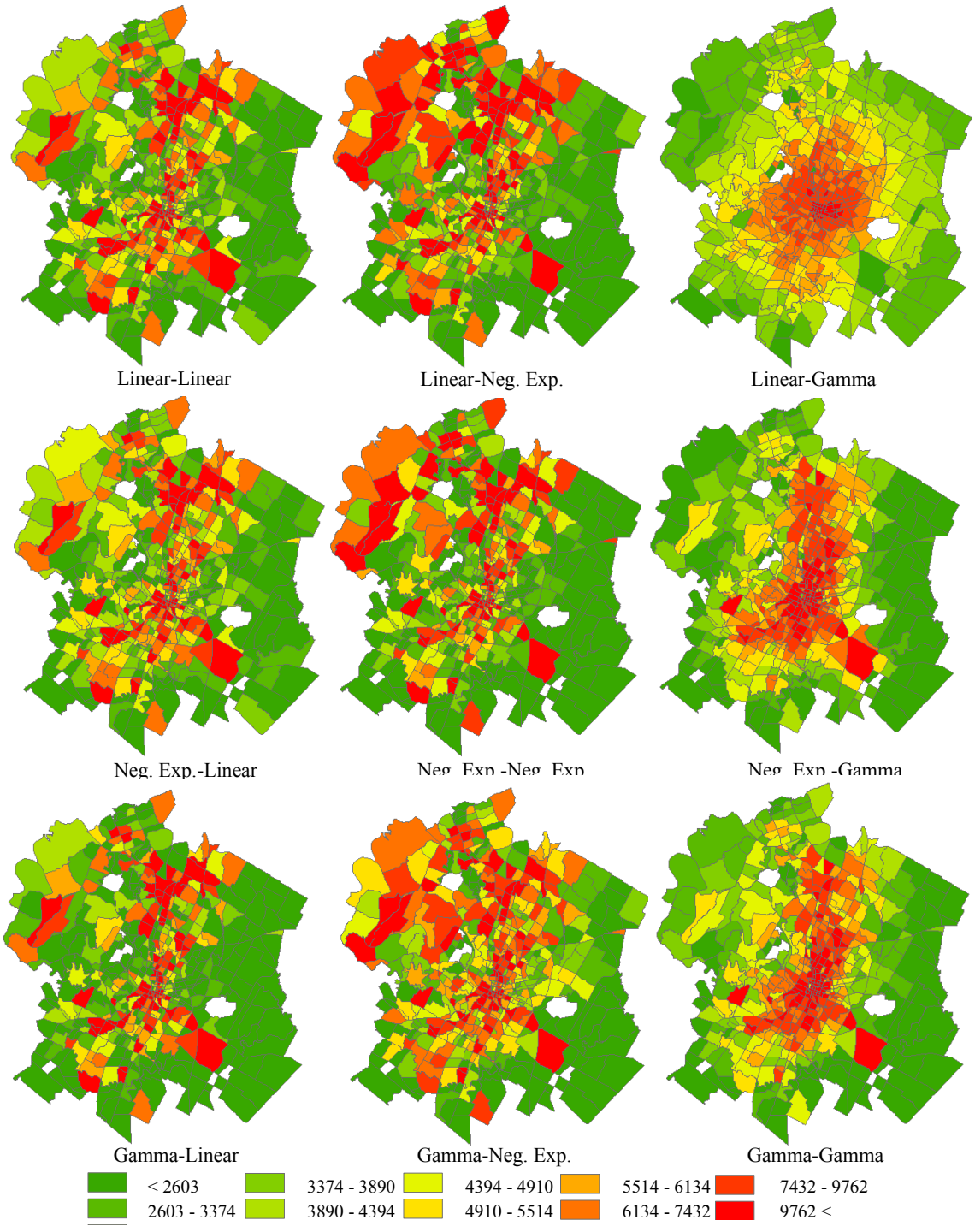


Figure 5.15: Trip Attractions for the Proposed Doubly-Constrained Gravity Models

	Linear-Linear		**Linear-Neg. Exp.**		Linear-Gamma	
	#	%	#	%	#	%
Y	209	40.19	223	**42.88**	57	10.96
C	124	23.85	100	19.23	69	13.27
N	187	35.96	197	37.88	394	75.77
Y+C	333	**64.04**	323	62.12	126	24.23
	Neg. Exp.-Linear		**Neg. Exp.-Neg. Exp.**		Neg. Exp.-Gamma	
	#	%	#	%	#	%
Y	192	36.92	229	**44.04**	131	25.19
C	131	25.19	103	19.81	118	22.69
N	197	37.88	188	36.15	271	52.12
Y+C	323	62.12	332	**63.85**	239	47.88
	Gamma-Linear		Gamma-Neg. Exp.		Gamma-Gamma	
	#	%	#	%	#	%
Y	265	**50.96**	170	32.69	162	31.15
C	90	17.31	117	22.50	106	20.38
N	165	31.73	233	44.81	252	48.46
Y+C	355	**68.27**	287	55.19	268	51.54
Y+C	355	**68.27**	287	55.19	268	51.54

Table 5.8: TAZ Attraction Rate Graphical Similarity Statistics Doubly-Constrained Gravity Models

Intensity Analysis

The intensity analysis described in Chapter 4 was performed on each of the nine doubly-constrained gravity models with the resulting graphics shown in Figures 5.16 through 5.24. For these graphics, the lighter striations show the areas of high flow between origin-destination TAZ pairs for the doubly-constrained model as well as the CAMPO model. In addition to the origin-destination intensity analysis, an intensity MAE analysis and an origin-destination trip frequency analysis was performed. The MAE

analysis shows the error magnitude from the proposed model, while the origin-destination trip frequency analysis shows where over and under estimation occurs within the proposed model. Through visual analysis, the following observations can be made:

1.) The models that used the linear component for long trips did not show similar flow rates for intrazonal trips compared to the CAMPO model. This was confirmed with the lighter color along the 45° line within the MAE intensity graphics for these models, which indicate a larger error.
2.) The models with the negative exponential long trips showed better intrazonal trip calculations and had color striations that better matched the CAMPO model. The MAE intensity graphic confirmed this with the presence of dark shading throughout the graphics.
3.) The models with the gamma long trips showed over calculation for the intrazonal trips, which was confirmed by the MAE intensity graphic that shows a pronounced 45° line.
4.) The best preforming model with respect to MAE intensity was the linear-negative exponential model (Figure 5.17).
5.) Examination of the OD trip frequency graphics for all of the models showed each models ability to closely represent the frequencies for lower values, but many had over or under estimation for higher frequencies.
6.) The model that best performed with respect to OD trip frequencies was the linear-negative exponential model (Figure 5.17), although there was significant over estimation for the tail of the curve.

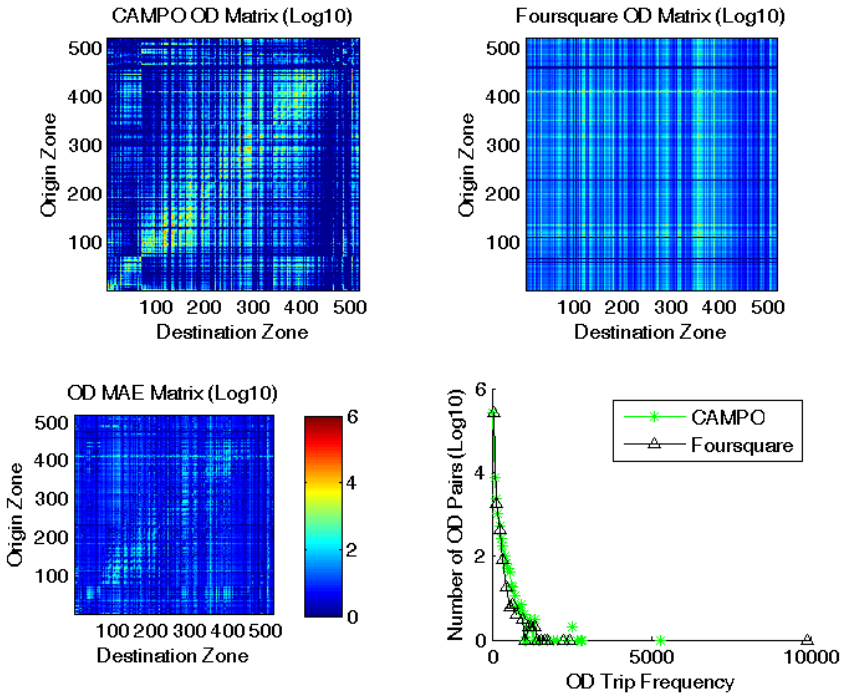


Figure 5.16: Intensity Diagrams for Linear-Linear Doubly-Constrained Gravity Model

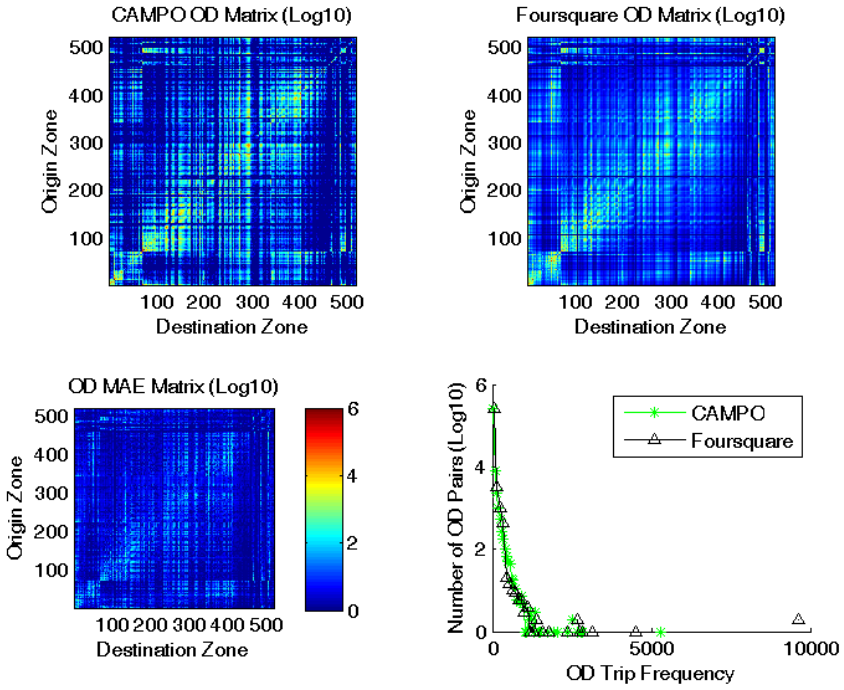


Figure 5.17: Intensity Diagrams for Linear-Negative Exponential Doubly-Constrained Gravity Model

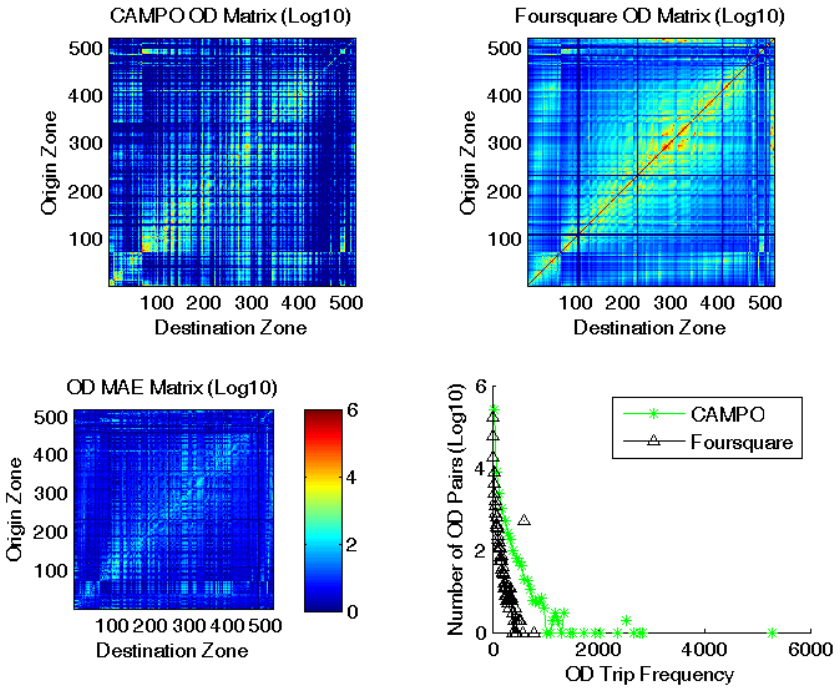


Figure 5.18: Intensity Diagrams for Linear-Gamma Doubly-Constrained Gravity Model

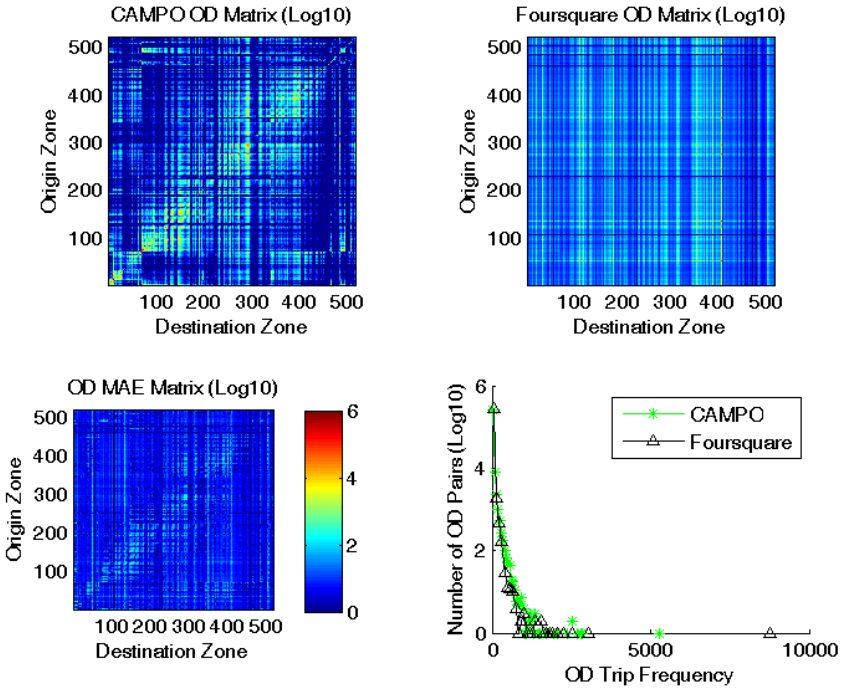


Figure 5.19: Intensity Diagrams for Negative Exponential-Linear Doubly-Constrained Gravity Model

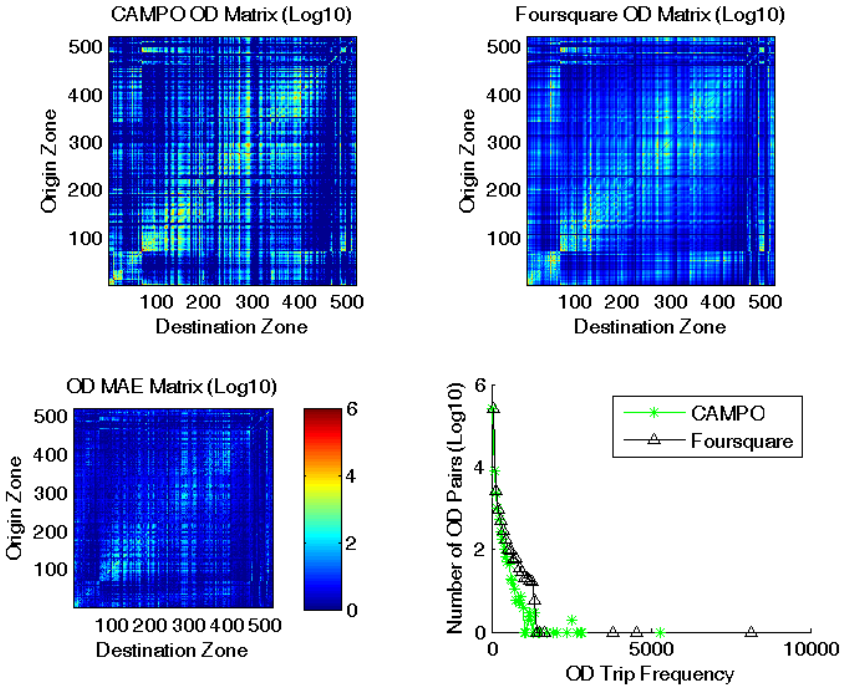


Figure 5.20: Intensity Diagrams for Negative Exponential-Negative Exponential Doubly-Constrained Gravity Model

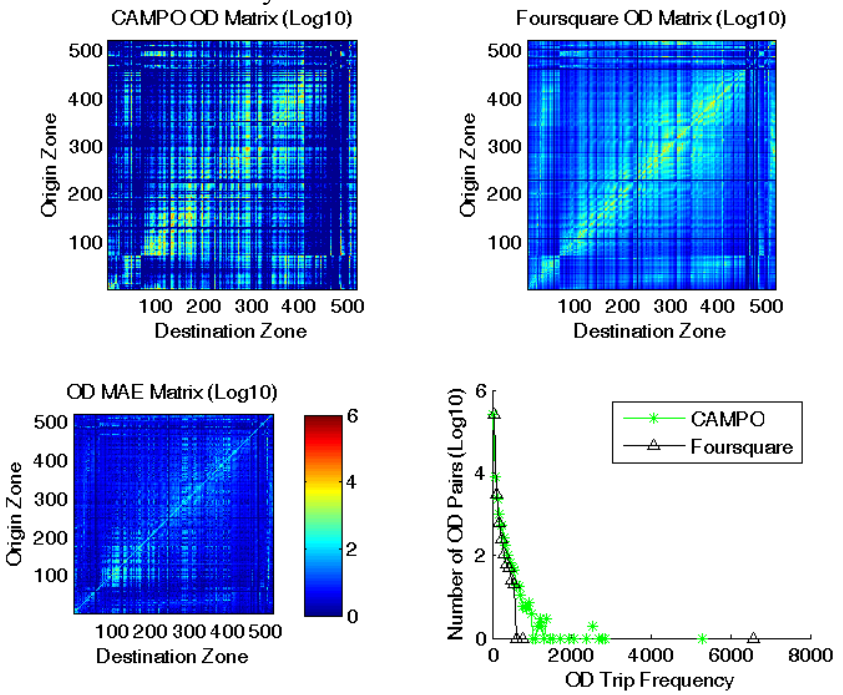


Figure 5.21: Intensity Diagrams for Negative Exponential-Gamma Doubly-Constrained Gravity Model

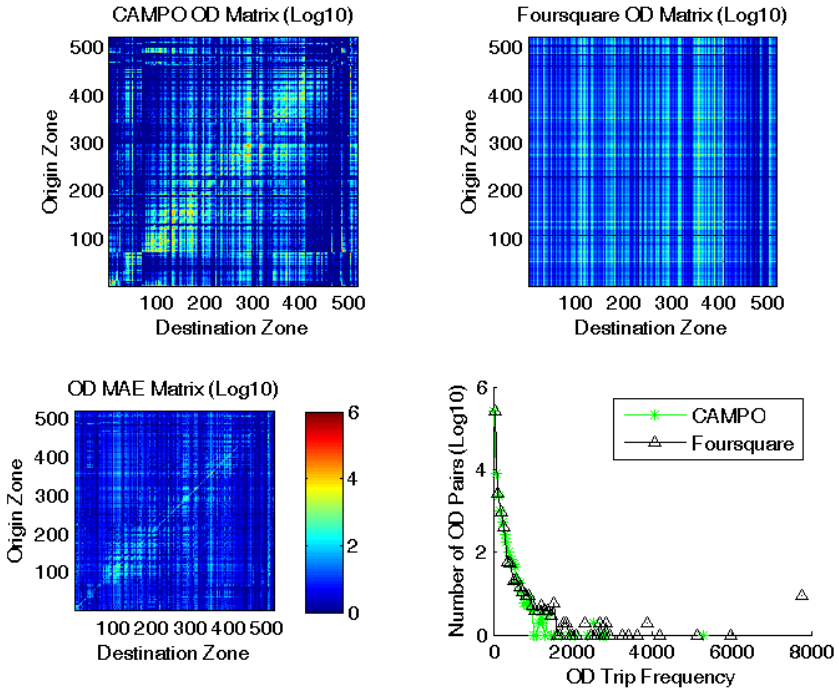


Figure 5.22: Intensity Diagrams for Gamma-Linear Doubly-Constrained Gravity Model

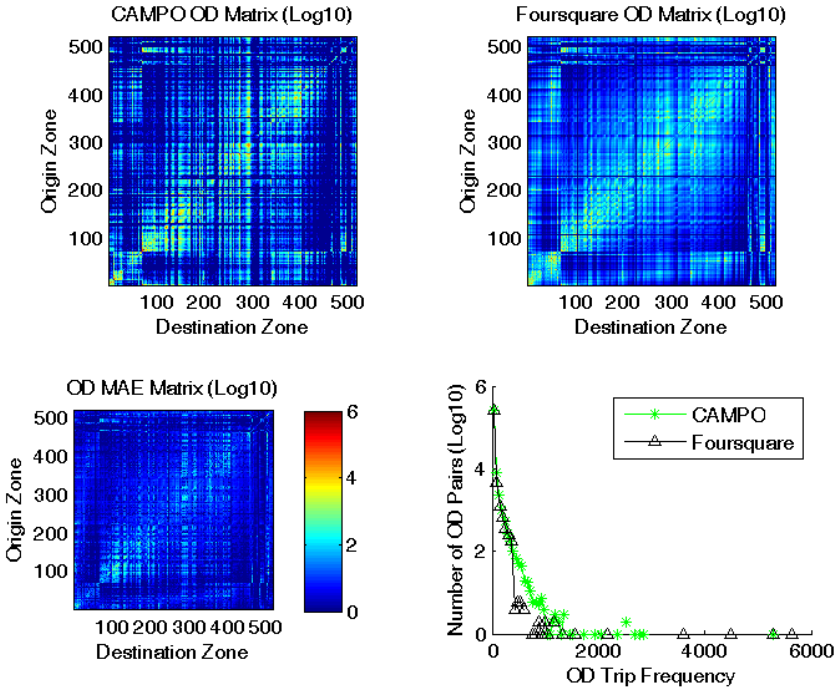


Figure 5.23: Intensity Diagrams for Gamma-Negative Exponential Doubly-Constrained Gravity Model

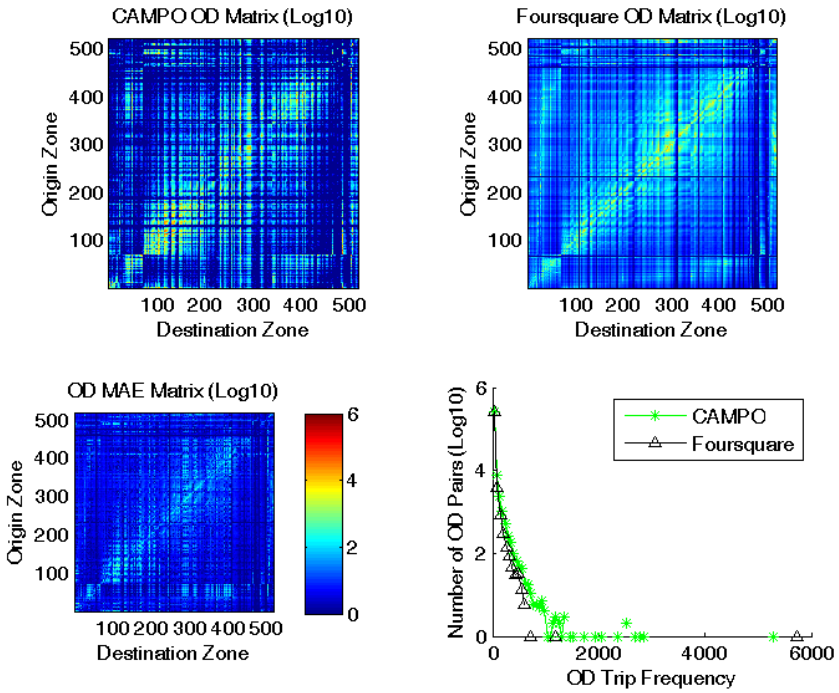


Figure 5.24: Intensity Diagrams for Gamma-Gamma Doubly-Constrained Gravity Model

Selection of "Best" Models

Based on the nine measurable criteria (CR, ME, MAE, FR, Swap Ratio, Production similarity – both versions, and Attraction similarity– both versions) used for analyzing the nine doubly-constrained models, a simple non-weighted rating system was used to determine which model or models performed the "best." Based on the average value for the nine criteria, the following models were found to be "best" performers and will be used for comparison with the peer-to-peer models:

1.) Linear - Negative exponential, Negative exponential - Negative exponential (tied for best)
2.) Gamma - Linear
3.) Gamma - Negative exponential

Peer-to-Peer Model Results

The peer-to-peer models have been analyzed using the methodology described in the previous chapter and shown above for the doubly-constrained gravity models. The nine peer-to-peer models will be discussed in this section to determine the "best" versions of the model that will be used for comparison to the doubly-constrained models.

Coincidence Ratio Analysis

As described in the doubly-constrained gravity component of this chapter, a coincidence ratio (CR) analysis of how "closely" the peer-to-peer models trip distributions came to matching the CAMPO trip distributions was done. Table 5.9 provides the CR values for each of the nine models. Similar to the doubly-constrained models, the table shows the limited sensitivity of the models to the short trip distance friction function used and higher sensitivity toward the long trip distance friction functions used. Models that have the linear long trip function performed significantly better than the other models with the best performing models as follows:

1.) Linear - Linear (0.9772)

2.) Gamma - Linear (0.9608)

3.) Negative exponential - Linear (0.8997)

Peer-to-Peer

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	**0.9772**	0.4613	0.4622
	Neg. Exp.	0.8997	0.4765	0.5102
	Gamma	0.9608	0.4904	0.5102

Table 5.9: Resulting Coincidence Ratio for Peer-to-Peer Models

Mean Error Analysis

Table 5.10 provides the mean error as calculated for each of the nine peer-to-peer models. While the values range between positive and negative, all are extremely small in magnitude indicating minimal bias exists with respect to the origin-destination matrix creation. The "best" performers are the following models:

1.) Negative exponential - Linear (6.5656E-15)

2.) Linear - Linear (1.7006E-14)

3.) Gamma - Negative exponential (3.7564E-14)

Peer-to-Peer

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	1.7006E-14	5.3547E-14	-7.8518E-14
	Neg. Exp.	**6.5656E-15**	-4.0026E-14	-7.8087E-14
	Gamma	9.3425E-14	3.7564E-14	6.8131E-14

Table 5.10: Resulting Mean Error for Peer-to-Peer Models

Mean Absolute Error Analysis

With respect to the MAE, the models that had the lowest error were those that used the linear long trip component (Table 5.11). The models that used the negative and gamma long trip components had similar errors that were noticeably larger than the linear long trip models. The following provides the ranking order for the linear long trip models:

1.) Negative exponential - Linear (9.3329)

2.) Gamma - Linear (9.5713)

3.) Linear - Linear (9.5806)

Peer-to-Peer

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	9.5806	12.2691	12.3202
	Neg. Exp.	**9.3329**	12.7081	12.0454
	Gamma	9.5713	12.4273	12.0607

Table 5.11: Resulting Mean Absolute Error for Peer-to-Peer Models

Frequency Ratio Analysis

Comparing the frequency of trips created by the proposed models to those from the CAMPO model was done to determine the method's abilities (Table 5.12). While all models appear to be in the 90% or greater rating, the models within the linear long trip grouping performed noticeably better than the other models. The following models are noted as the "best" performers:

1.) Negative exponential - Linear (0.9720)

2.) Gamma - Linear (0.9715)

3.) Linear - Linear (0.9695)

Peer-to-Peer

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	0.9695	0.9284	0.9251
	Neg. Exp.	**0.9720**	0.9012	0.9044
	Gamma	0.9715	0.9217	0.9044

Table 5.12: Resulting Frequency Ratio for Peer-to-Peer Models

Swap Ratio Analysis

For the swap ratio analysis, the comparison resulted in similar trends as the MAE analysis. The models with the linear function for long trips had the lowest values (Table

5.13). The models with the negative exponential and gamma function for long trips were markedly worse performers, with the negative exponential-negative exponential model preforming the worst. The best performing models are ranked as follows:

1.) Negative exponential - Linear (27.2053)

2.) Linear - Gamma (27.3535)

3.) Linear - Linear (27.4773)

Peer-to-Peer

		Long Trips		
		Linear	Neg. Exp.	Gamma
Short Trips	Linear	27.4773	28.9969	28.9246
	Neg. Exp.	**27.2053**	30.0939	29.8166
	Gamma	27.3535	28.8518	29.8148

Table 5.13: Resulting Swap Ratio for Peer-to-Peer Models

Trip Length Distribution Analysis

As was done for the doubly-constrained models, a graphical representation of trip length distributions was created in MATLAB for each of the nine peer-to-peer models (Figures 5.25 to 5.33). Examining the graphics reveals that the models with the linear friction function used for the long trips have the closest trip length distribution (Figures 5.25, 5.28, and 5.31). The models with the negative exponential and gamma friction functions for the long trips show significant under estimation for shorter distance trips and significant over estimation for longer distance trips. The "best" performing models are ranked as follows:

1.) Negative exponential - Linear

2.) Linear - Linear

3.) Gamma – Linear

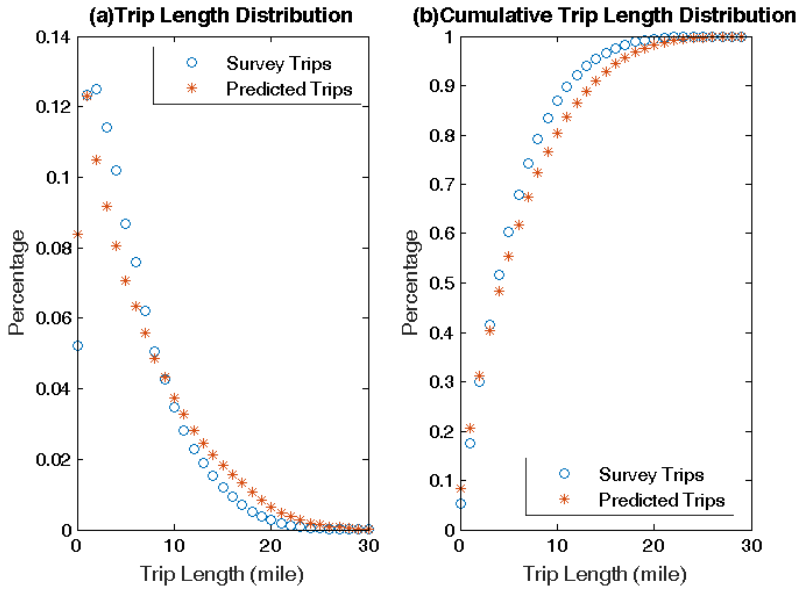


Figure 5.25: Trip Length Distributions for Linear-Linear Peer-to-Peer Model

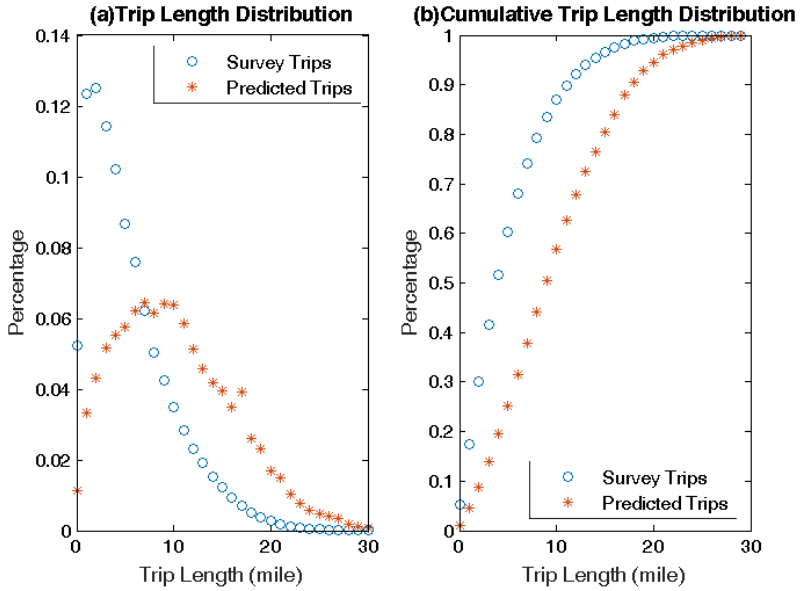


Figure 5.26: Trip Length Distributions for Linear-Negative Exponential Peer-to-Peer Model

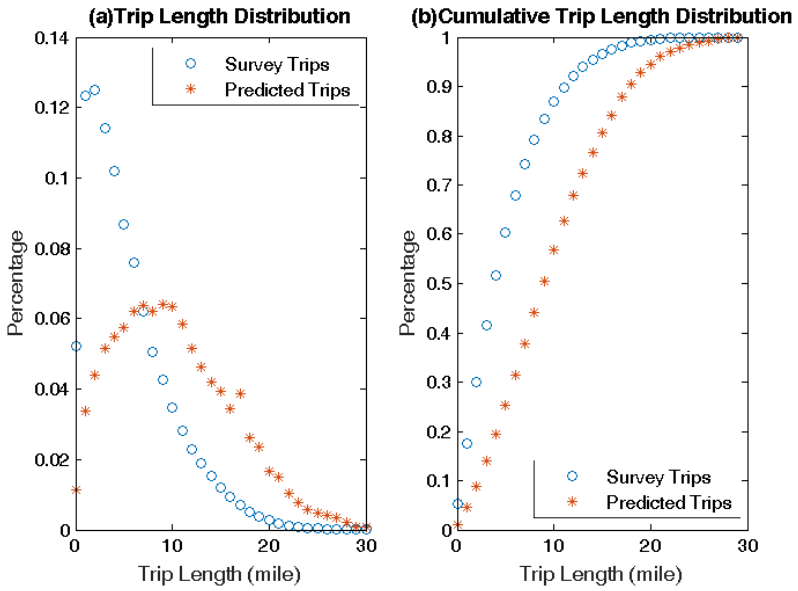


Figure 5.27: Trip Length Distributions for Linear-Gamma Peer-to-Peer Model

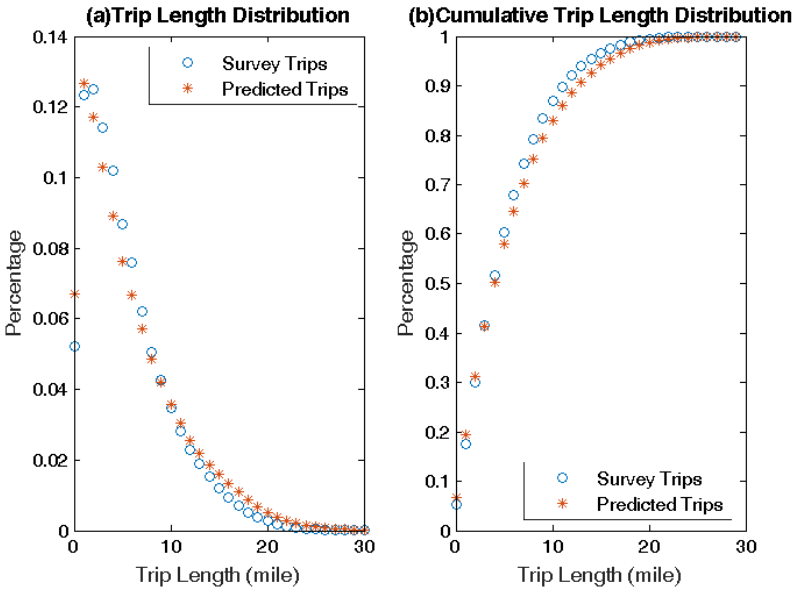


Figure 5.28: Trip Length Distributions for Negative Exponential-Linear Peer-to-Peer Model

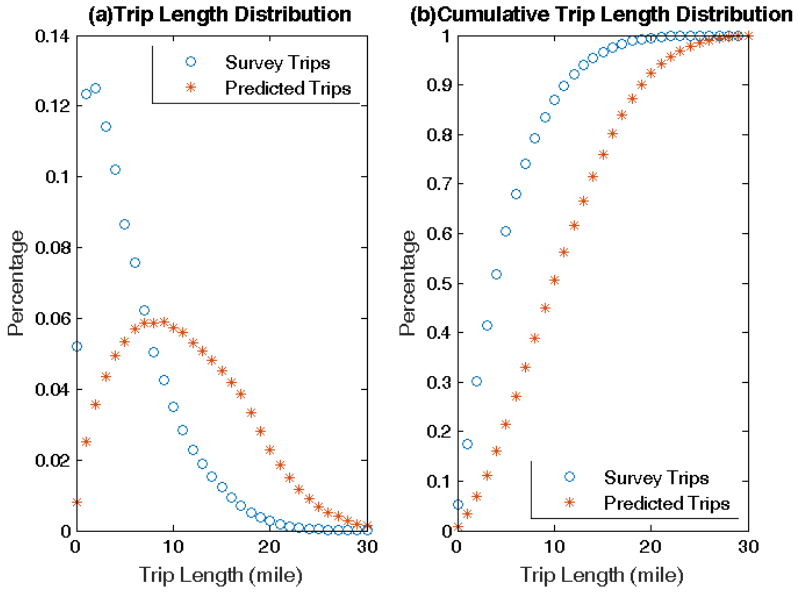


Figure 5.29: Trip Length Distributions for Negative Exponential-Negative Exponential Peer-to-Peer Model

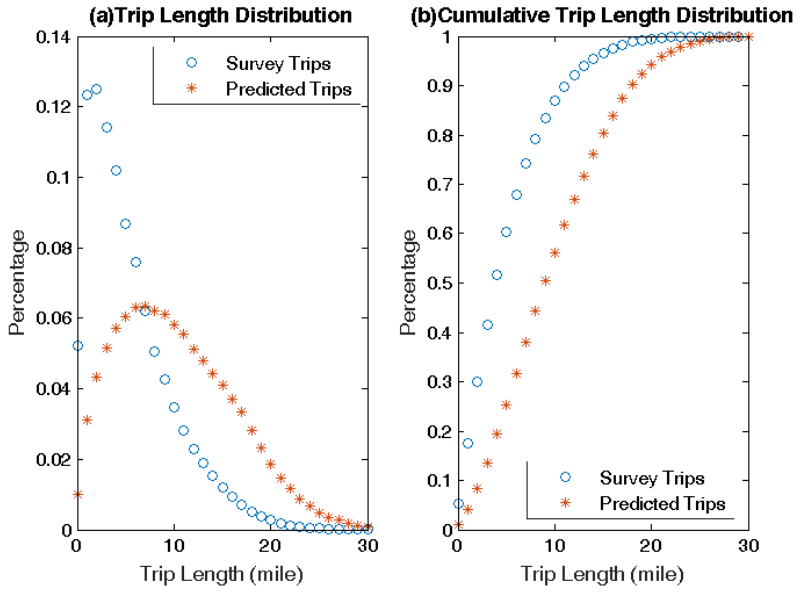


Figure 5.30: Trip Length Distributions for Negative Exponential-Gamma Peer-to-Peer Model

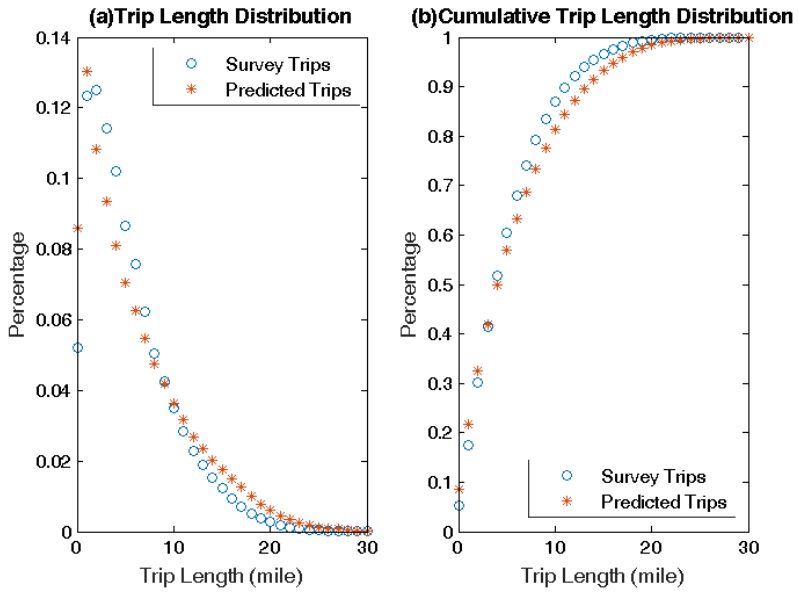


Figure 5.31: Trip Length Distributions for Gamma-Linear Peer-to-Peer Model

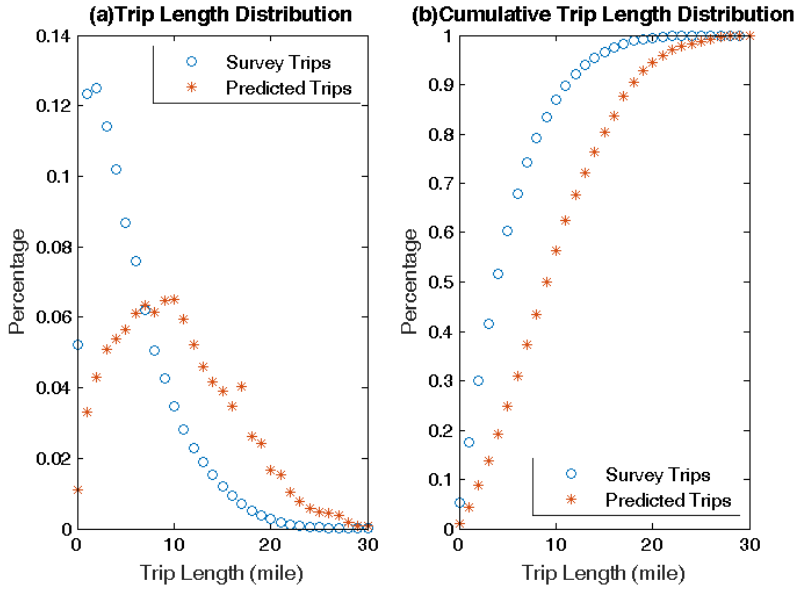


Figure 5.32: Trip Length Distributions for Gamma-Negative Exponential Peer-to-Peer Model

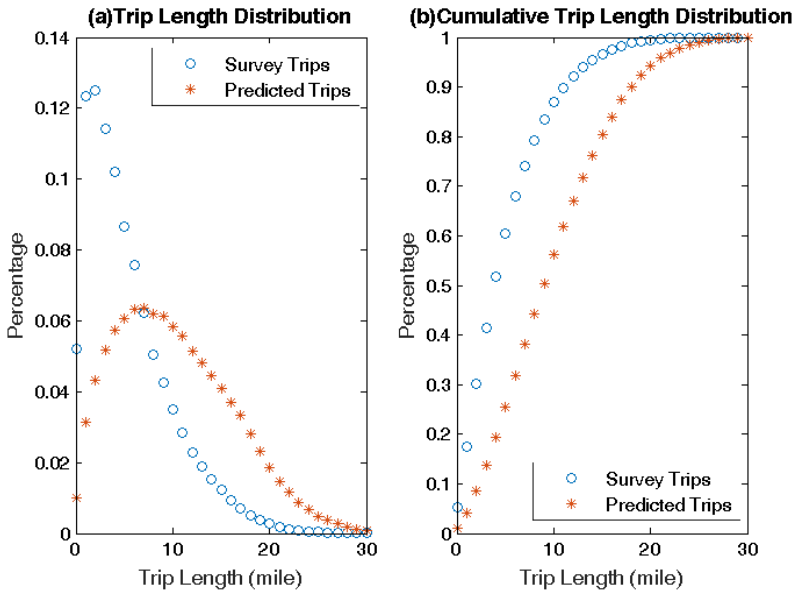


Figure 5.33: Trip Length Distributions for Gamma-Gamma Peer-to-Peer Model

Production and Attraction Analysis

Trip productions and attractions were examined using the same methodology as was done for the doubly-constrained gravity models. The ArcGIS created figures were created for the nine peer-to-peer models using the same color scaling as described previously. Figures 5.12 and 5.14 above should be used as the reference CAMPO production and attractions figures.

Figure 5.34 provides the TAZ productions generated by each of the proposed peer-to-peer models. The same comparison effort with respect to matching, given a "Y," one shade darker or lighter, "C," and neither, "N," was done. Table 5.14 provides the statistics for each of models. Based on these statistics, the models with the most TAZ with the most categorically the same productions, "Y," are the following:

1.) Linear - Gamma

2.) Negative exponential - Linear

3.) Gamma - Negative exponential

With respect to the models with the most similar categorization, “Y” and “C,” the models that ranked the highest were:

1.) Negative exponential - Negative exponential, Gamma - Negative exponential (tied)

2.) Linear - Gamma

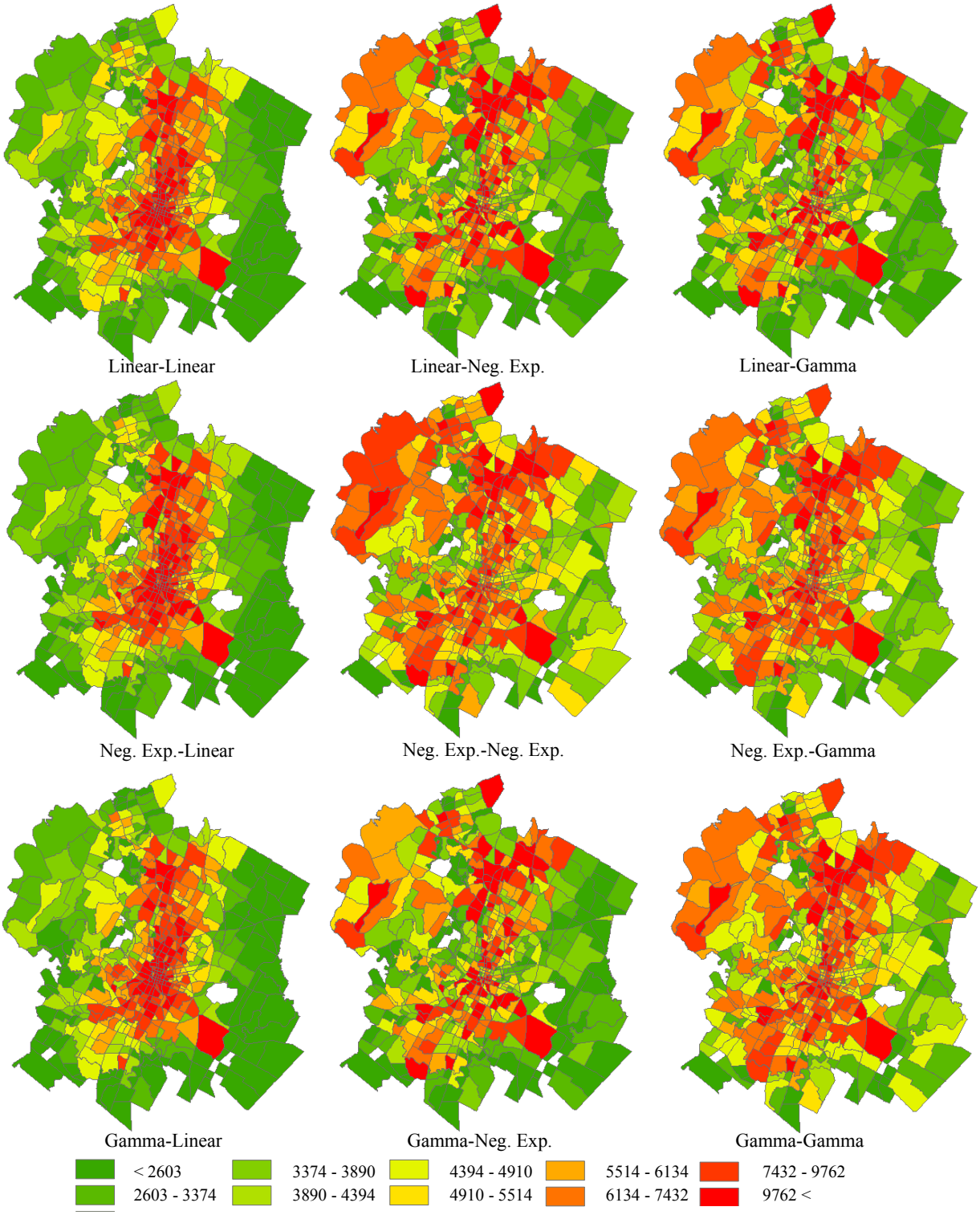


Figure 5.34: Trip Productions for the Proposed Peer-to-Peer Models

	Linear-Linear		**Linear-Neg. Exp.**		**Linear-Gamma**	
	#	%	#	%	#	%
Y	137	26.35	129	24.81	138	**26.54**
C	106	20.38	140	26.92	126	24.23
N	277	53.27	251	48.27	256	49.23
Y+C	243	46.73	269	**51.73**	264	**50.77**
	Neg. Exp.-Linear		Neg. Exp.-Neg. Exp.		Neg. Exp.-Gamma	
	#	%	#	%	#	%
Y	139	**26.73**	98	18.85	102	19.62
C	106	20.38	144	27.69	149	28.65
N	275	52.88	278	53.46	269	51.73
Y+C	245	47.12	242	46.54	251	48.27
	Gamma-Linear		**Gamma-Neg. Exp.**		Gamma-Gamma	
	#	%	#	%	#	%
Y	136	26.15	146	**28.08**	102	19.62
C	108	20.77	123	23.65	148	28.46
N	276	53.08	251	48.27	270	51.92
Y+C	244	46.92	269	**51.73**	250	48.08

Table 5.14: TAZ Production Rate Graphical Similarity Statistics Peer-to-Peer Models

Following the steps described previously, the trip attractions were examined. Figure 5.35 provides the attractions generated by each of the proposed peer-to-peer models, while Table 5.15 provides a breakdown of these statistics for the productions. Examination of the nine models showed that the models with the most TAZs with the same categorization, "Y," of attractions were as follows:

1.) Gamma - Negative exponential

2.) Negative exponential - Linear, Gamma - Linear (tied)

With respect to the models with the most similar categorization, "Y" and "C," the models that ranked the highest were:

1.) Linear - Gamma

2.) Gamma - Negative exponential

3.) Linear - Negative exponential

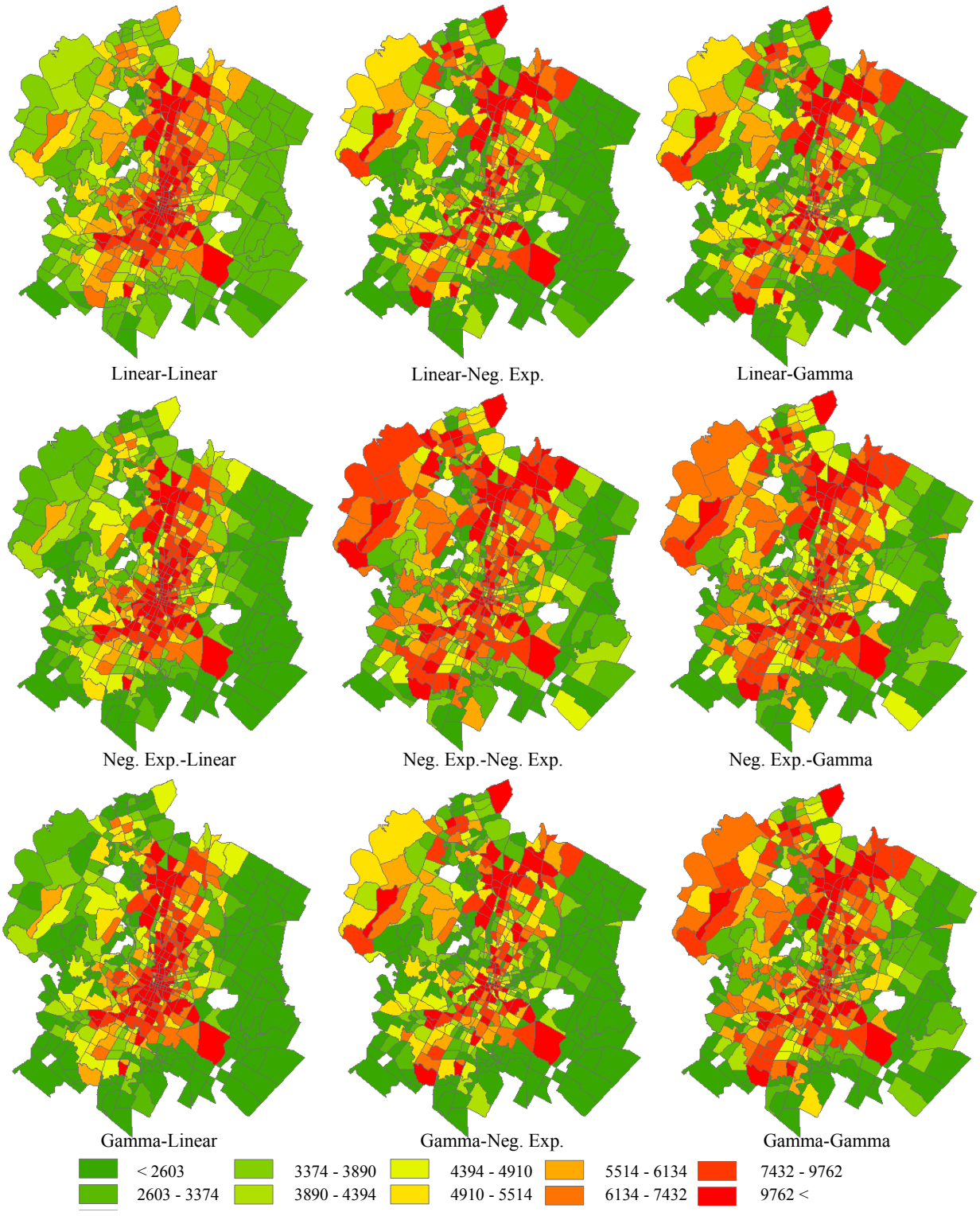


Figure 5.35: Trip Attractions for the Proposed Peer-to-Peer Models

	Linear-Linear		**Linear-Neg. Exp.**		**Linear-Gamma**	
	#	%	#	%	#	%
Y	227	43.65	223	42.88	222	42.69
C	102	19.62	110	21.15	113	21.73
N	191	36.73	187	35.96	185	35.58
Y+C	329	63.27	333	**64.04**	335	**64.42**
	Neg. Exp.-Linear		Neg. Exp.-Neg. Exp.		Neg. Exp.-Gamma	
	#	%	#	%	#	%
Y	228	**43.85**	171	32.88	180	34.62
C	100	19.23	121	23.27	126	24.23
N	192	36.92	228	43.85	214	41.15
Y+C	328	63.08	292	56.15	306	58.85
	Gamma-Linear		**Gamma-Neg. Exp.**		Gamma-Gamma	
	#	%	#	%	#	%
Y	228	**43.85**	233	**44.81**	180	34.62
C	99	19.04	101	19.42	126	24.23
N	193	37.12	186	35.77	214	41.15
Y+C	327	62.88	334	**64.23**	306	58.85

Table 5.15: TAZ Attraction Rate Graphical Similarity Statistics Peer-to-Peer Models

Intensity Analysis

Each of the peer-to-peer models was scrutinized using the intensity analysis described in Chapter 4 (Figures 5.36 through 5.44). These graphics show areas of higher flow between origin-destination TAZ pairs via the lighter striations. Similarly the MAE intensity analysis shows the error magnitude for the proposed model, while the origin-destination trip frequency analysis shows where over and under estimation occurs within the proposed peer-to-peer models. Through visual analysis, the following observations can be made:

Each of the peer-to-peer models was scrutinized using the intensity analysis described in Chapter 4 (Figures 5.36 through 5.44). These graphics show areas of higher flow between origin-destination TAZ pairs via the lighter striations. Similarly the MAE intensity analysis shows the error magnitude for the proposed model, while the origin-destination trip frequency analysis shows where over and under estimation occurs within the proposed peer-to-peer models. Through visual analysis, the following observations can be made:

1.) The models that used the linear component for long trips showed similar flow rates for intrazonal trips compared to the CAMPO model. Additionally, the striations within these models have similar shading as is shown in the CAMPO model. This is confirmed by the CR values from the earlier analysis for these models and through the MAE intensity graphic which shows the presence of dark shading throughout the graphics. However, the presence of some lightness along the 45° intrazonal line indicates that errors exist within these models.

2.) Models with the negative exponential and gamma long trip functions do not predict the intrazonal trips well, which is confirmed by the lighter coloring along the 45° line within the MAE intensity graphics, which indicate a larger error.

3.) The best performing model with respect to MAE intensity was the negative exponential-linear model (Figure 5.39).

4.) Examination of the OD trip frequency graphics for the linear short trip models (Figures 5.36, 5.37, and 5.38) showed each model's ability to closely represent the frequencies for lower values. However, the model for the linear long trip friction function (Figure 5.36) showed slight under

estimation for the larger values. The other two models, negative exponential and gamma, had over estimation with respect to the larger values.

5.) For the negative exponential and gamma short trip functions, the models with linear long trip functions (Figures 5.39 and 5.42) performed better with respect to the lower frequency values and had underestimation for the larger frequency values.

6.) The gamma-negative exponential model (Figure 5.43) showed good estimation of lower frequencies, but had over estimation for larger values especially the most extreme value.

7.) The other models showed slight under and over estimation for all of the frequencies.

8.) The model that best performed with respect to OD trip frequencies was the linear-linear model (Figure 5.36)

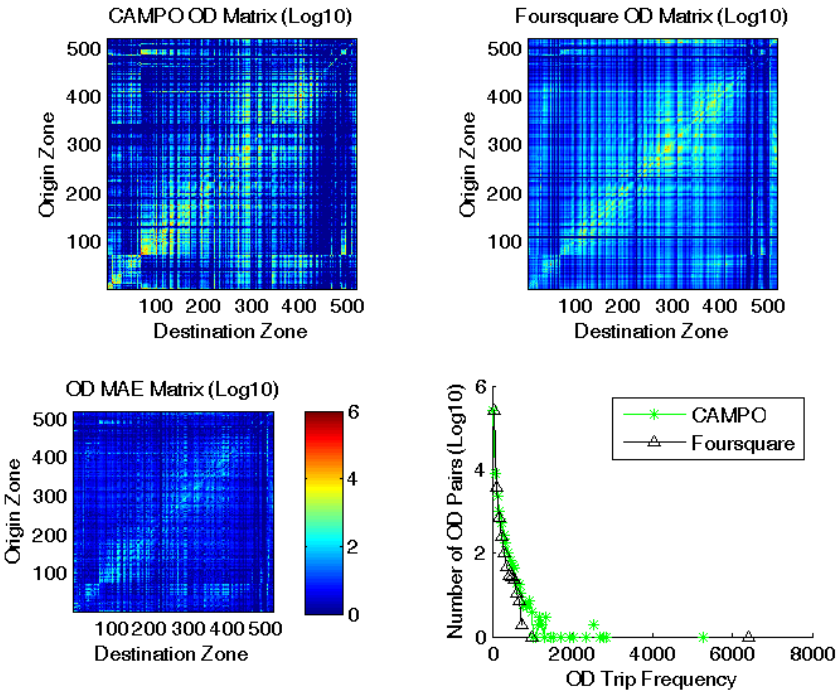


Figure 5.36: Intensity Diagrams for Linear-Linear Peer-to-Peer Model

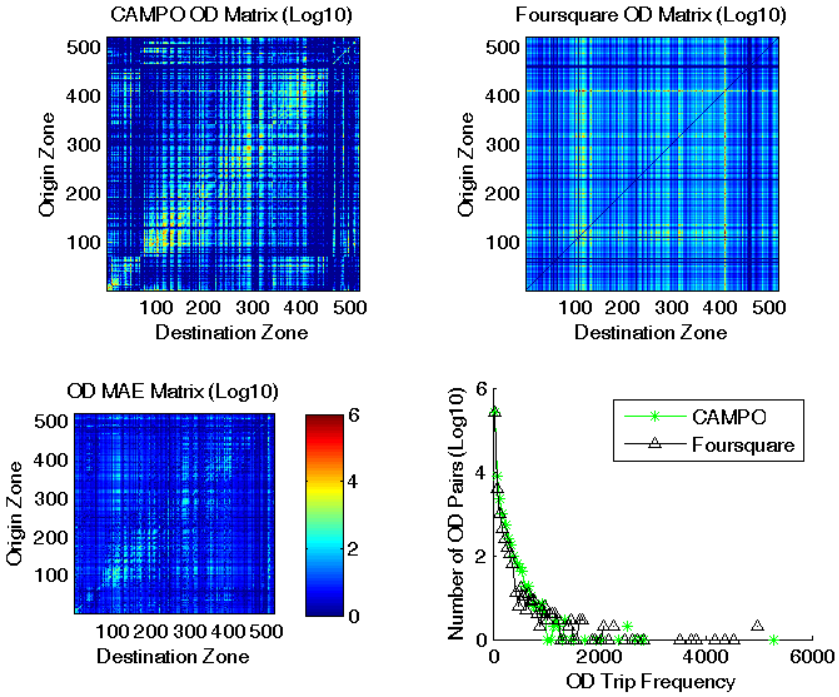


Figure 5.37: Intensity Diagrams for Linear-Negative Exponential Peer-to-Peer Model

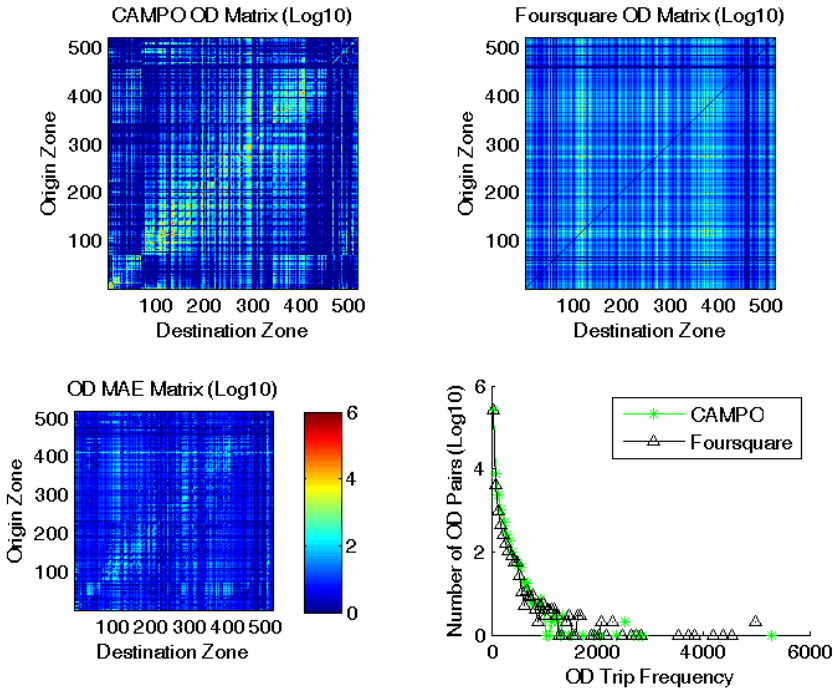


Figure 5.38: Intensity Diagrams for Linear-Gamma Peer-to-Peer Model

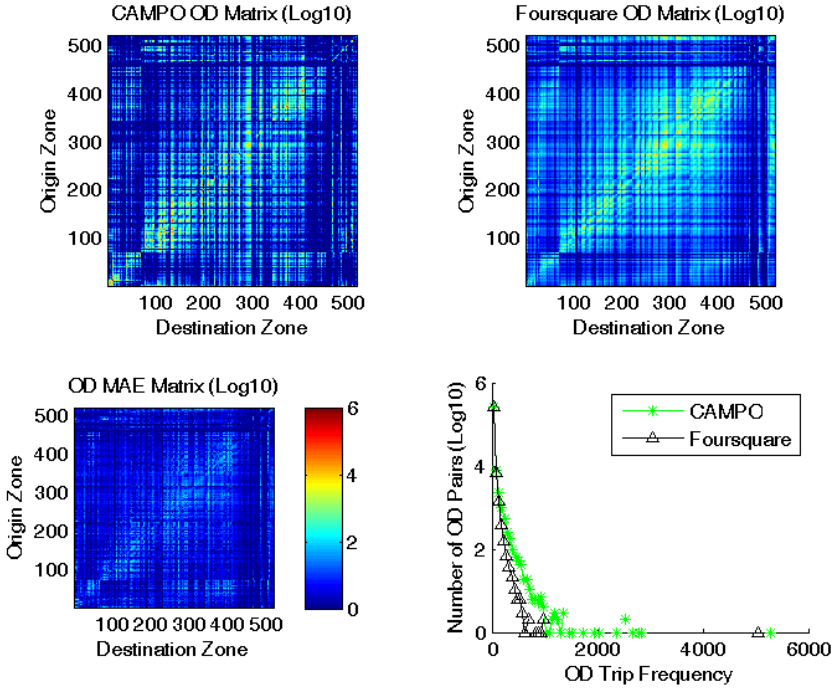


Figure 5.39: Intensity Diagrams for Negative Exponential-Linear Peer-to-Peer Model

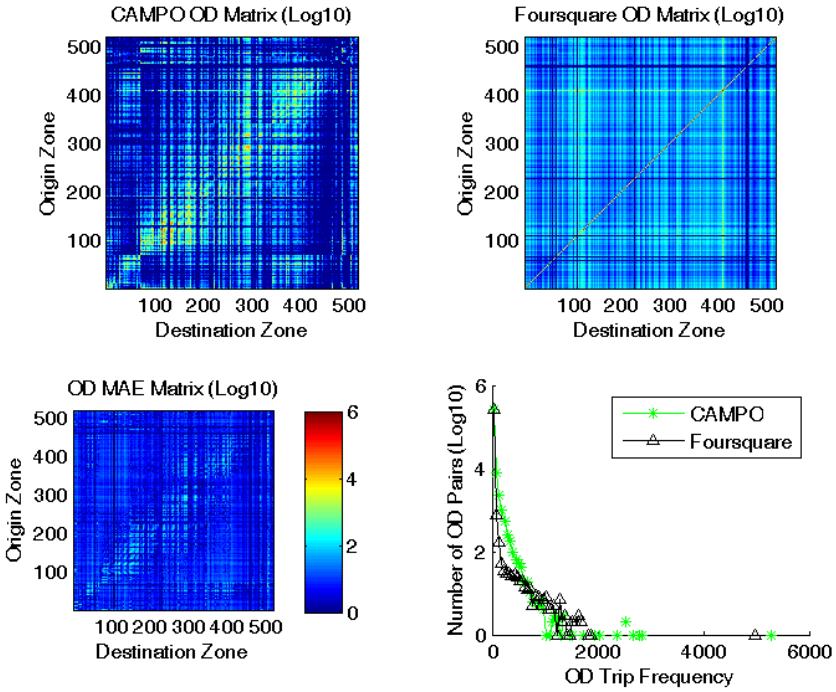


Figure 5.40: Intensity Diagrams for Negative Exponential-Negative Exponential Peer-to-Peer Model

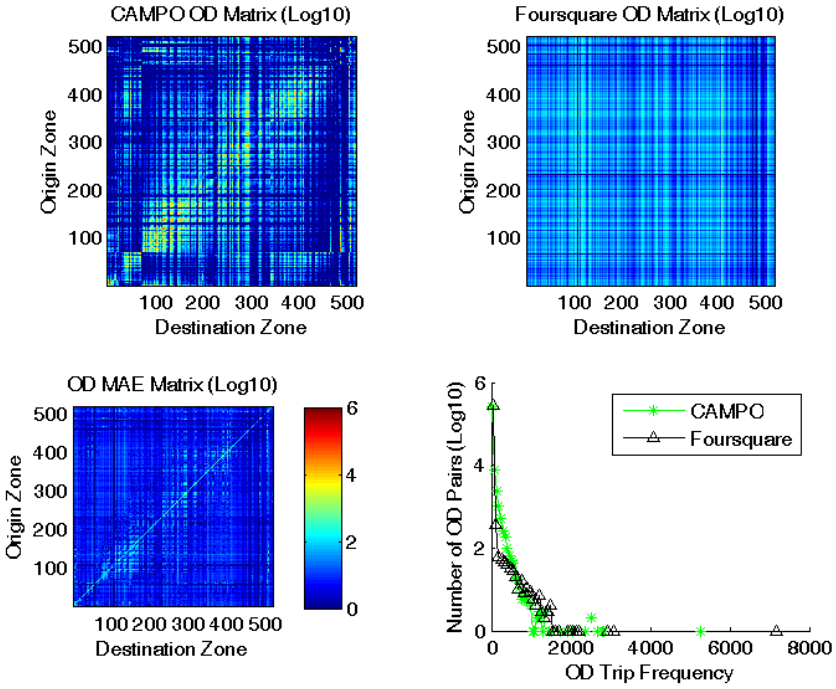


Figure 5.41: Intensity Diagrams for Negative Exponential-Gamma Peer-to-Peer Model

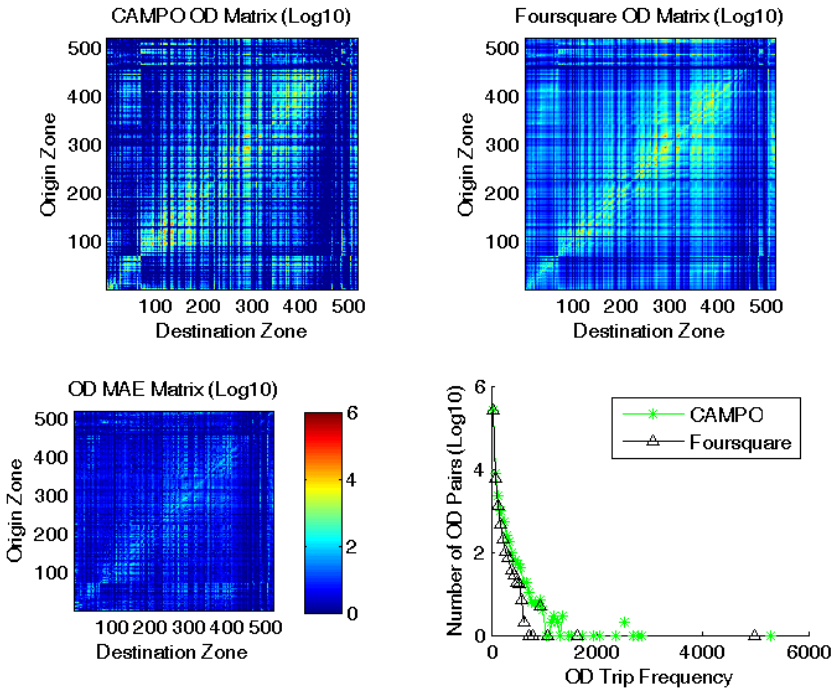


Figure 5.42: Intensity Diagrams for Gamma-Linear Peer-to-Peer Model l

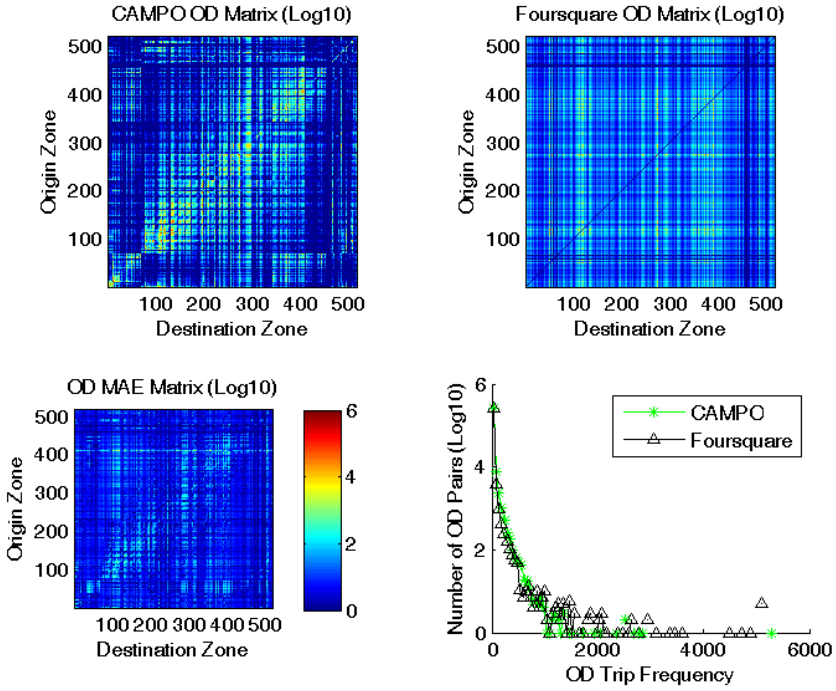


Figure 5.43: Intensity Diagrams for Gamma-Negative Exponential Peer-to-Peer Model

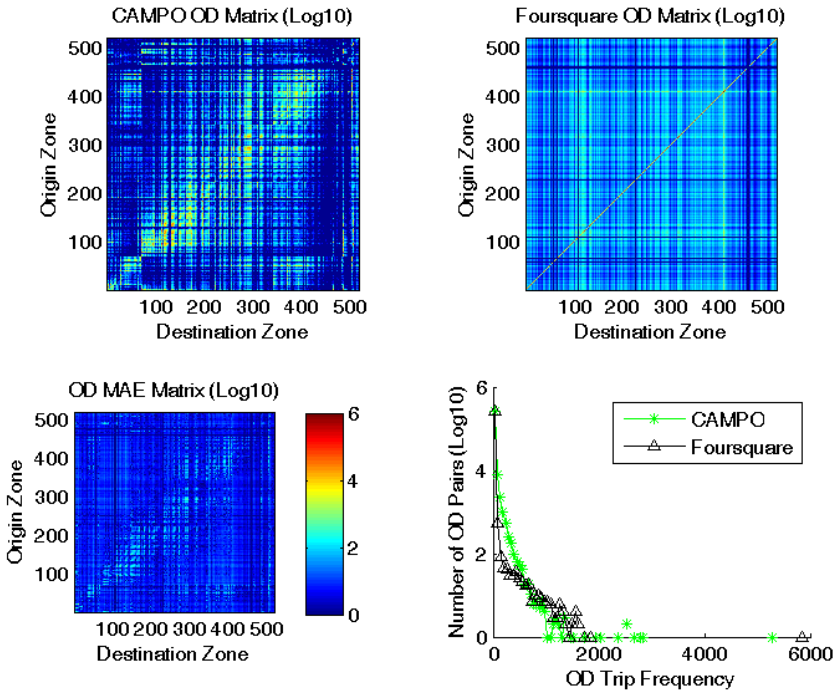


Figure 5.44: Intensity Diagrams for Gamma-Gamma Peer-to-Peer Model

Selection of "Best" Models

As was done for the doubly-constrained gravity model selection, the nine measurable criteria (CR, ME, MAE, FR, Swap Ratio, Production similarity – both versions, and Attraction similarity– both versions) were used to determine the "best" peer-to-peer models via the simple non-weighted rating system. Based on the average value for the nine criteria, the following peer-to-peer models were found to be "best" performers and will be used for comparison with the doubly-constrained gravity models:

1.) Negative exponential - Linear

2.) Gamma - Negative exponential

3.) Linear - Linear

4.) Gamma - Linear

Best Performing Model Comparisons

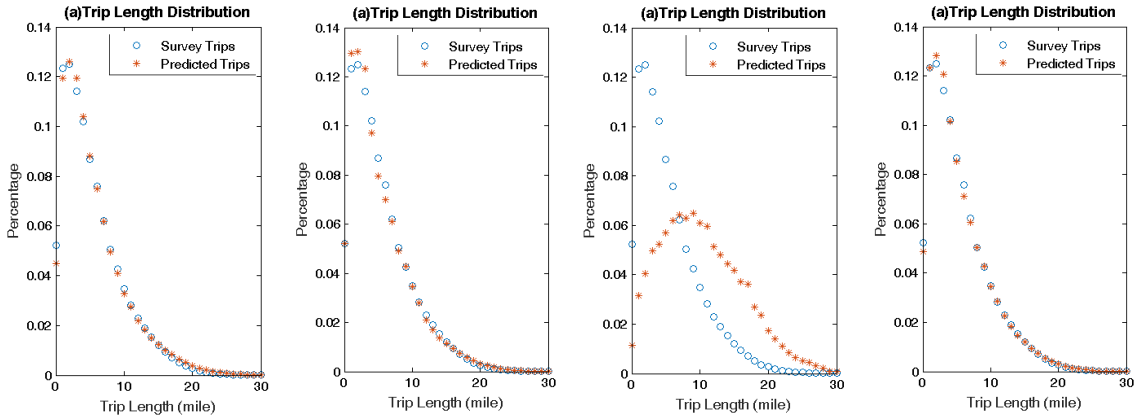
Using the previously identified doubly-constrained gravity and peer-to-peer models that were "best" performers, this section will compare these models to one another and identify the strengths and weakness of each. Using the following as the major areas of comparison, each estimation model will be analyzed: CR, MAE, FR, trip length distributions, production and attraction comparisons, and intensity analysis.

With respect to the CR, MAE, and FR values, Table 5.16 shows the values for each of the "best" models and provides a comparative rank. A comparison of CR for the "best" models reveals that the two best performing models are the peer-to-peer linear-linear (0.9772) and gamma-linear (0.9608). The third best performing model is the doubly-constrained gravity linear-negative exponential model (0.9576). An examination of the MAE values shows that the top three performing models are all within the peer-to-peer group: negative exponential-linear (9.3329), gamma-linear (9.5806), and linear-linear (9.5806). This same trend is seen in the values for FR which has the same peer-to-peer models as the top three: negative exponential-linear (0.9720), gamma-linear (0.9715), and linear-linear (0.9695).

	Model Name	CR	Rank	MAE	Rank	FR	Rank
Doubly-Constrained	**Linear - Neg. Exp.**	**0.9576**	**3**	9.9869	4	0.9565	5
	Neg. Exp. - Neg. Exp.	0.9283	5	10.5308	6	0.9287	7
	Gamma - Linear	0.4961	7	13.0517	8	0.9477	6
	Gamma - Neg. Exp.	0.9449	4	10.1379	5	0.9588	4
Peer-to-Peer	**Linear - Linear**	**0.9772**	**1**	**9.5806**	**3**	**0.9695**	**3**
	Neg. Exp. - Linear	0.8997	6	**9.3329**	**1**	**0.9720**	**1**
	Gamma - Linear	**0.9608**	**2**	**9.5713**	**2**	**0.9715**	**2**
	Gamma - Neg. Exp.	0.4904	8	12.427	7	0.9217	8

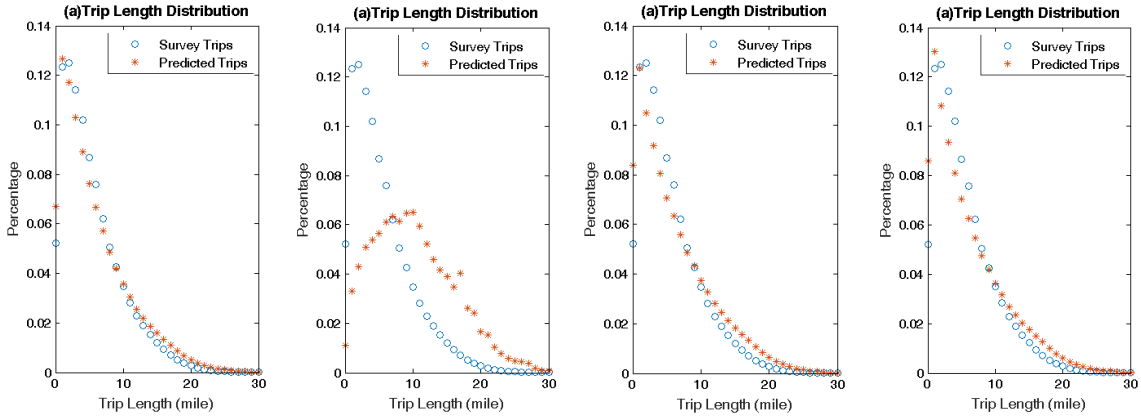
Table 5.16: Best Models Comparisons – CR, MAE, and FR

Figure 5.45 shows the trip length distributions for each of the comparison models and Figure 5.46 shows the cumulative trip length distributions for each of the comparison models. Using a visual inspection of the graphics in each figure, it appears that the top three performing models for the trip length distributions are from the doubly-constrained gravity models and are ranked as follows: gamma - negative exponential, linear - negative exponential, and negative exponential - negative exponential. For the cumulative trip length distributions, the top three models are the same three models from the doubly-constrained gravity models but are ranked as follows: negative exponential - negative exponential, linear - negative exponential, and gamma - negative exponential.



Linear - Neg. Exp. | Neg. Exp. - Neg. Exp. | Gamma - Linear. | Gamma - Neg. Exp.

Doubly-Constrained Gravity Models



Neg. Exp. - Linear | Gamma - Neg. Exp. | Linear - Linear | Gamma - Linear

Peer-to-Peer Models

Figure 5.45: Trip Length Distributions for Comparison Models

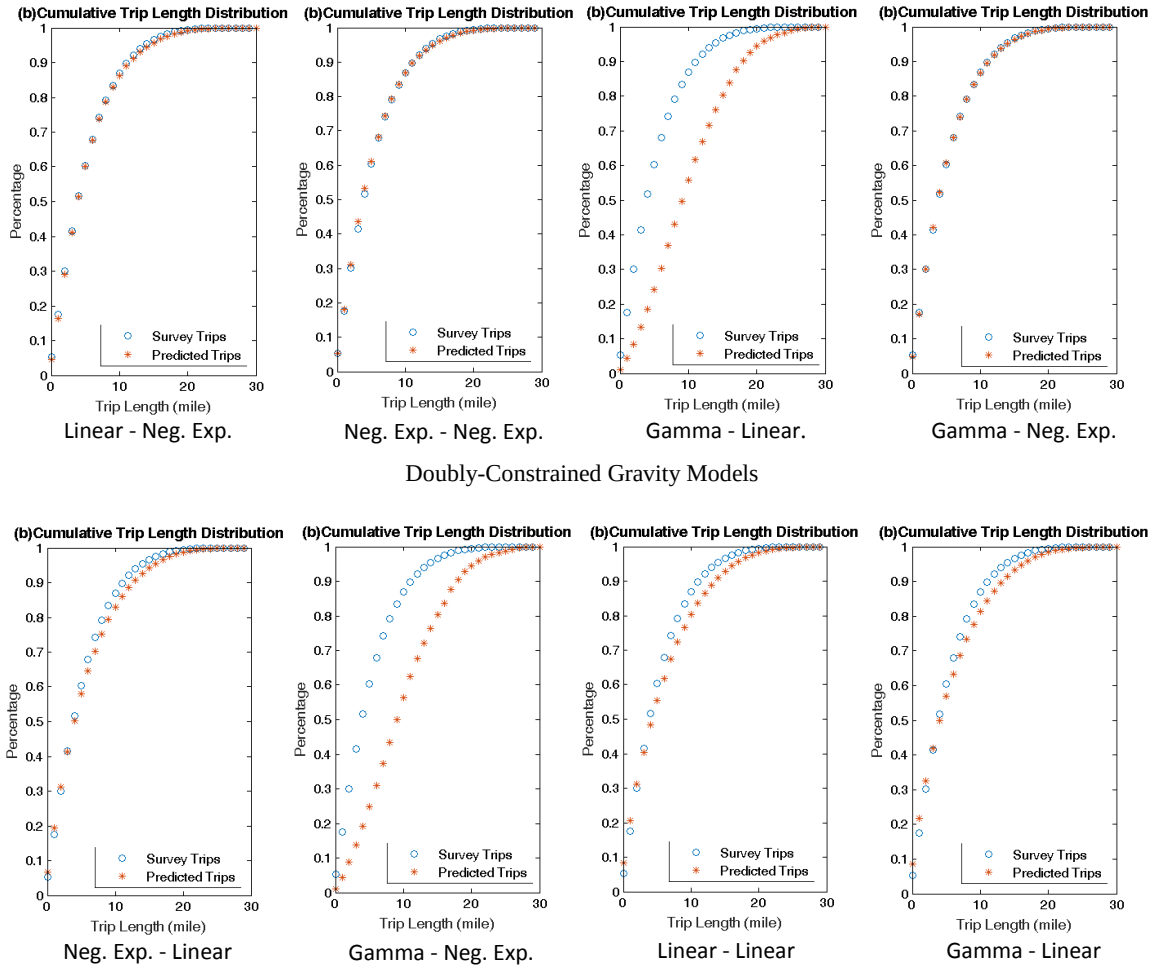


Figure 5.46: Cumulative Trip Length Distributions for Comparison Models

Using the trip productions and attractions graphics that were crated in ArcGIS previously, Table 5.17 presents the results for the productions and attractions analysis. The "Y" category shows the total number of TAZs that matched the exact color category and the "Y+C" category shows the total number of TAZs that were the same plus those within one shade darker or lighter. Examining first the productions, it is notable that out of the 520 TAZs the largest number of TAZs that any of the "best" models had that matched was 146 or 28.08% found from the peer-to-peer gamma-negative exponential

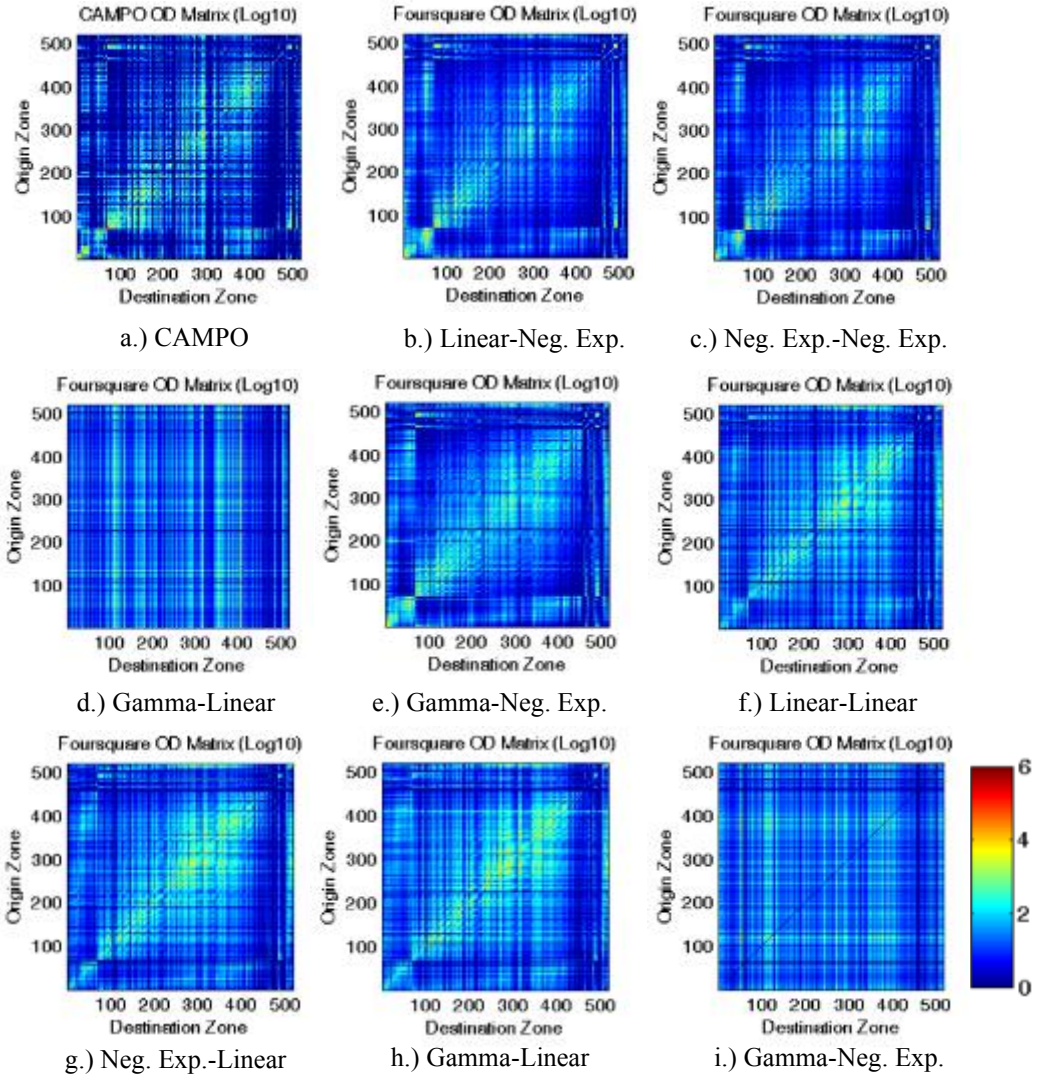
model. With respect to the combination of matches and within one shade TAZs, the production results of the models were better with the highest amount of matches coming from the peer-to-peer gamma-negative exponential model at 269 or 51.73%. All models examined had better results with respect to the attractions. The highest number of TAZ matches was found from the doubly-constrained gravity gamma-linear model, which had a total of 265 or 50.96%. For the combination of matches and within one shade TAZs, the same doubly-constrained model had the largest number, 355 or 68.27%. The second and third place results are highlighted within Table 5.17. It should be noted that the peer-to-peer gamma-negative exponential model was the best and second best model for all categories of productions and attractions, respectively.

		Productions				Attractions			
	Model Name	**Y**	**Rank**	**Y+C**	**Rank**	**Y**	**Rank**	**Y+C**	**Rank**
Doubly-Constrained	Linear - Neg. Exp.	104	6	224	7	223	7	323	7
	Neg. Exp. - Neg. Exp.	110	5	226	6	**229**	**3**	**332**	**3**
	Gamma - Linear	95	7	**248**	**2**	**265**	**1**	**355**	**1**
	Gamma - Neg. Exp.	91	8	197	8	170	8	287	8
Peer-to-Peer	**Linear - Linear**	**137**	**3**	243	5	227	6	329	4
	Neg. Exp. - Linear	**139**	**2**	**245**	**3**	228	4	328	5
	Gamma - Linear	136	4	244	4	228	4	327	6
	Gamma - Neg. Exp.	**146**	**1**	**269**	**1**	**233**	**2**	**334**	**2**

Table 5.17: Best Models Comparisons – Productions and Attraction Matching

A visual examination of the intensity graphics, Figure 5.47, for all of comparison models gives additional insight into the "best" models. Models b, c, d, and e are the models from the doubly-constrained gravity models and f, g, h, and i are models from the peer-to-peer modeling group. Model a is the CAMPO matrix that is used for comparison.

Examining the overall color striations, the doubly-constrained models appear to better predict the overall zone flows throughout the study area. The "best" performing models are the doubly-constrained linear-negative exponential and negative exponential-negative exponential models which have more similar striations throughout the matrices than the other comparison models. It should be noted that the peer-to-peer models do have areas of noticeably more pronounced consistent striations when compared to the CAMPO model, but the shading throughout is less consistent than the doubly-constrained models. This indicates further examination into the locational data characteristics where the peer-to-peer model was more successful should be done in the future. When the examination of the intrazonal trips (the 45° line within the matrix) was done, the peer-to-peer models linear-linear, negative exponential-linear, and gamma-linear have closer coloring to the CAMPO model than the other matrices. This may be a function of the inclusion of venues in the calculation of intrazonal trips that does not exist within the doubly-constrained formulation. The doubly-constrained models, with the exception of the gamma-linear model, have some similarities to the CAMPO model with respect to intrazonal trip estimations, but are not as pronounced as those seen in the peer-to-peer models indicating a limitation to the models' capabilities with respect to these trips.



a.) CAMPO b.) Linear-Neg. Exp. c.) Neg. Exp.-Neg. Exp.

d.) Gamma-Linear e.) Gamma-Neg. Exp. f.) Linear-Linear

g.) Neg. Exp.-Linear h.) Gamma-Linear i.) Gamma-Neg. Exp.

Figure 5.47: OD Intensity Comparison Matrices

The visual examination of the MAE Intensity graphics (Figure 5.48) reveals that all of the models have errors along the intrazonal trip line, indicating a potential need for improvement in the methodologies to better address these trips. For this figure the doubly-constrained models are models a, b, c, and d, while the peer-to-peer models are e, f, g, and h. Looking at the overall gradation of the figures, the doubly-constrained gravity

model that uses the linear-negative exponential friction function appears to perform the "best". However, there are noticeable areas of larger errors present particularly along the intrazonal line and between the 300 and 450 TAZs. The next "best" performer was identified as the peer-to-peer model that uses the negative exponential-linear friction function. This model has fairly consistent color throughout indicating more consistency within the errors produced.

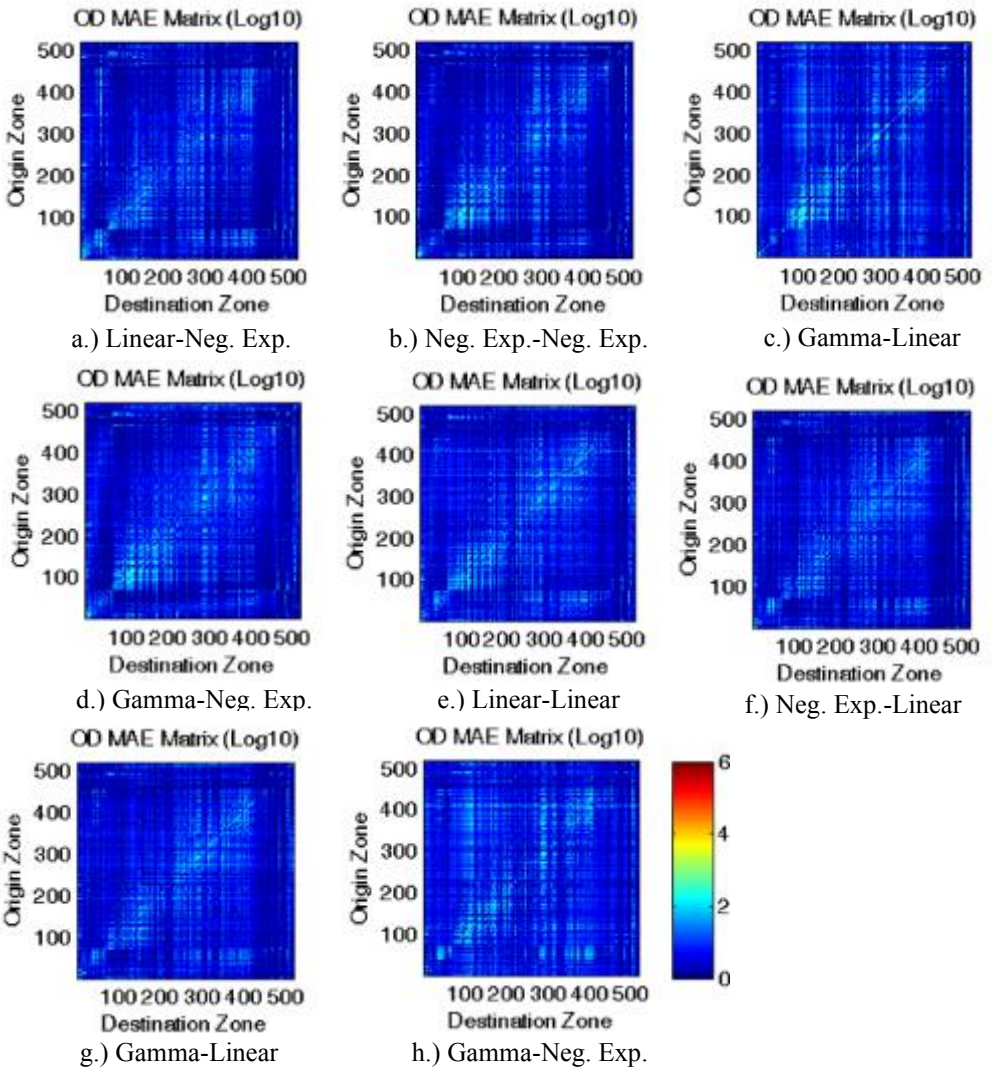


Figure 5.48: MAE Intensity Comparison Matrices

A similar visual examination of the origin-destination (OD) trip frequency intensities (Figure 5.49) was performed as was done previously. This examination revealed that the doubly-constrained gravity model with the linear-negative exponential friction function was the closest at predicting OD trip frequencies compared to the other "best" models. The second and third best were found within the peer-to-peer models and were the linear-linear and gamma-linear, respectively. The linear-linear model was noted to have under estimation, while the gamma-linear model had slightly more under estimation than the linear-linear model. These results indicated the potential for slight improvements within all of the models with respect to trip frequency estimation, since each model showed fairly close estimation on average.

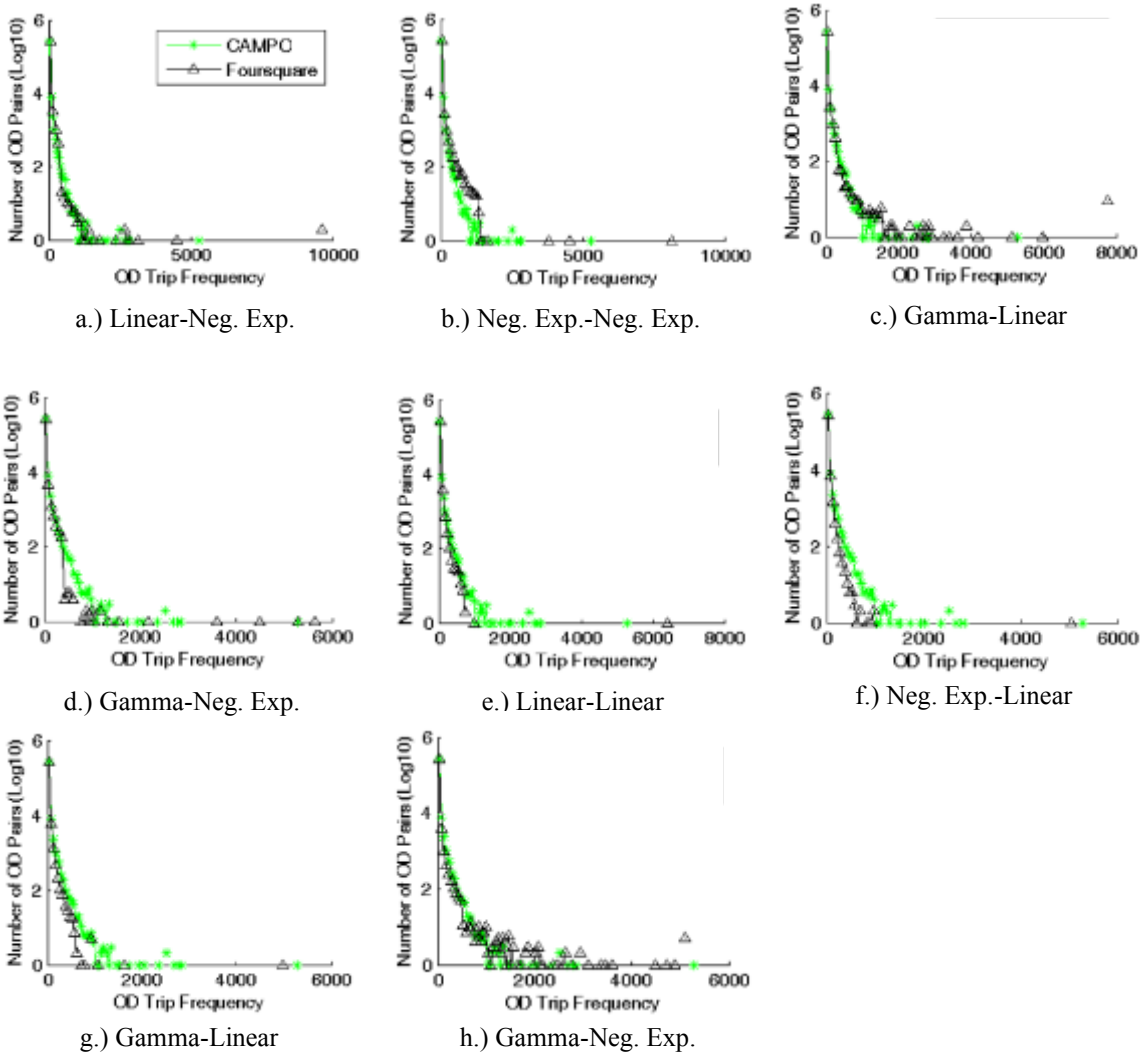


Figure 5.49: OD Trip Frequency Intensity Comparison

Similar to the steps undertaken to determine the best models from the base models of doubly-constrained gravity and peer-to-peer, the quantifiable characteristics of CR, MAE, FR, production estimation matches (both versions) and attraction estimation matches (both versions) were used to determine trends within the "best" models. Using the same non-weighted average of rankings, the following conclusions were made:

1.) The peer-to-peer models that used the friction functions of linear-linear, negative exponential-linear, and gamma-linear performed better than the other "best" performing models. This negative exponential-linear model was ranked first with an average score of 3.143. Second and third ranked models were the gamma-linear (3.429) and linear-linear (3.571), respectively.
2.) Despite considerably low CR values, the peer-to-peer gamma-negative exponential and the doubly-constrained gravity gamma-linear models showed strength in their ability to predict productions and attraction rates per TAZ fairly accurately. These models also ranked in fourth (4.143) and fifth place (4.571), respectively.
3.) All of the peer-to-peer models ranked higher in the average rankings than the doubly-constrained models.

The visual analysis of the trip distribution and intensity graphics did not always support these previous conclusions. With the exception of the peer-to-peer models abilities to better predict OD intrazonal trip intensities, the doubly-constrained models appeared to be more successful in their abilities to coincide with the CAMPO model on average.

CONCLUDING STATEMENTS

The analysis done within the chapter provides some practical insight into the performance of the doubly-constrained gravity and peer-to-peer models capabilities compared to the CAMPO model. The following are the resulting themes from this analysis:

1.) Based on the results of the coincidence ratio, the doubly-constrained gravity model had better results with the use of the negative exponential friction function for the long trip component of the two-regime friction function.
2.) The coincidence ratio also showed the peer-to-peer model's strength with respect to the use of the linear friction function for the long trip component of the two-regime friction function.
3.) The peer-to-peer modeling methodology showed strength in estimation of intrazonal trips.

The noted differences between the proposed models and the CAMPO model can potentially be attributed to the growth in population within Austin between the two study years. Additionally, according to the FRED Economic Data website (2015) there has been an increase of approximately 150,000 employed (non-farm) individuals between the two study periods. These factors also can be seen in changes in land use and in the on-going development of the city.

Additionally, it is also worth noting that variation existed in computation time for each of the models. The doubly-constrained gravity model took an average of seven hours per run, while the peer-to-peer had only an average of two hours per run. This computational efficiency of the peer-to-peer model could be desirable for municipalities who have greater need for efficiency.

Finally, it should be noted that CAMPO has an updated effort currently underway that will create trip distributions at the parcel level using survey data from 2010. Future efforts should look to comparing the results of this dissertations methodology against the result of the new CAMPO TDM model for further understanding of the proposed method's capabilities.

Chapter 6: Conclusion

This dissertation effort explored the use of Location-Based Social Network (LLBSN) data and peer-to-peer modeling for transportation planning. Two directions of focus were proposed in Chapter 1:

1. How impactful is LBSN as a data source? Does the data have the capabilities to accurately represent the demographics of the Austin area, and if not what are the limitations? Is the data capable of being a standalone dataset or is it better suited as a supplementary source?
2. What are the benefits (if any) from using many-to-many, specifically peer-to-peer, modeling with respect to origin-destination matrices?

The previous chapters of this dissertation attempted to provide answers to these questions through the detailed examination of the two thrusts. This chapter will provide concluding remarks to summarize the dissertation efforts, what the impacts from this effort are, and future directions the research could go.

LBSN AS A DATA SOURCE

Chapter 3 provided a detailed examination of the Foursquare dataset selected for use. Venue characteristics were examined to better understand the data's strengths and limitations. Venues were examined from the category, time of day, and day of week perspectives in an effort to determine if biases existed within the data and to determine how check-in trends related to individual travel within the study area. These efforts are a first attempt to illicit nuances that exist within the dataset.

It was noted that the data included a significant number of venues in the Shops & Services as well as Professional & Other Places categories when compared to the other categories. However, the number of check-ins for these locations was overly distributed.

Instead, the Food and Shops & Services categories were noted as having a significant number of check-ins. This over representation may need to be mitigated against should a municipality desire to use LBSN as a data source. This could be done by providing a weighting system that would place emphasis on venues that would be typical commuter venues (i.e., professional, work).

With respect to day of week and time of day, the examination of the data revealed a notable trend of check-ins on Tuesdays for weekdays for all categories of venues. For each category, the venue check-ins were examined for each day of the week to look for any unique trends. One possible reason for this is that according to a survey conducted by Accountemps Tuesdays are the most productive day of the week (Brooks 2015). For the most part, venue check-ins were found to be consistent for Monday through Sunday, with the noted increase on Tuesdays. There were three categories that did not follow this trend: the Nightlife Spots category, the Professional & Other, and Residences. The Nightlife Spots category had a significant increase on Sundays, while the Professional & Other category had a significant decrease on Sunday. For the Residence category, there was a notable increase for Sundays, which was surprising. The time analysis revealed the likelihood of check-ins occurring during peak hours throughout the weekdays. Further analysis examined the time periods each category was checked-into, and provided additional insight into trends for each category. With the exception of late night residential check-ins, most categories had peak check-ins at anticipate time periods (i.e., Professional & Other Places during the A.M. Peak). A visual analysis for location of check-ins was also conducted to observe the time and location for check-ins, as well as identification of venues with the most check-ins. These analyses provided additional insight into the datasets spatial-temporal coverage with respect to each category, and added in identifying locations with unusual or extreme check-in rates.

The Foursquare data was also examined from a user demographic stand point. This was done to identify biases within the data that may exist such as age, gender, and income. One major limitation from the existing dataset is the lack of individual user information. This is due to the methodology that was selected for data collection. However, purchasing data from a data distributor would allow for user information to be included in the data and would greatly increase the value of the data. Typically the user information does not use an individual's personal identification (i.e., name), but instead uses a userID, an alphanumeric identifier, which could be used to track the individuals check-ins throughout the study period. This in turn could provide insight into check-in rates for users of Foursquare as well as trip chaining that may be occurring. Some information was able to be extrapolated from the dataset through the joining of an external dataset and base assumptions. This included the average household size and mean income, which was fairly comparable to the comparison data source. Examination into income groups revealed a bias toward the $20,000 to $75,000 income categories, which matches the comparative demographics for Foursquare as a whole.

Finally, the Foursquare dataset was compared to the land use data for the study region to determine how the venues composition within each TAZ related. Based on this analysis, the Foursquare data fell short in the majority of TAZs within the study area from the four categories examined. Additionally, it was noted that the Foursquare dataset does not have a category for mining, which was one of the land use categories identified.

In addition to the above noted limitation, the current data source is limited by the platform change from one application to two applications that operate in particularly different capacities, as described in Chapter 3. This change may require an individual to access both platforms data in order to get the level and type of data used within this dissertation. Another limitation of note is the numbers and locations of check-ins are

heavily dependent on user's willingness to do so. The same limitation exists with respect to the venues within the data source; users must chose to create them. Furthermore, the tendency of users to check-in to leisure and recreational type venues more than work related venues provides a skew in the dataset. However, this trend in the data does provided insights into the activities users participate in more so than other traditional data sources. Finally, with respect to land use, the limited overall representation of land uses could be problematic for examination at the parcel level, which some municipalities are moving toward.

PEER-TO-PEER MODELING

This dissertation effort provides a first attempt at using many-to-many modeling for transportation planning. Chapter 4 provided the methodology used and Chapter 5 offered a case study using Austin, TX as the study area. A comparison was done between the doubly-constrained gravity model and the peer-to-peer model to assess the strengths and weaknesses of each model. Nine different friction functions were used for each model and the "best" performers from each base model were compared to each other based on multiple criteria. It was found that the doubly-constrained gravity model performed "best" with the friction functions that had the negative exponential formula for the long trip component of the two-regime friction function. For the peer-to-peer model, the "best" performers were the models with the linear friction function for the long trips. In a side by side comparison of the "best" model revealed that the peer-to-peer models had higher coincidence ratio values than the doubly-constrained models.

An examination of the production and attraction rates produced by each model was also conducted and compared to the Capital Area Metropolitan Area's model. While the results of the production and attraction rate examination is dependent on the models,

it also revealed areas where there were a greater number of check-ins (e.g., the airport TAZ), which directly impacts rates of productions and attractions. An examination of the origin-destination matrices from each model also revealed the peer-to-peer models capability to predict the intrazonal trips better than the doubly-constrained model.

As was noted in the chapter discussion, difference between the proposed models and the comparison model may be from how the models are calculated, but could also be from the changes in the Austin population between the study dates. Austin's added population resulted in increases in employment as well as changes in land use and increased development throughout the study region. Moreover, the comparison of a venue based analysis to a TAZ analysis could suffer from the modifiable area unit problem, which encompasses the aggregation and zoning issues that result in validity errors.

When examining the computational efficiency of each model, it was found that the peer-to-peer methodology was significantly more efficient than the doubly-constrained model. This is particularly important for municipalities who may not have a dedicated machine for running analysis. Additionally, the ability to update the model quickly with new data may be desired by municipalities and the speed of the peer-to-peer modeling would be appropriate.

FUTURE RESEARCH

Based on the results of the examination of the dataset and the peer-to-peer modeling capabilities, there are areas that should be investigated further to additionally validate the use of LBSN data and the many-to-many modeling. With respect to the LBSN data, analysis to examine in greater depths the location venues to one another could be done. An algorithm could be developed to determine the likelihood of travel

between venues based on distance and time. Having the capability to do this may user statistics could also be explored with respect to the dataset. This could include possible inferences on user income based on restaurant venues, if additional information from other LBSN sources like Yelp were added to the dataset. Similarly, vehicle ownership could be inferred from checking-into parking facilities. The current analysis of the Foursquare data identified user demographics that were under represented, which could serve as a catalyst for targeted survey that would address these biases. Having to only use more traditional survey methods on smaller groups would lessen the financial burden on municipalities.

The fields of dynamic traffic assignment (DTA) and activity-based modeling should be further explored with respect to this data source type. DTA requires data that includes a time component, which is readily available within the dataset. Similarly, activity-based modeling requires data that contains time-based trips with trip purposes, both of which can be gleaned from the Foursquare data. The ability of the dataset to provide insights into special events and their impacts on transportation patterns could also be examined with respect to both fields. The insights potentially attained could aid a municipality's ability to keep the transportation network flowing effectively and safely.

Finally, the data set could be explored with respect to other supplemental data sets. The combination of Waze, a community-based traffic app that provides route details that consider existing traffic conditions, and Foursquare could provide additional data on route choices made by users, even revealing trends in routes that are have limited use or are avoided due to frequent traffic or incidents. The CityBikes app, which contains bike sharing data for select cities, could be used to with the Foursquare data to further understand the bike mode usage within a municipality. Ridescout, an app that provides information about transportation options between two locations, could be used to

determine modal options between venues within a municipality. Other non-transportation related applications could be explored for their additive impacts. For example, Groupon data could be mined for times and locations of usage. Along these lines, OpenTable provides users a platform for making reservations at restaurants and could also be used to determine future locations that individuals would be attending with the added benefit of knowing how many individuals would be doing so. Google Places would be another robust dataset that could be complementary to the Foursquare dataset since some venues could be in one dataset and not the other. However, data user agreements may limit this interaction since these are competitive apps. Apps like Twitter, Instagram, and Tumblr could be explored for additional value that could be added to the Foursquare data with respect to the content of each tweet and photo.

With respect to the peer-to-peer modeling, CAMPO is currently transitioning to a parcel level analysis, for which the proposed model may have better matching characteristics based on how venues are handled within the model. Along this spectrum, other many-to-many models could be explored for their capabilities including the business-to-customer and social forces models. The peer-to-peer model could also be further refined to account for venue check-in trends. Users who check-into restaurants for lunch are not likely to check-in to another restaurant until dinner. The current model treats all venues within a TAZ equally, but assigning weights based on the likelihood of check-in to a particular category may address this.

Due to the capability of easy data collection, dynamic origin-destination models could be another area for exploration. Determining the proposed model's transferability and demographical differences of LBSN data collected from other municipalities of varying sizes could also be explored in future work. Recent work by Ziemke, Nagel, and Bhat explored model transferability with respect to activity-based modeling finding it

feasible to transfer a Dallas-Ft. Worth model to Berlin, Germany. Beyond the examination of transferability of the model to municipalities of various sizes, investigation on the modeling capabilities with respect to suburban and rural areas could be done. These areas may have considerable differences in traffic patterns and planning needs that both the data source and methodology could provide greater insights than current methods. Finally, there is movement within the industry toward activity-based modeling. The Foursquare dataset has the potential to provide an exceptionally rich dataset that could easily be used for activity-based planning, especially if purchased data is used, and imposes very little burden on the participating individuals.

CONCLUDING STATEMENTS

This dissertation provides novel insights about the Foursquare data collected for the Austin area. The data source has been shown to be robust, easy to attain, and is capable of providing enormously detailed spatial and temporal information for a given area. Based on the findings of this dissertation, the use of location-based social networking data for transportation planning is recommended to be as a supplemental data source to traditional methods or in conjunction with other social networking platforms.

In addition to the exploration of the dataset, this dissertation examined using the peer-to-peer modeling methodology from the many-to-many modeling structure for transportation planning. This original effort demonstrated the ability of peer-to-peer modeling to closely approximate an existing gravity-based model used by the local metropolitan planning organization. Peer-to-peer models were should be have better capabilities in predicting intra-zonal trips, which was found to be a limitation of the comparable doubly-constrained gravity model. Additionally, peer-to-peer models are recommended when friction functions include a linear component for long trips, as they

were found to be superior to the doubly-constrained model. Furthermore, when time is a limiting factor, peer-to-peer models are computationally more efficient that the doubly-constrained gravity models. Finally, with respect to the productions and attraction rates for the examined models, the peer-to-peer (with linear functioned long trips) and doubly constrained (with gamma functioned long trips) models provide more reasonable information based on where the current population and businesses are located in comparison to the 2005 CAMPO models. This is likely due to the dataset used within this examination being more current and because of the nature of check-ins occurring at popular locations.

Appendix A

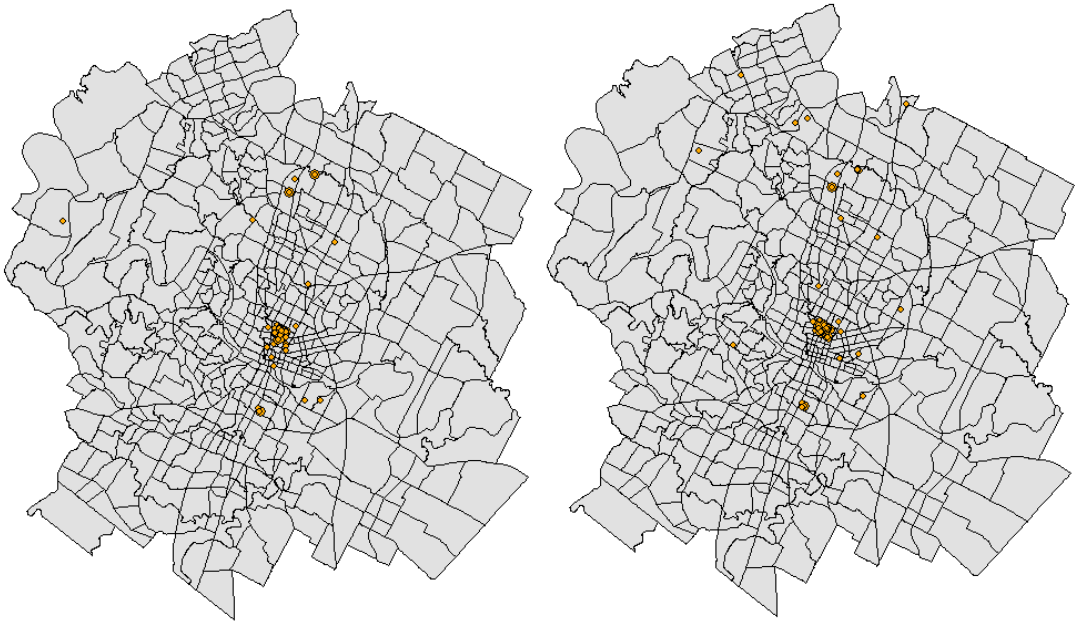
Supporting Graphics

Supplementary graphic created within the various analysis efforts of this dissertation have been included within this appendix to provide further details of the data source exploration and on each of the examined models.

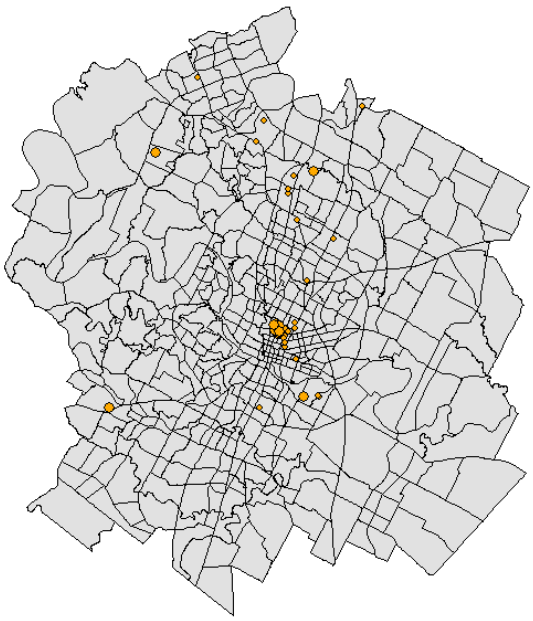
FOURSQUARE VENUE CHARACTERISTICS

Colleges & Universities:

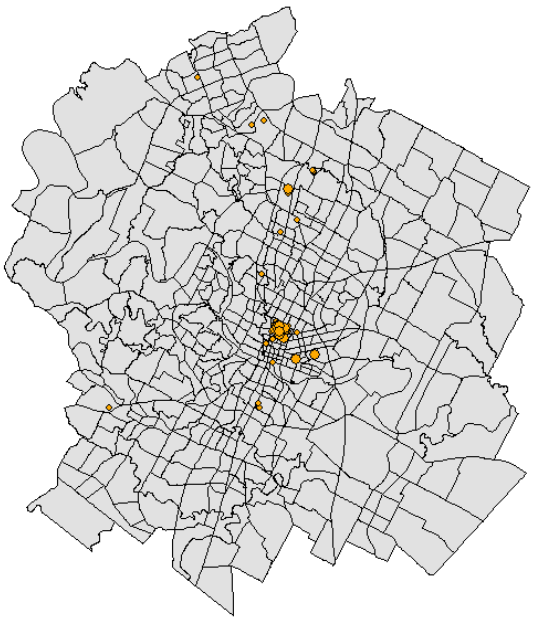
Mid-Morning



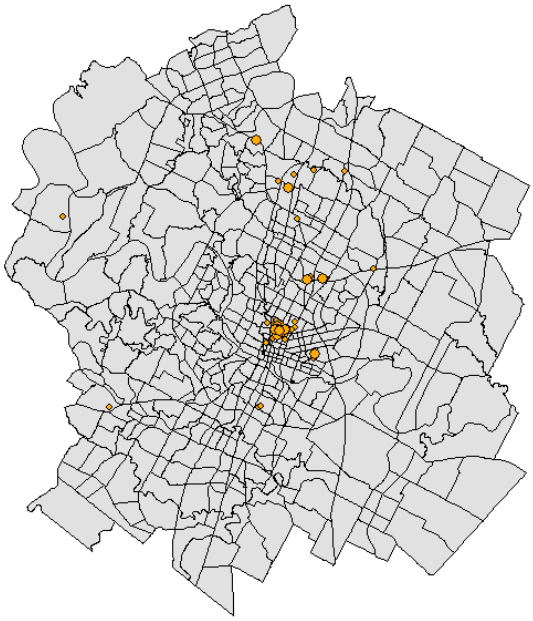
Monday Tuesday



Wednesday

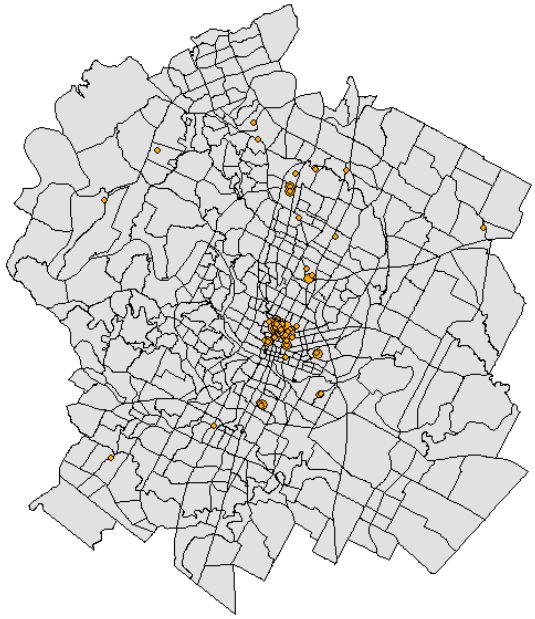


Thursday

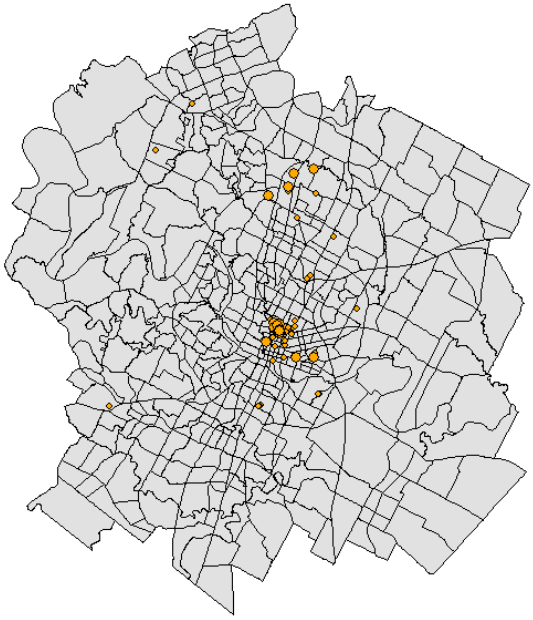


Friday

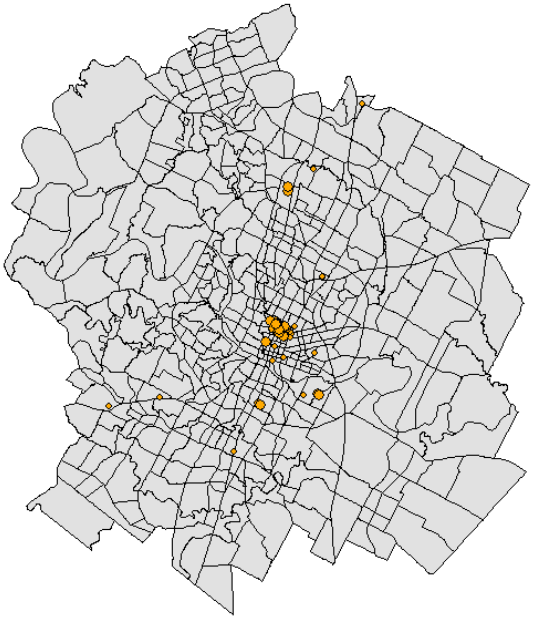
Lunch Hour



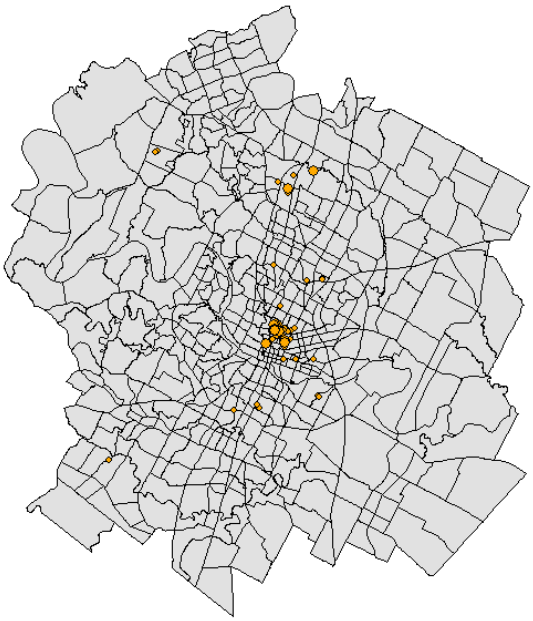
Monday



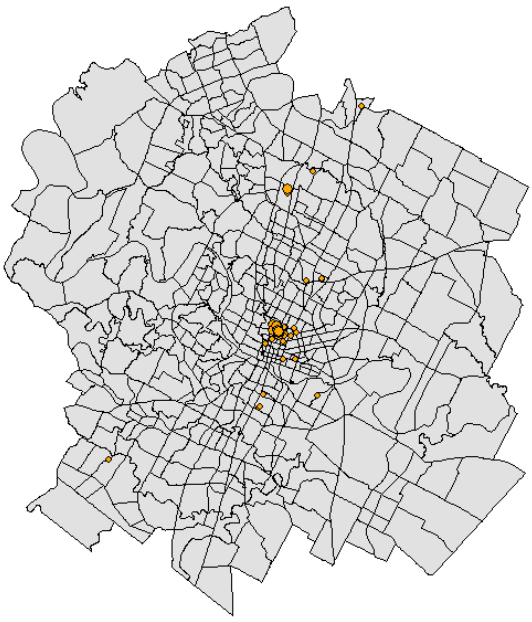
Tuesday



Wednesday

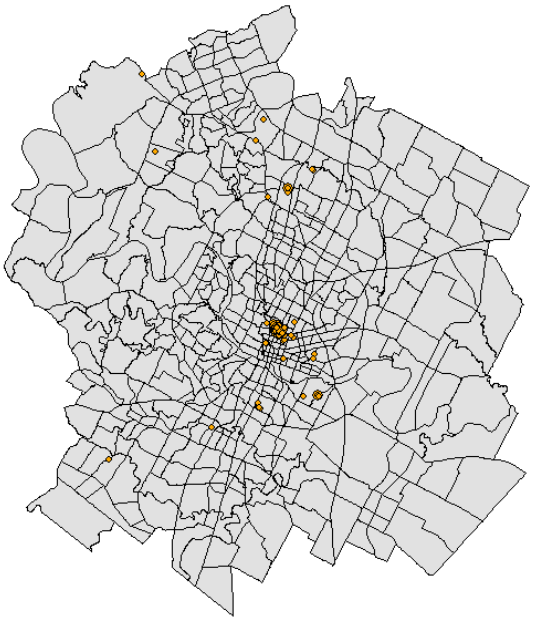


Thursday

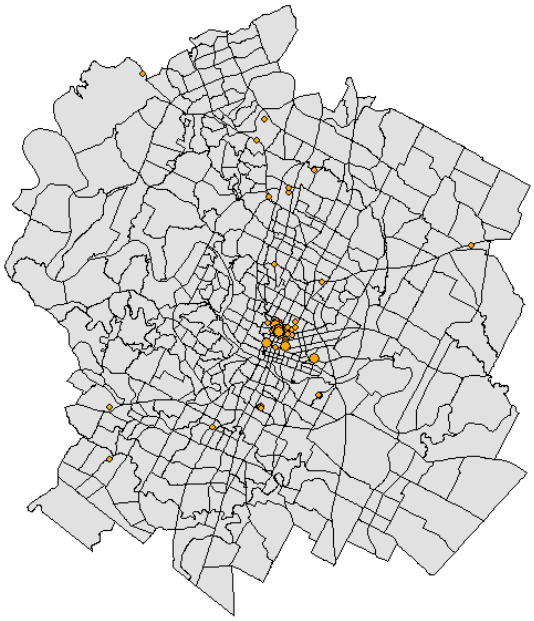


Friday

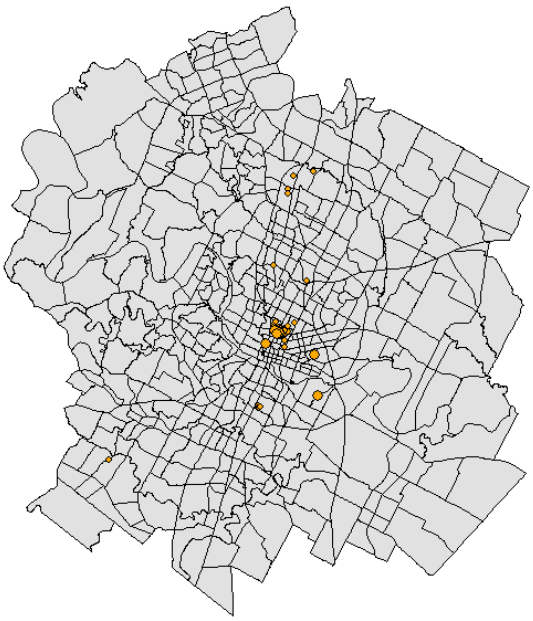
Mid-Afternoon



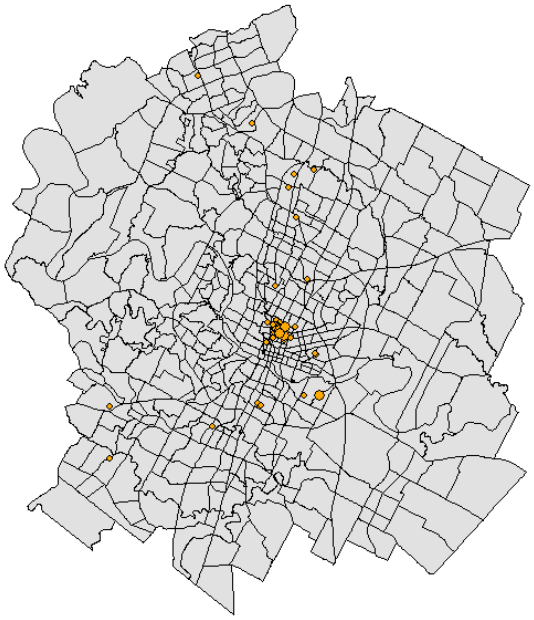
Monday



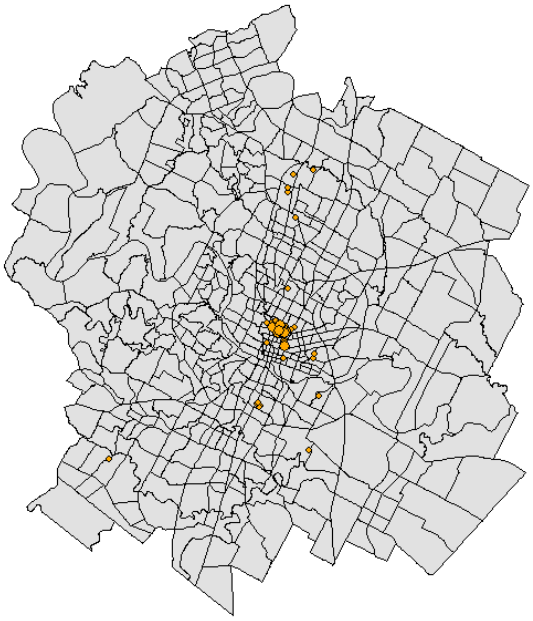
Tuesday



Wednesday



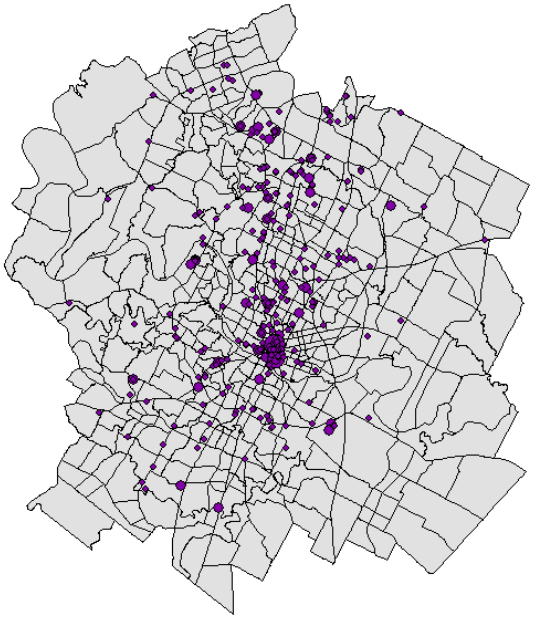
Thursday



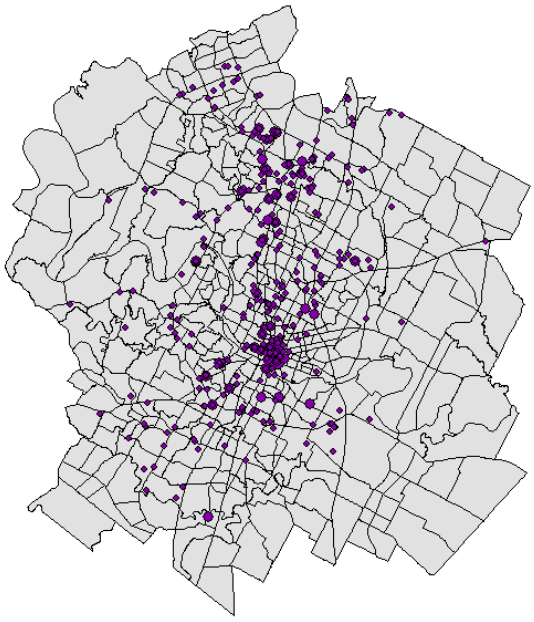
Friday

Professional & Other Places:

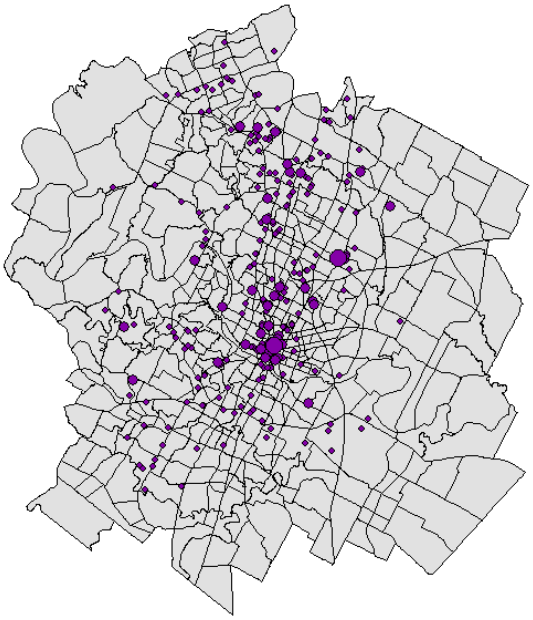
Mid-Morning



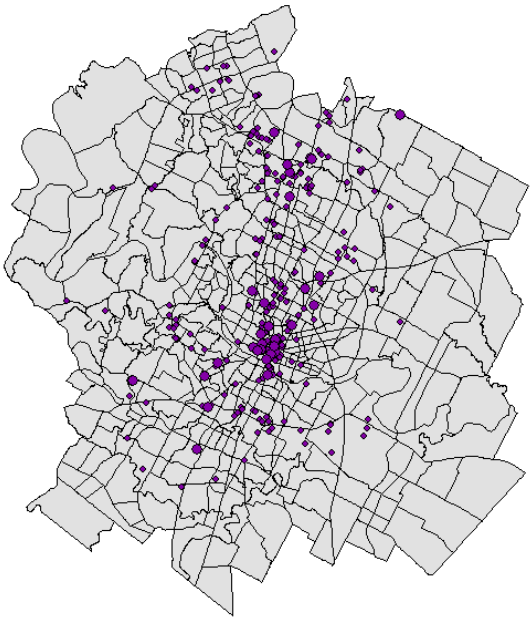
Monday



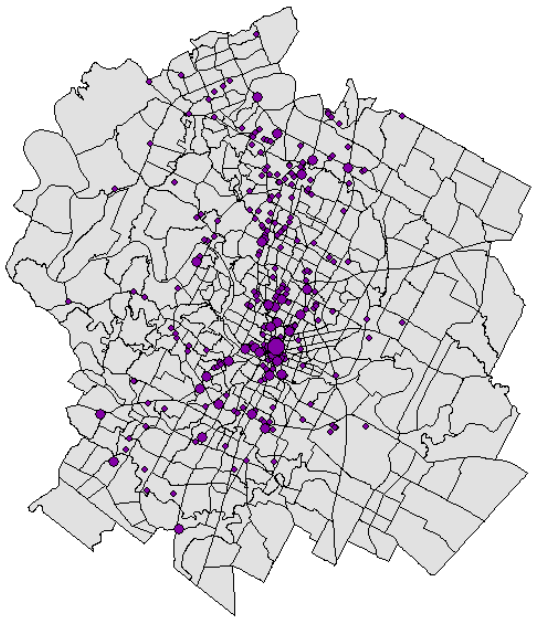
Tuesday



Wednesday

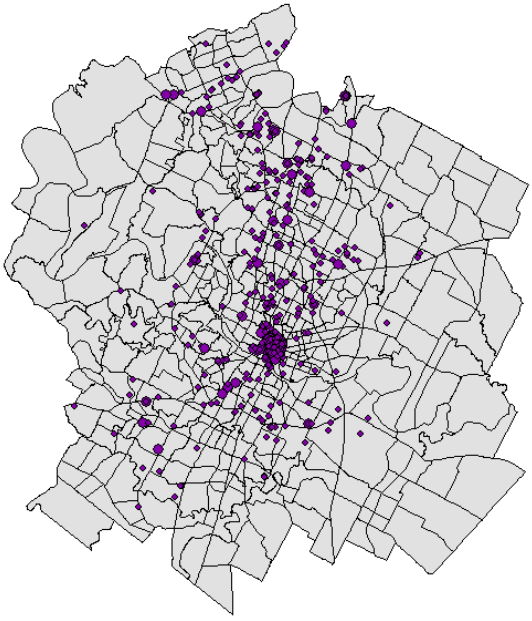


Thursday

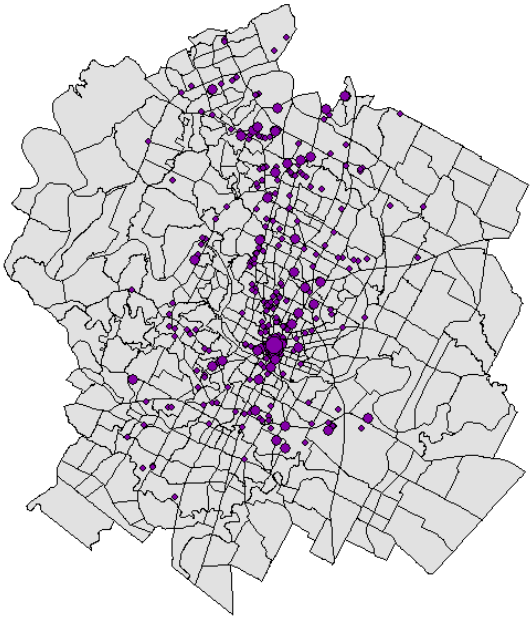


Friday

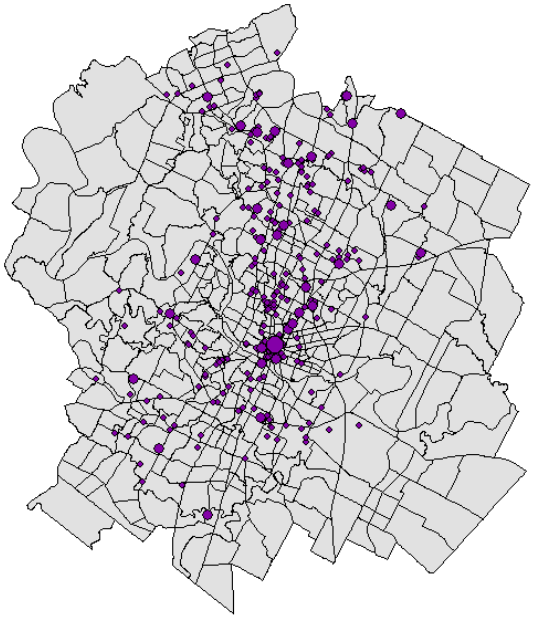
Lunch Hour



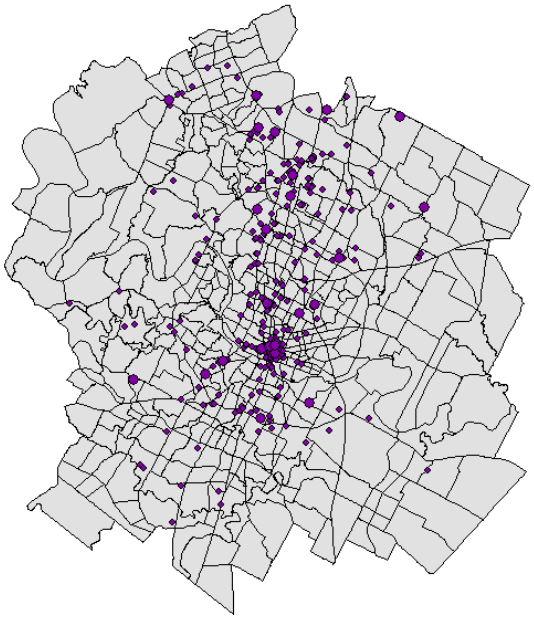
Monday



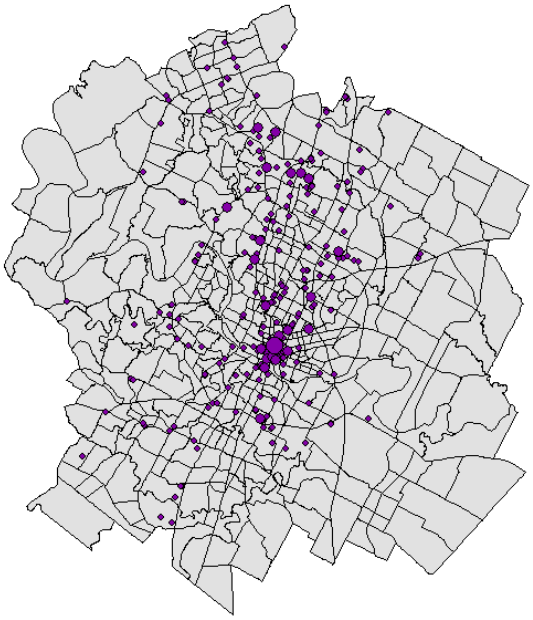
Tuesday



Wednesday

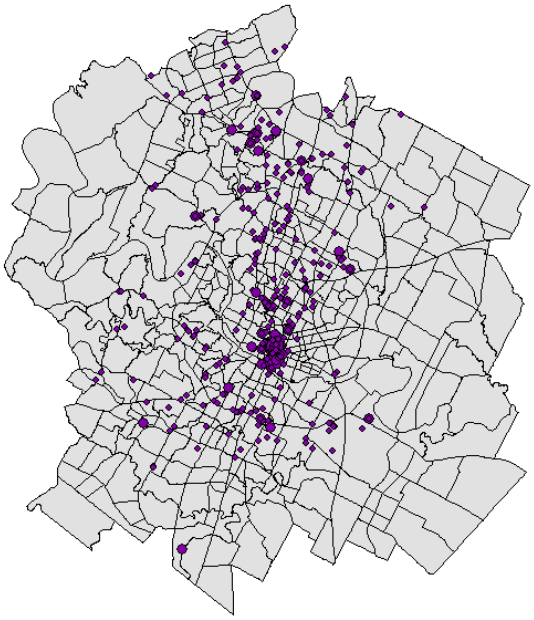


Thursday

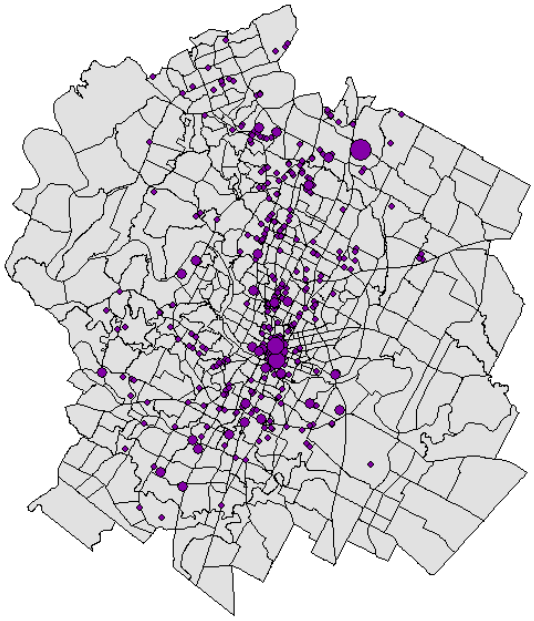


Friday

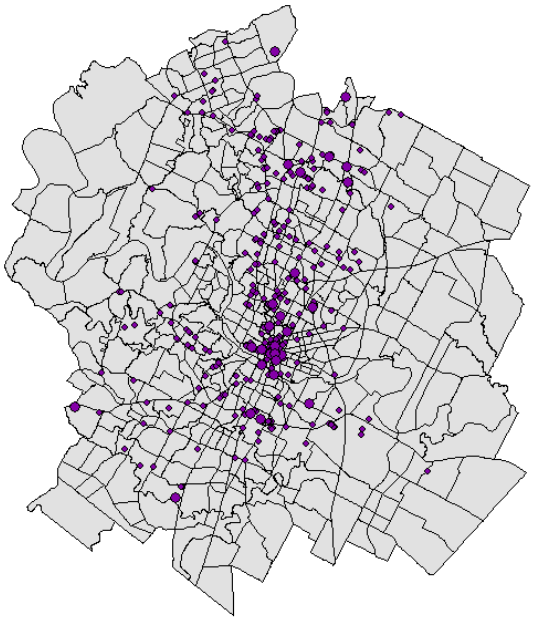
Mid-Afternoon



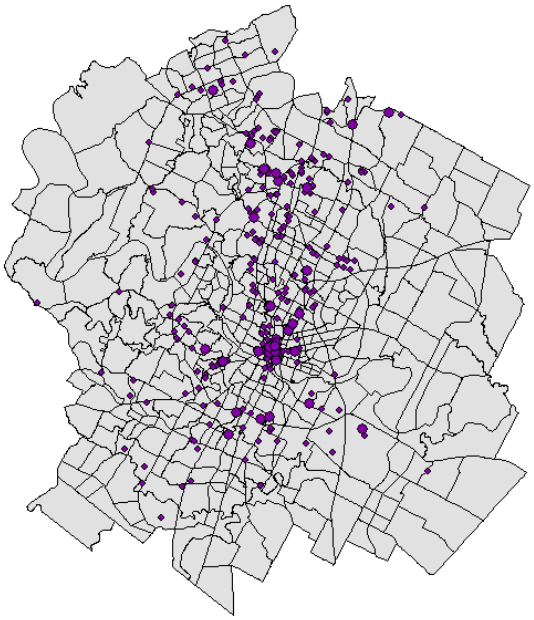
Monday



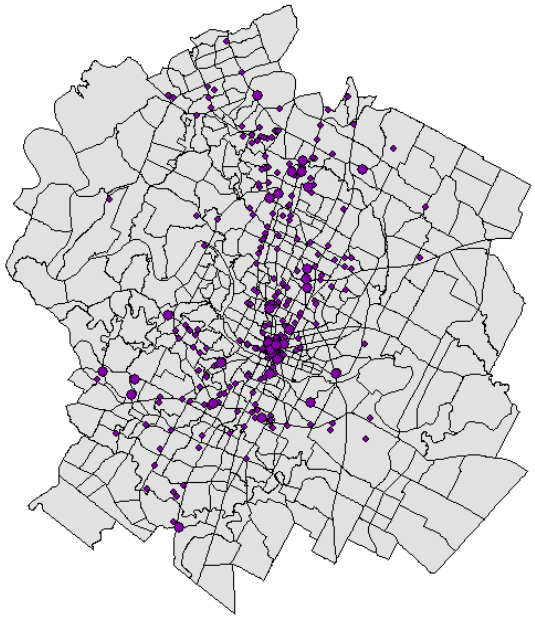
Tuesday



Wednesday



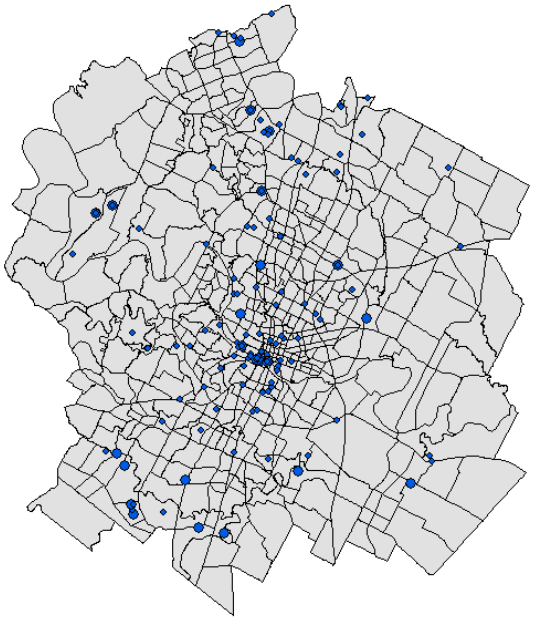
Thursday



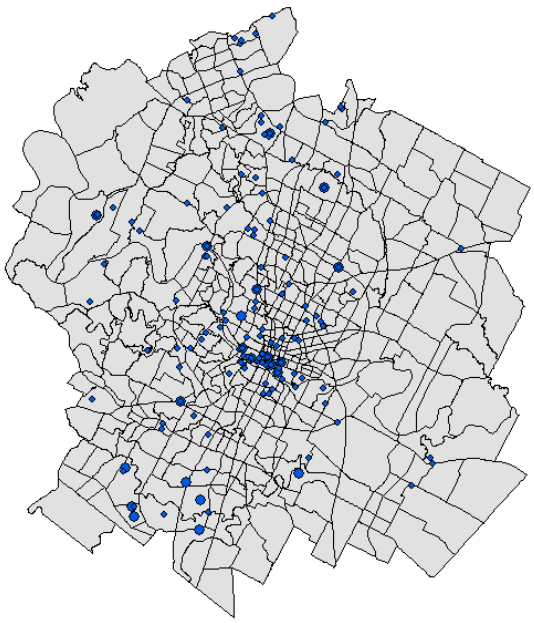
Friday

Great Outdoors:

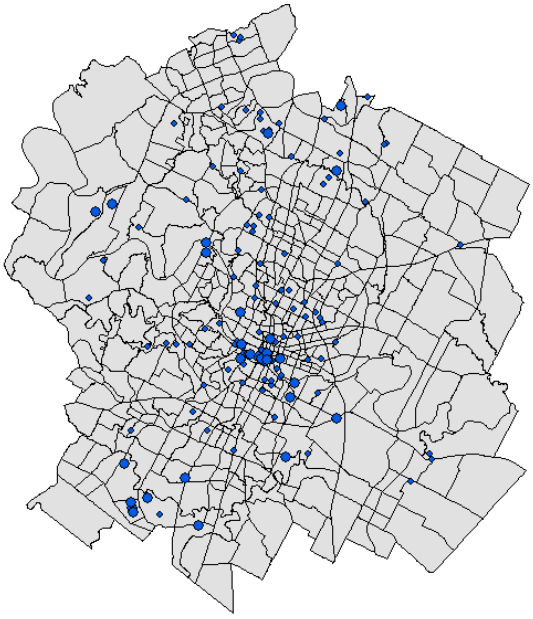
AM Peak



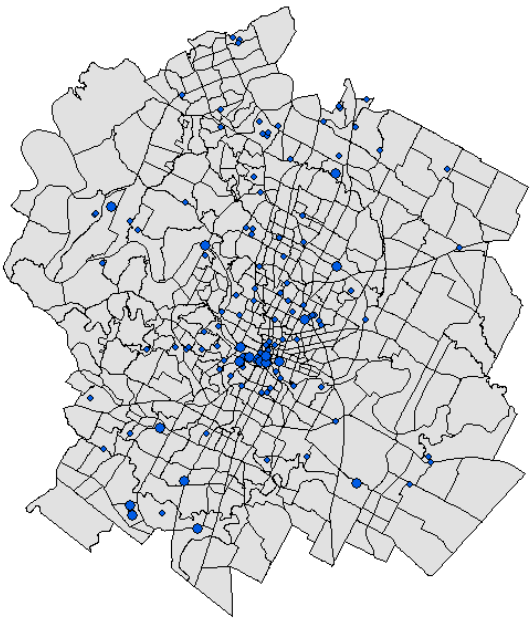
Monday



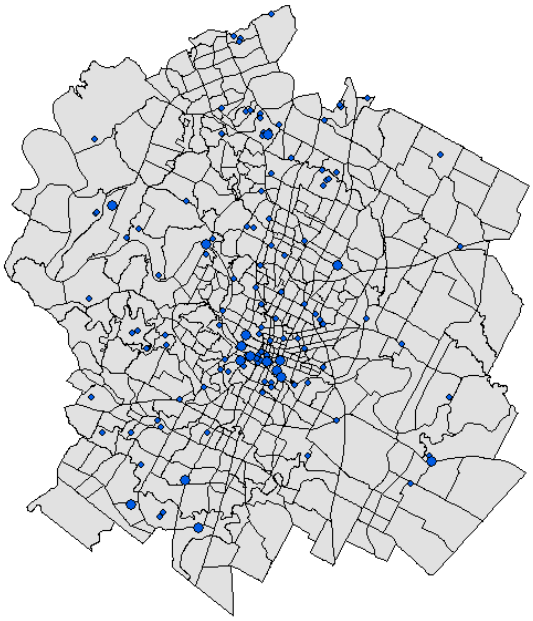
Tuesday



Wednesday

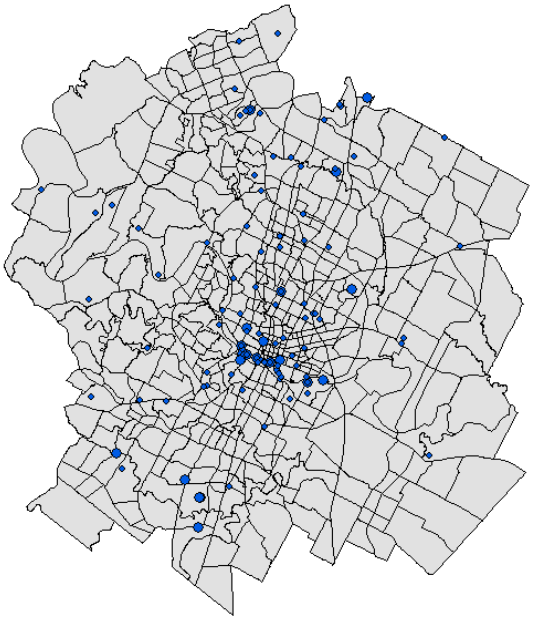


Thursday

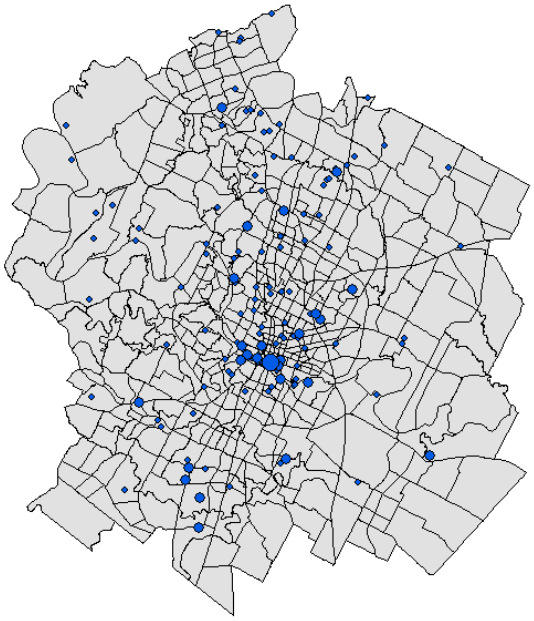


Friday

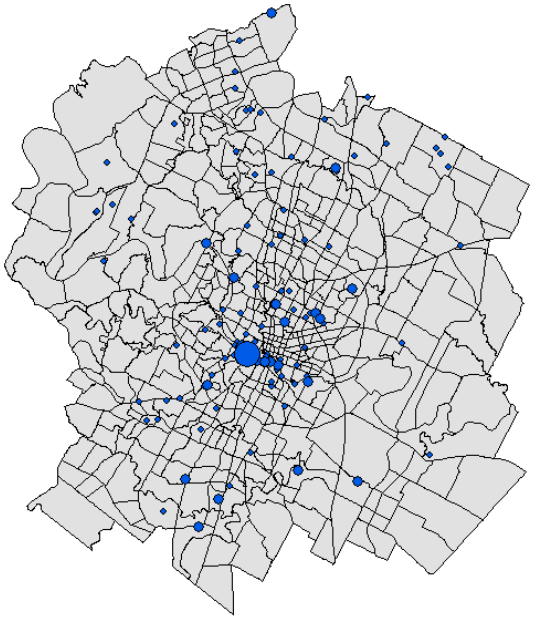
Evening



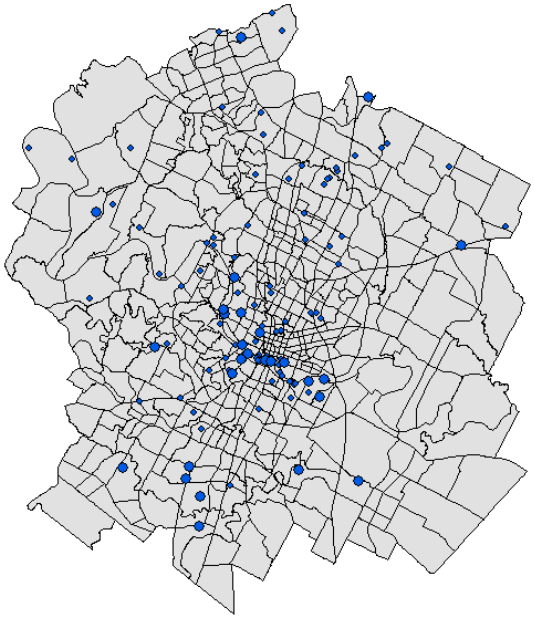
Monday



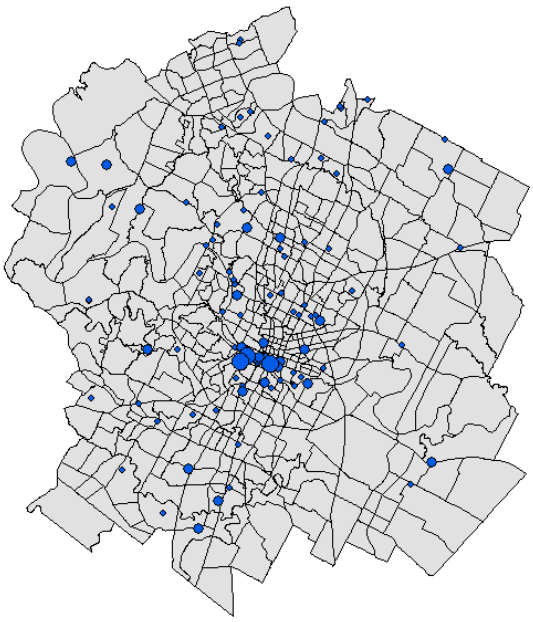
Tuesday



Wednesday



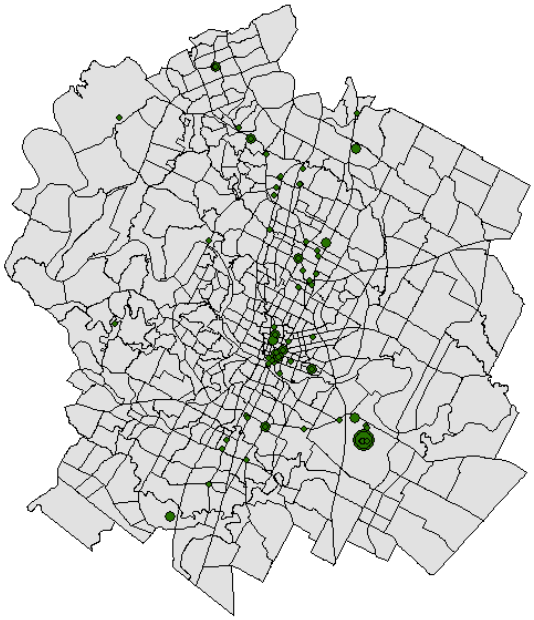
Thursday



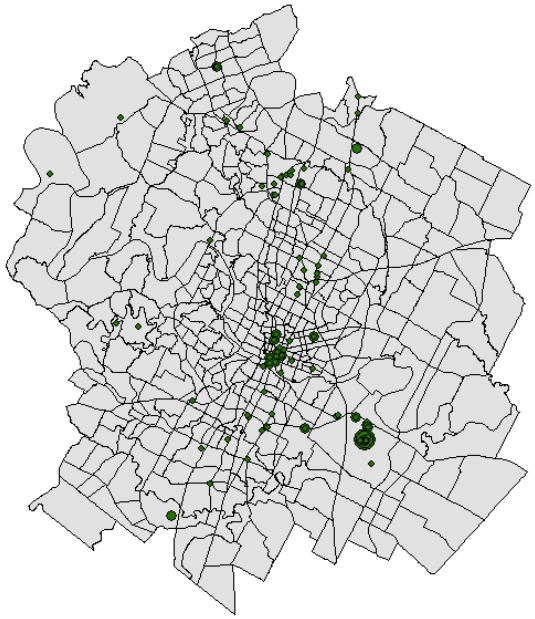
Friday

Travel & Transport:

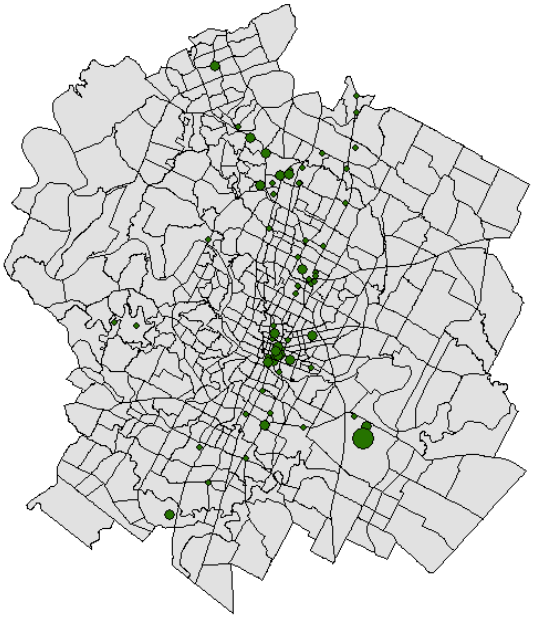
AM Peak



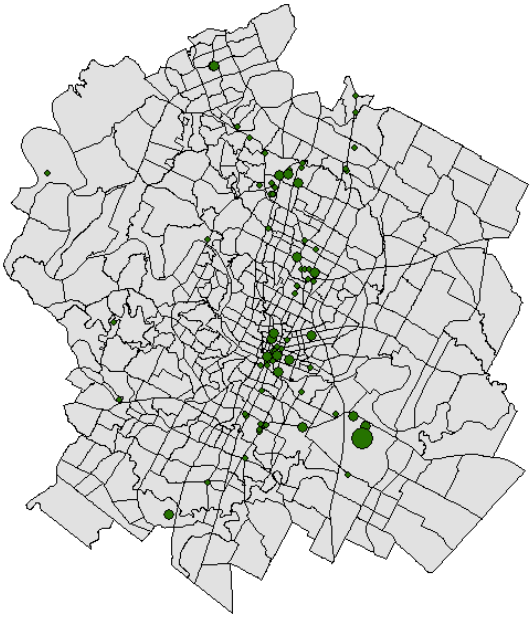
Monday



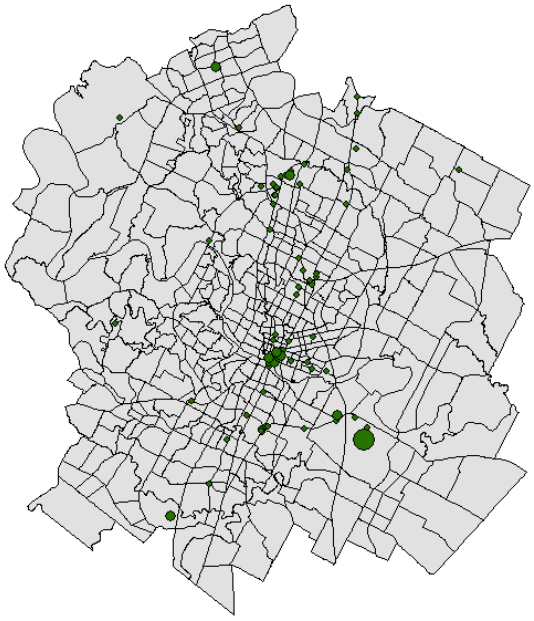
Tuesday



Wednesday



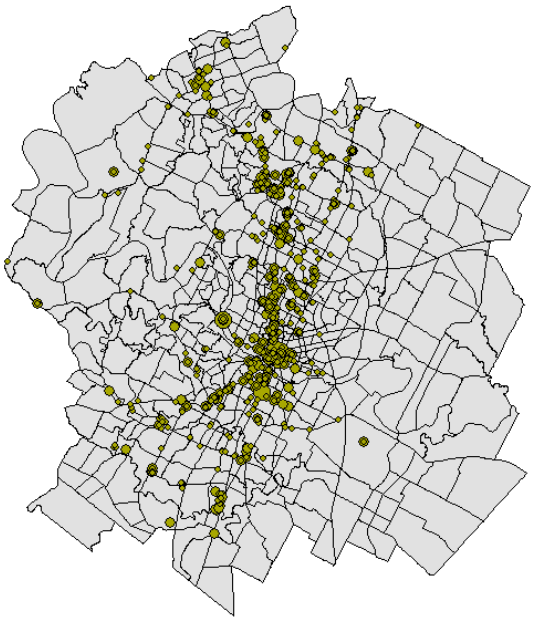
Thursday



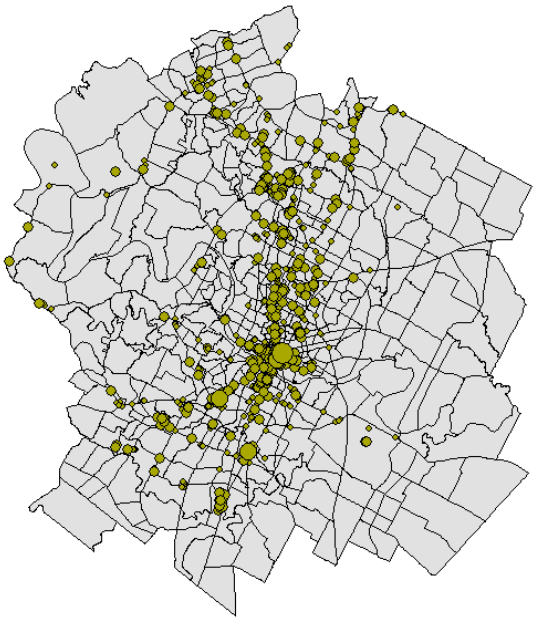
Friday

Food: Mid-Morning,

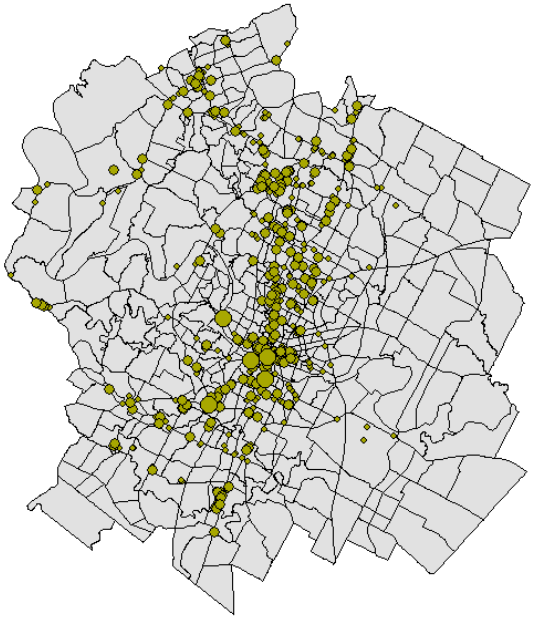
PM Peak



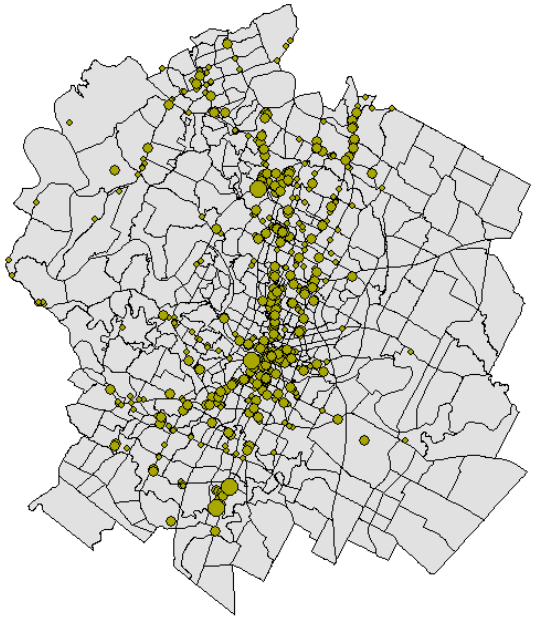
Monday



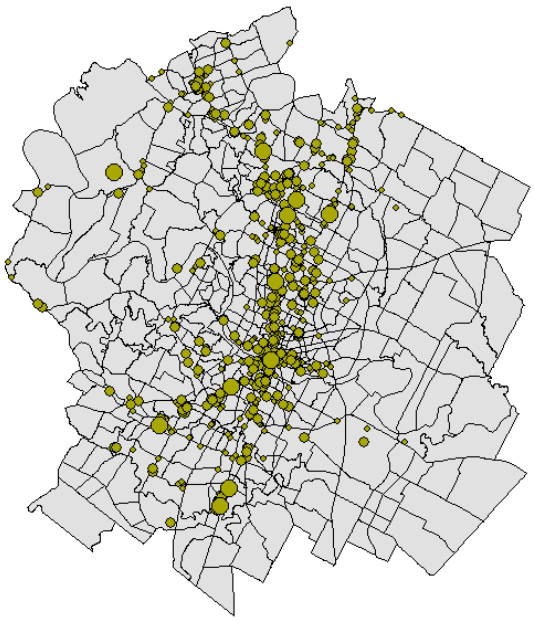
Tuesday



Wednesday

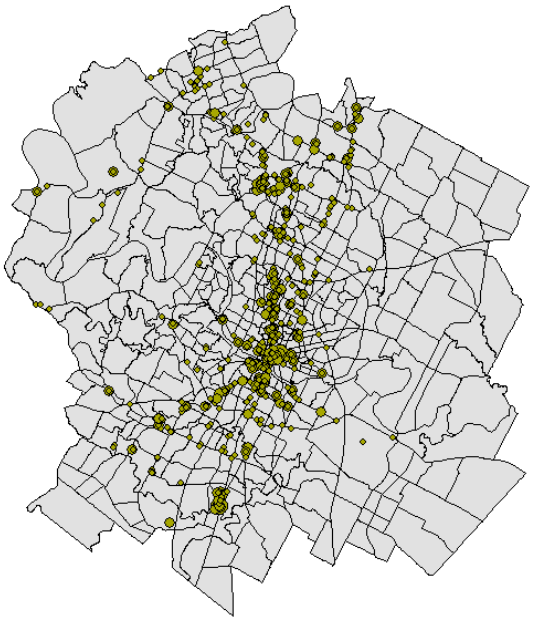


Thursday

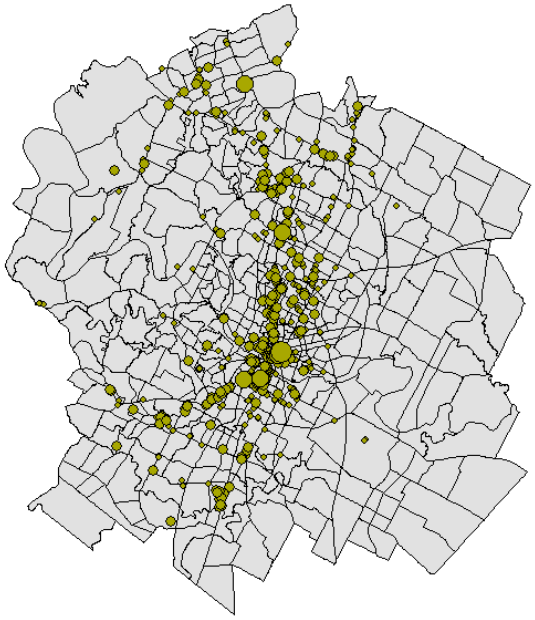


Friday

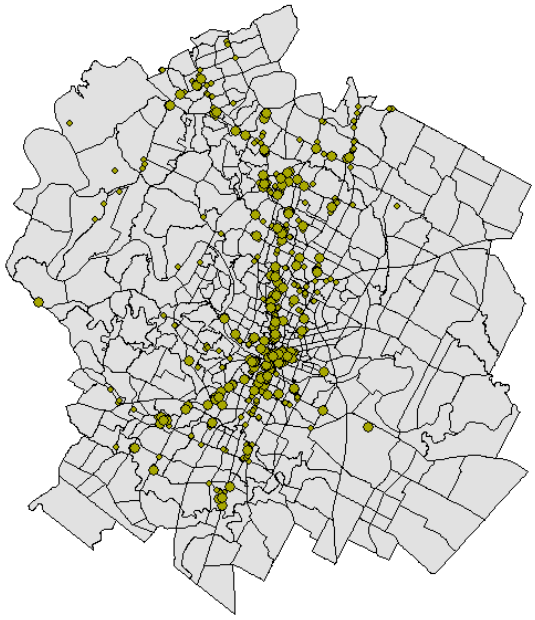
Evening



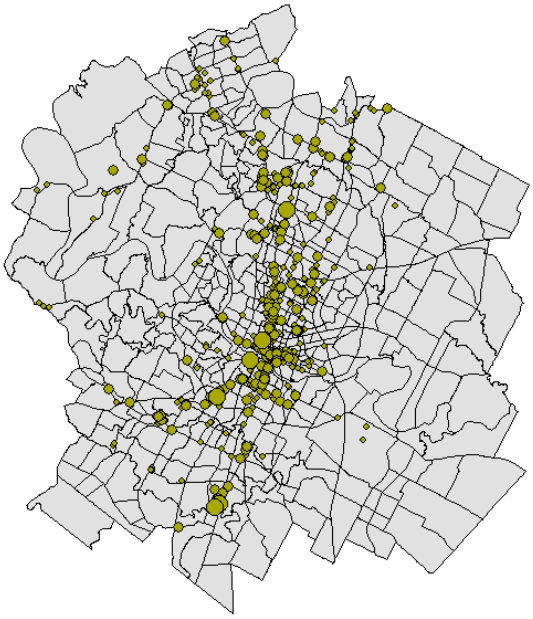
Monday



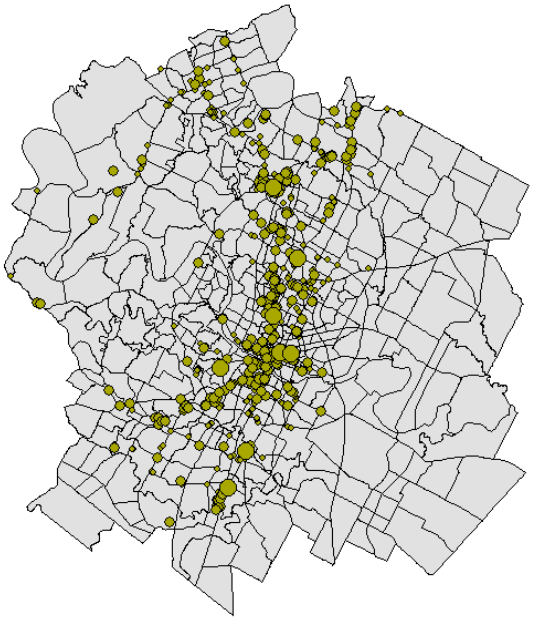
Tuesday



Wednesday



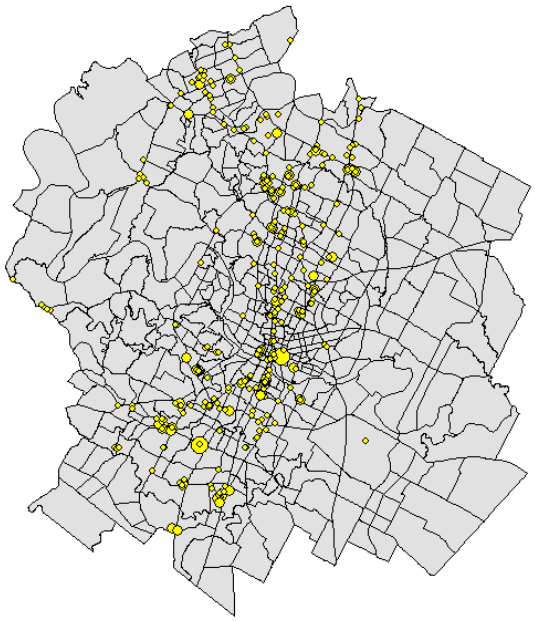
Thursday



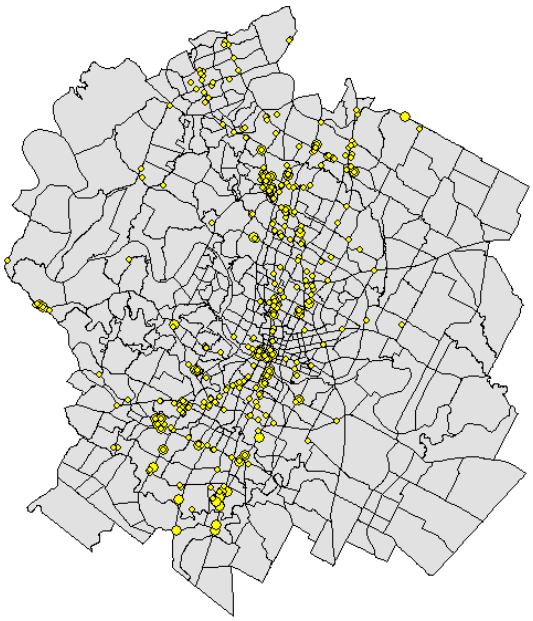
Friday

Shops & Services:

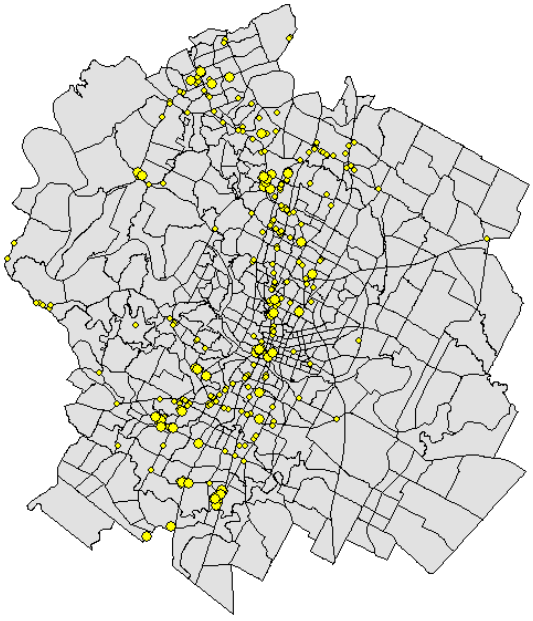
Mid-Morning



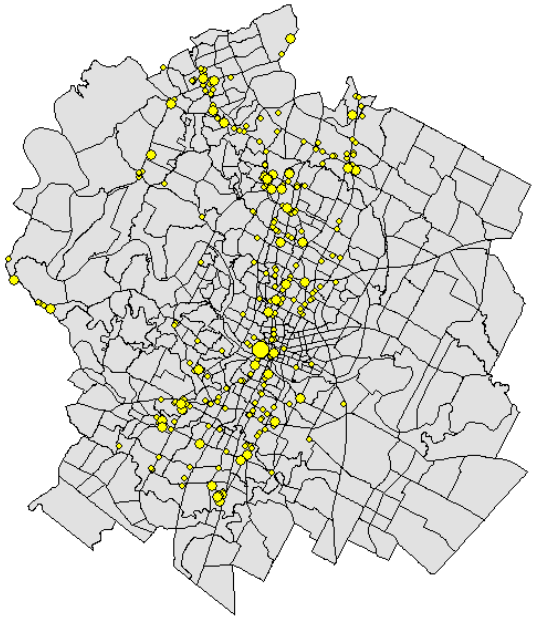
Monday



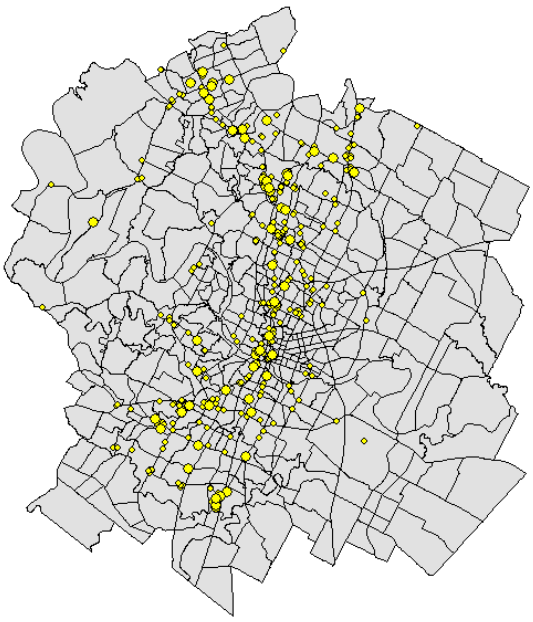
Tuesday



Wednesday

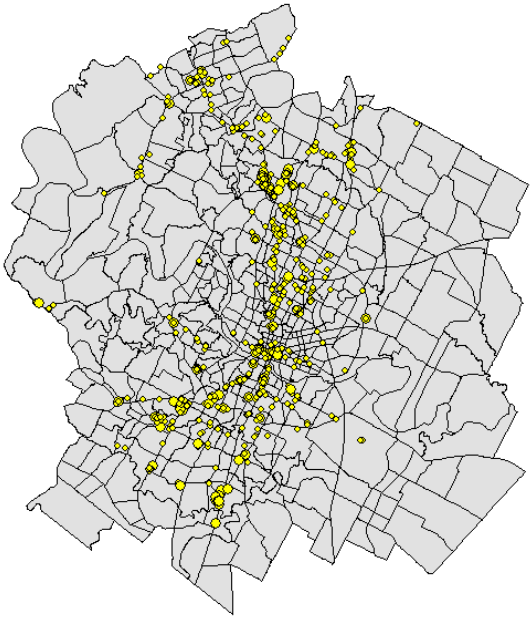


Thursday

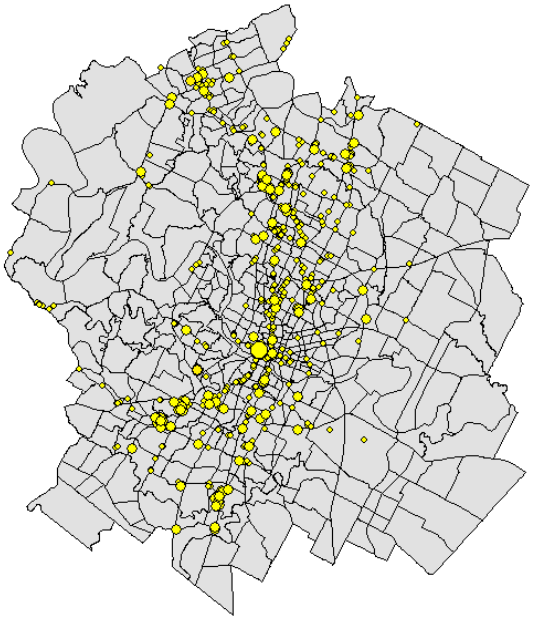


Friday

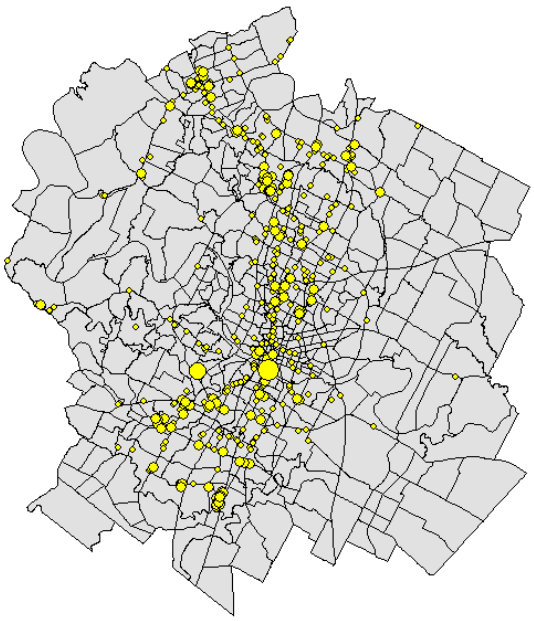
Mid-Afternoon



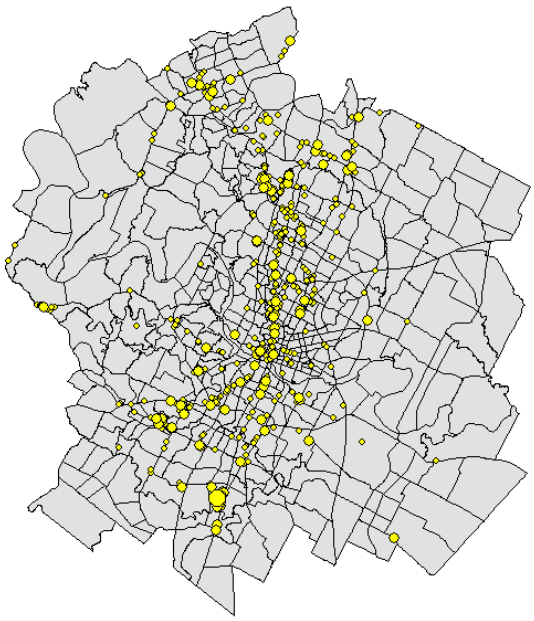
Monday



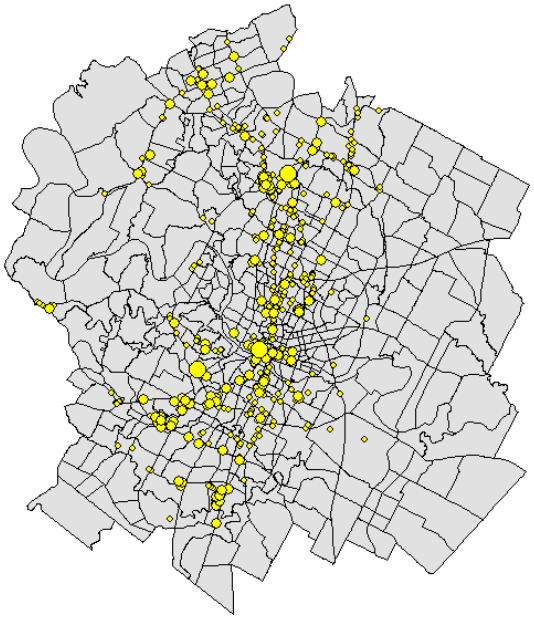
Tuesday



Wednesday

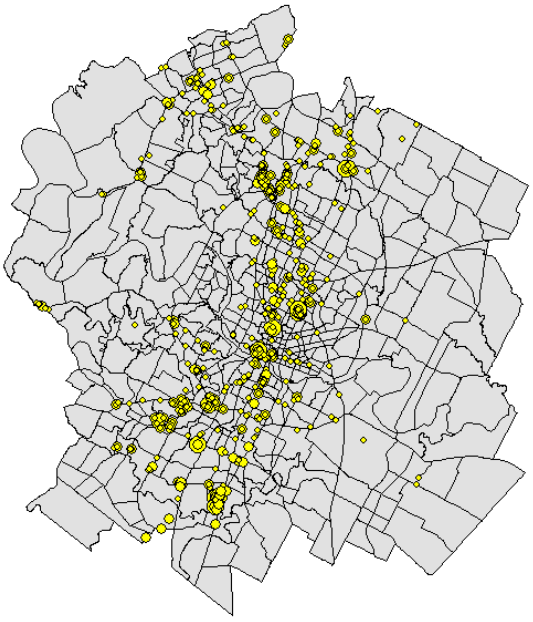


Thursday

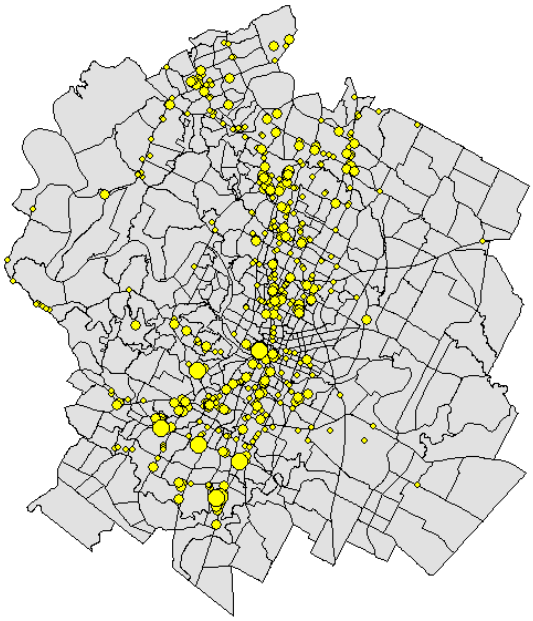


Friday

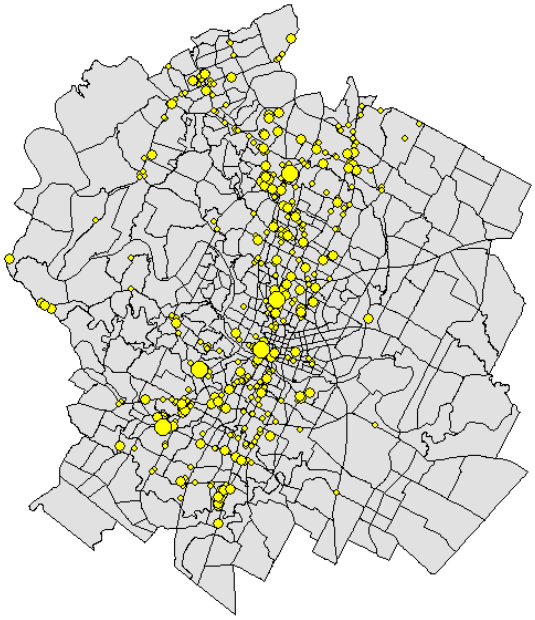
PM Peak



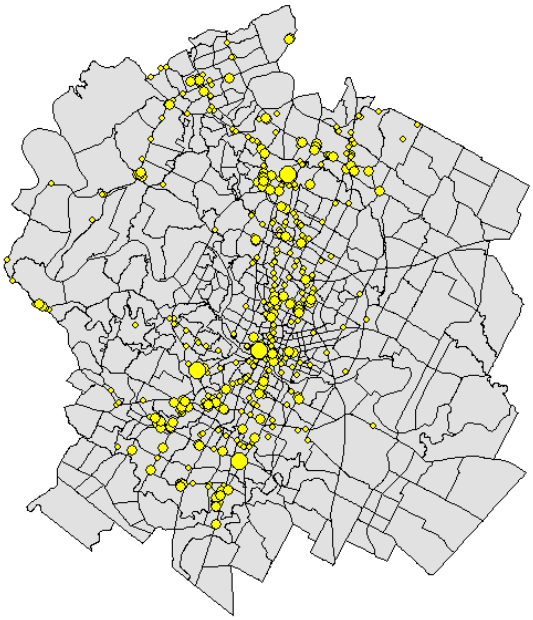
Monday



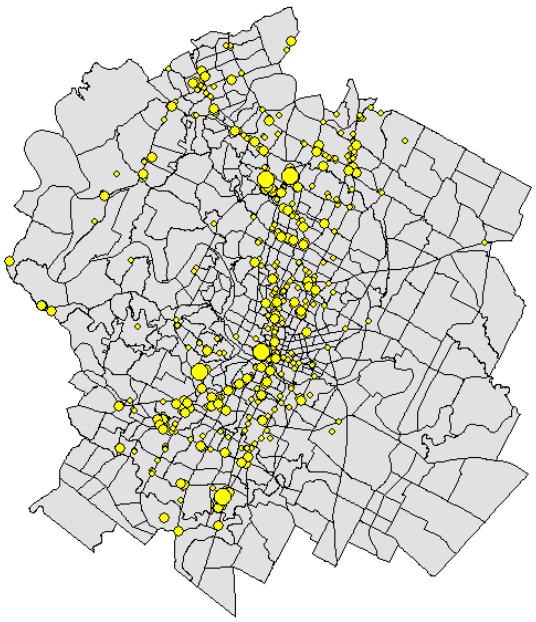
Tuesday



Wednesday

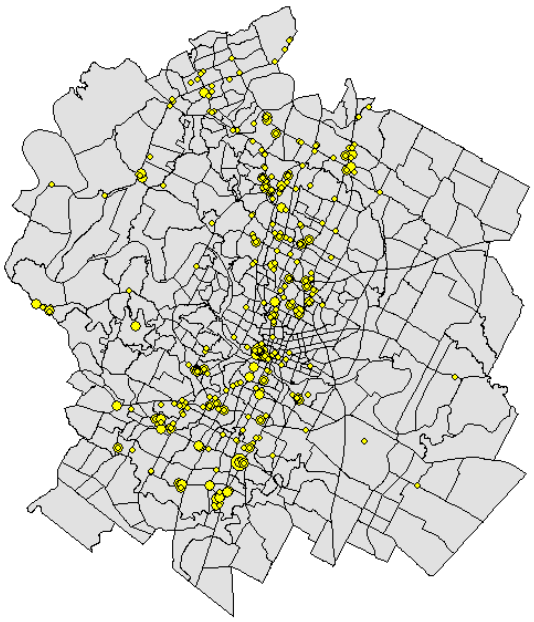


Thursday

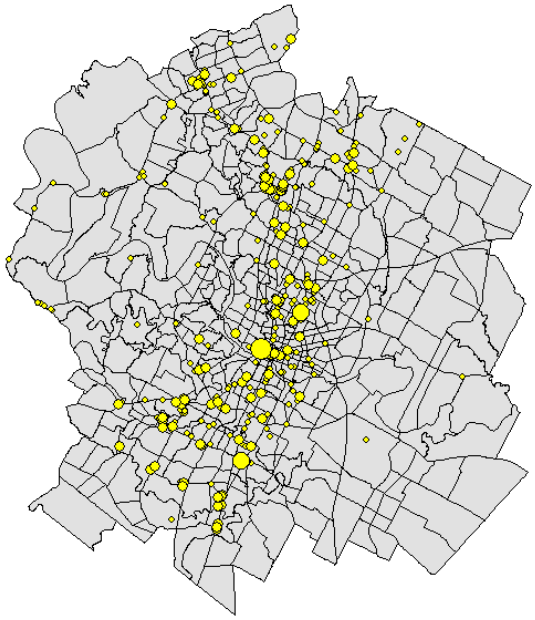


Friday

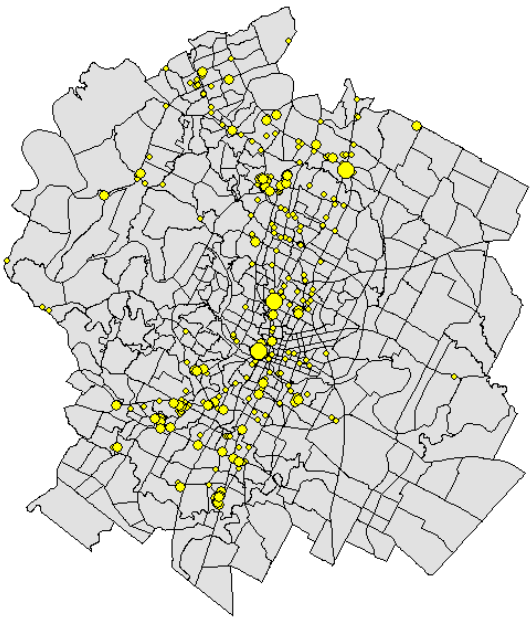
Evening



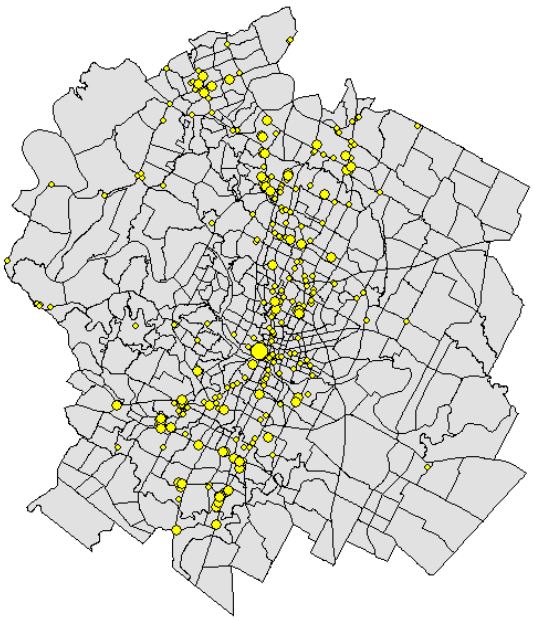
Monday



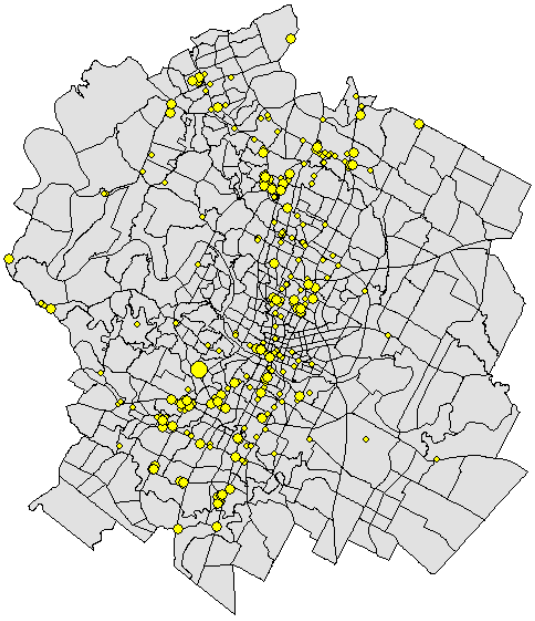
Tuesday



Wednesday



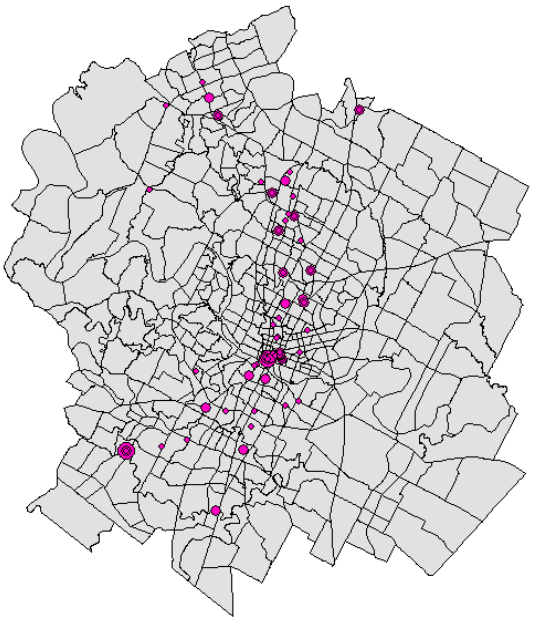
Thursday



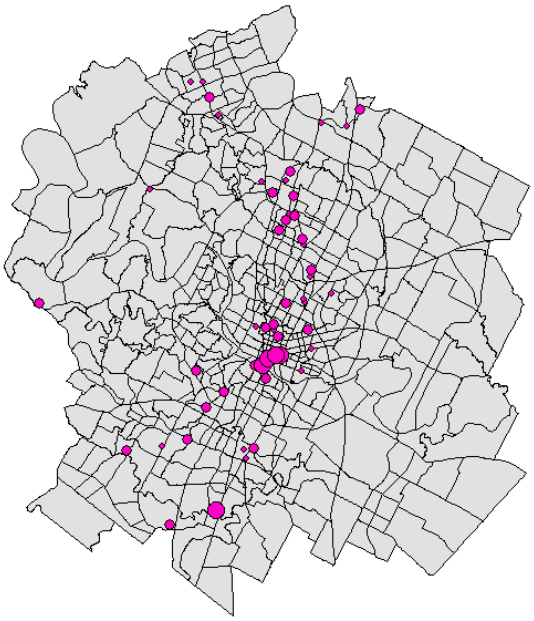
Friday

Art & Entertainment:

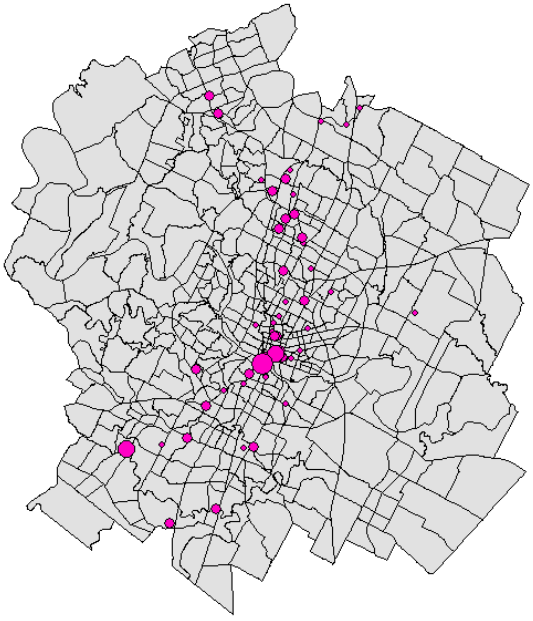
Evening



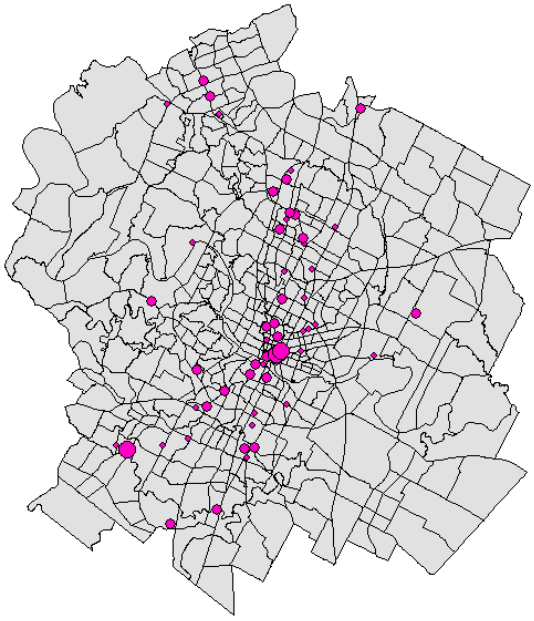
Monday



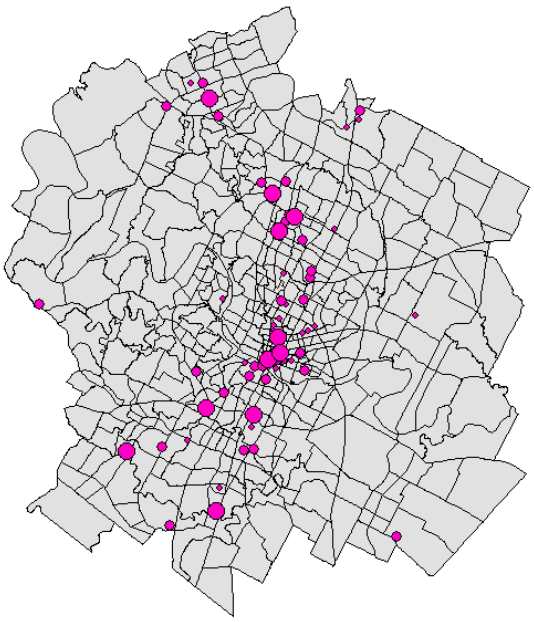
Tuesday



Wednesday



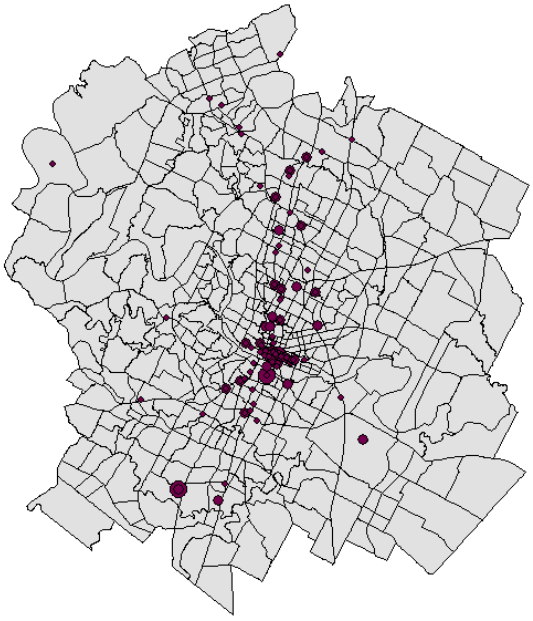
Thursday



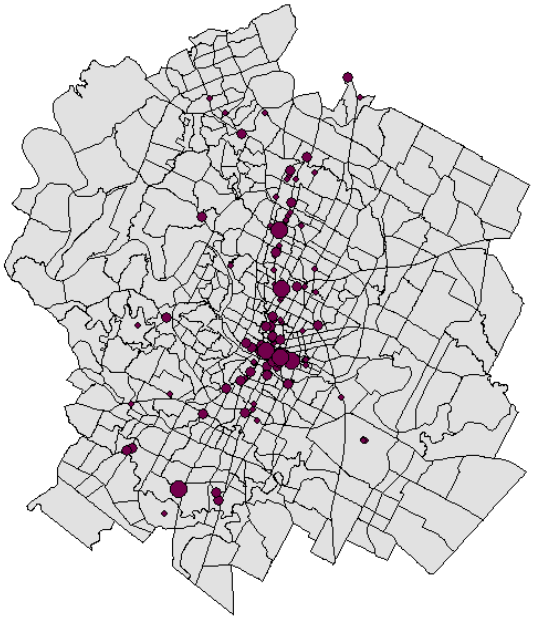
Friday

Nightlife Spots:

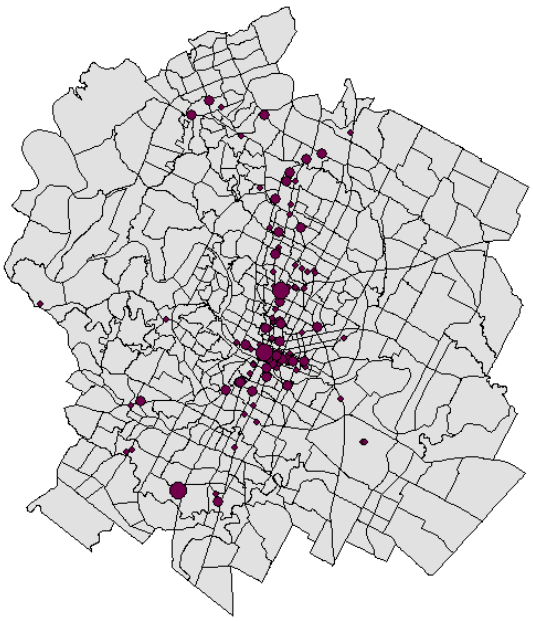
PM Peak



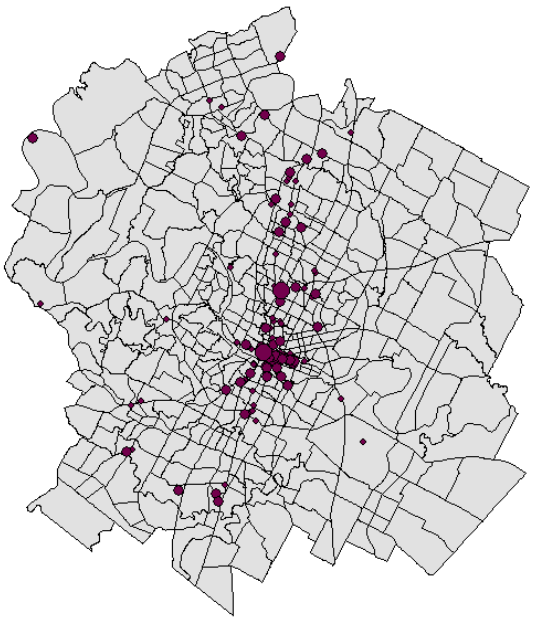
Monday



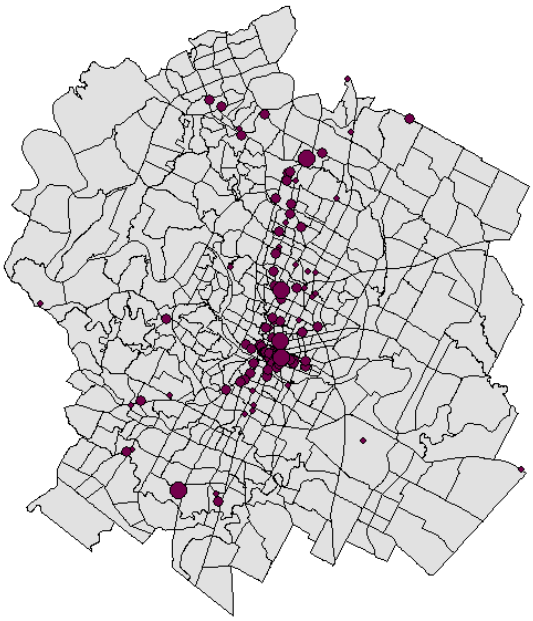
Tuesday



Wednesday

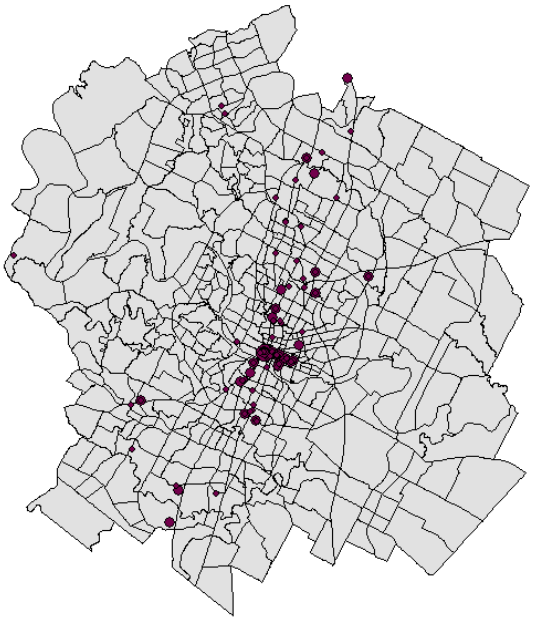


Thursday

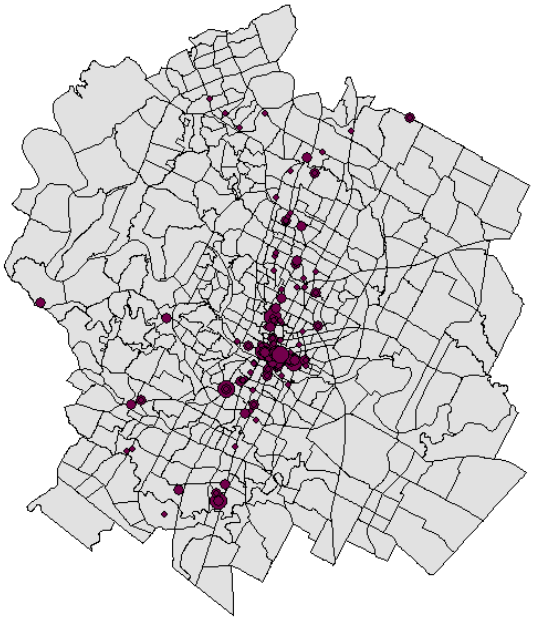


Friday

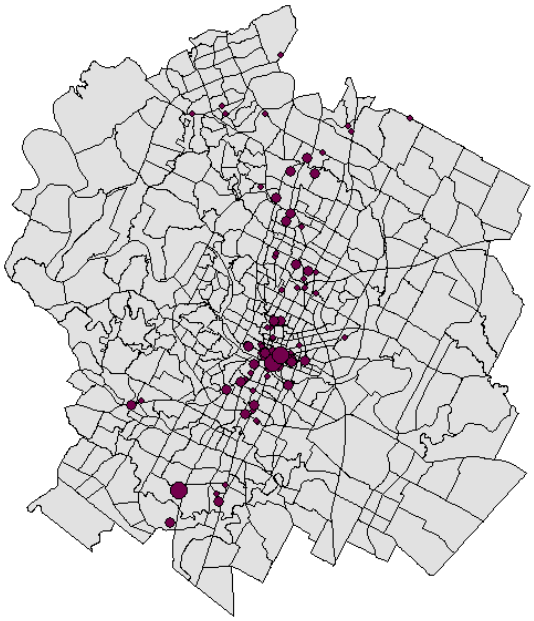
Late Night



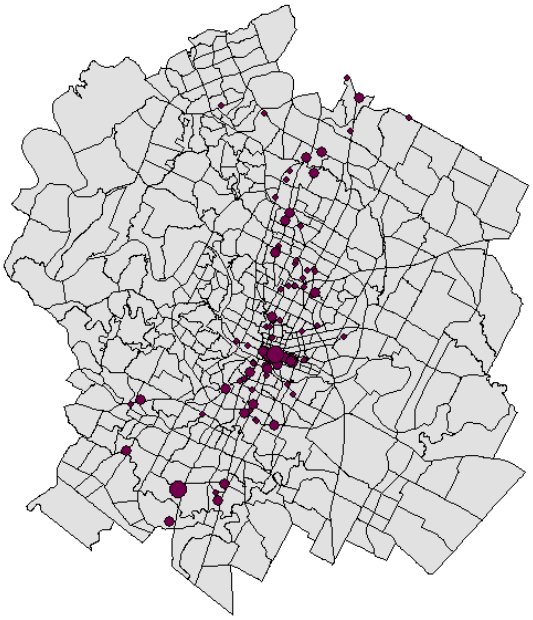
Monday



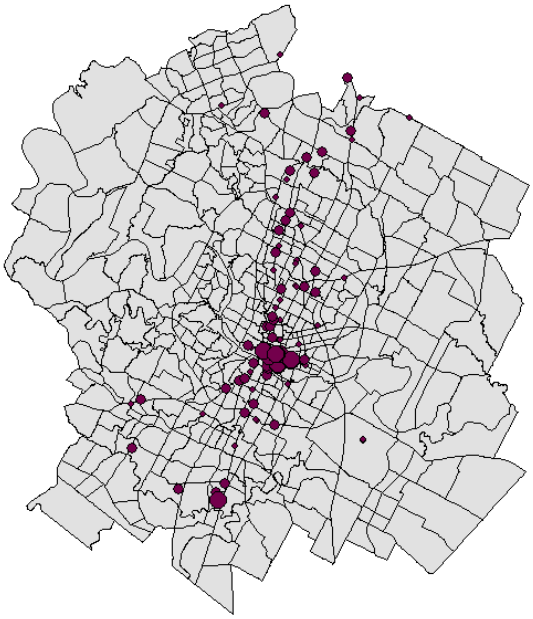
Tuesday



Wednesday



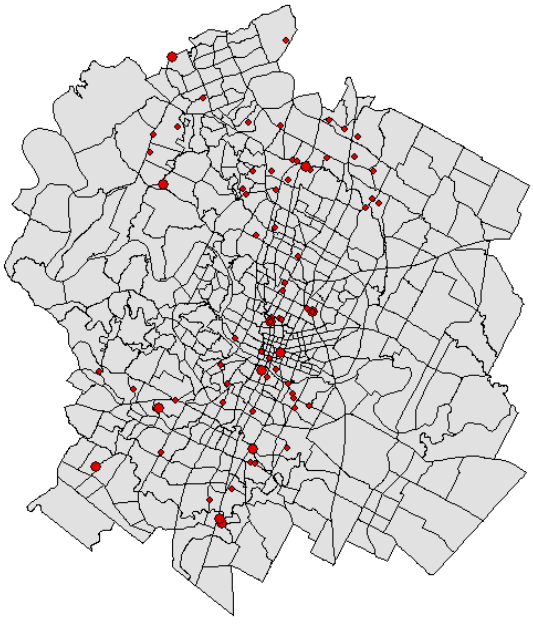
Thursday



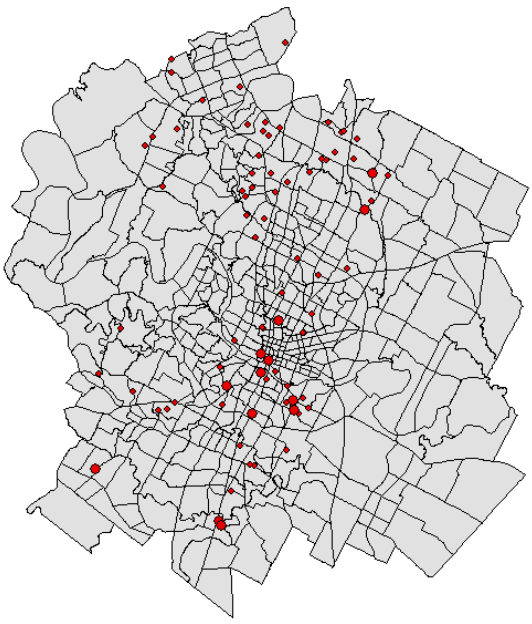
Friday

Residences:

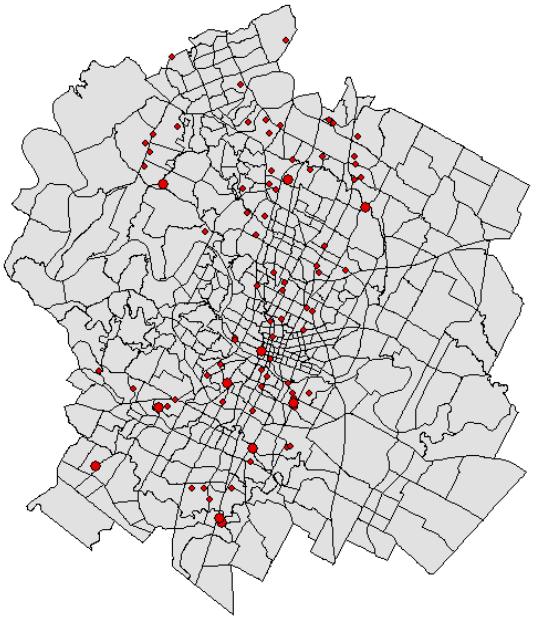
Evening



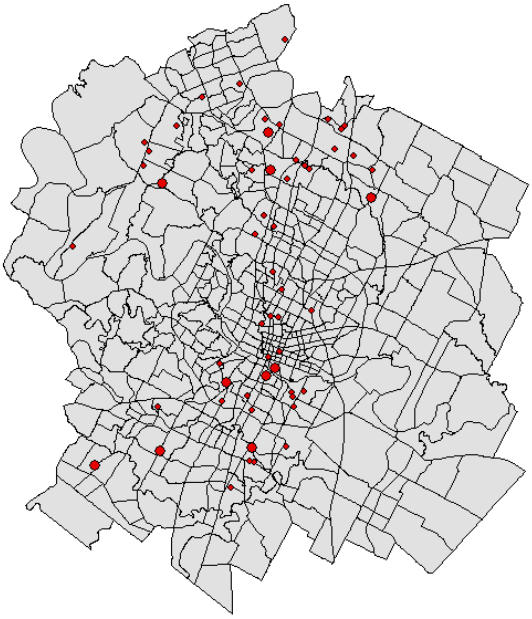
Monday



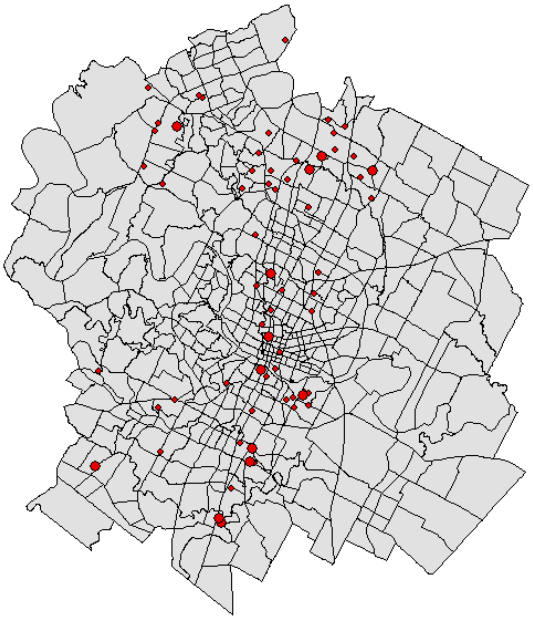
Tuesday



Wednesday

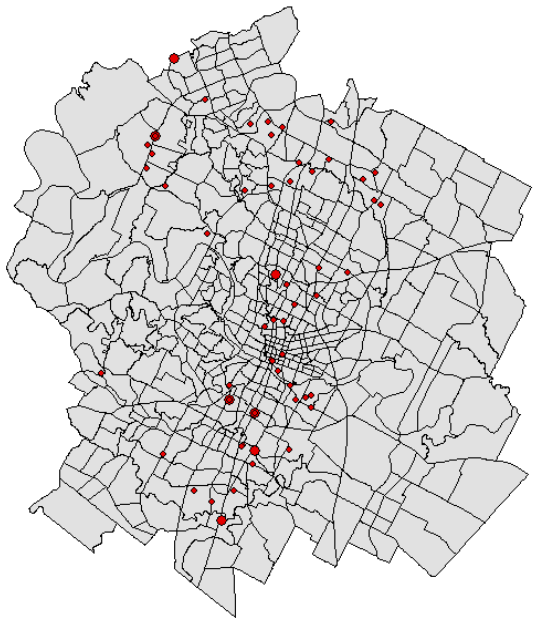


Thursday

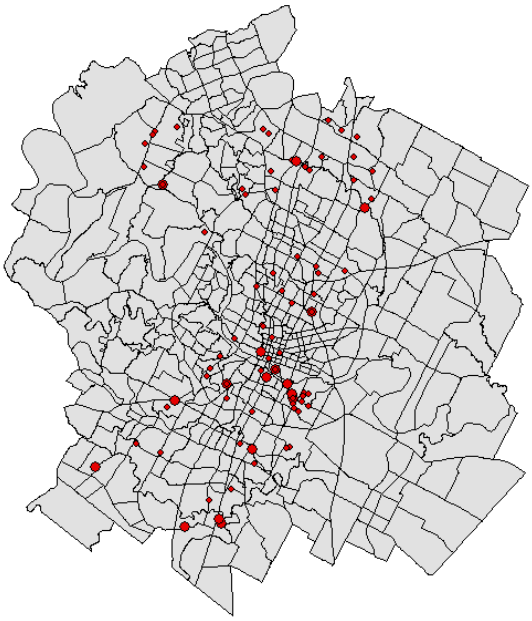


Friday

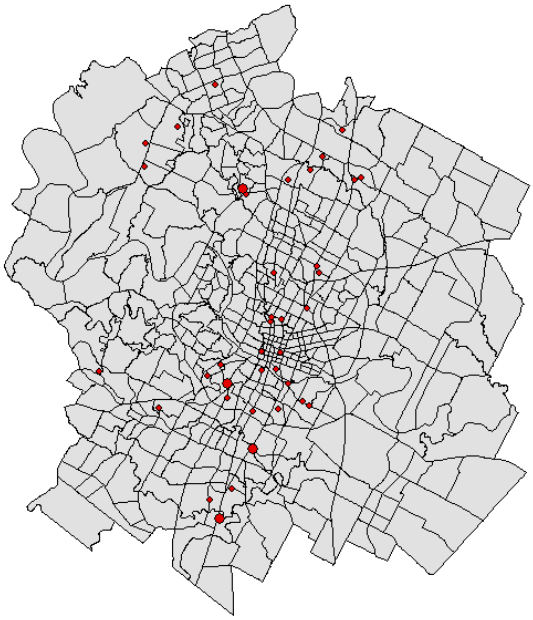
Late Night



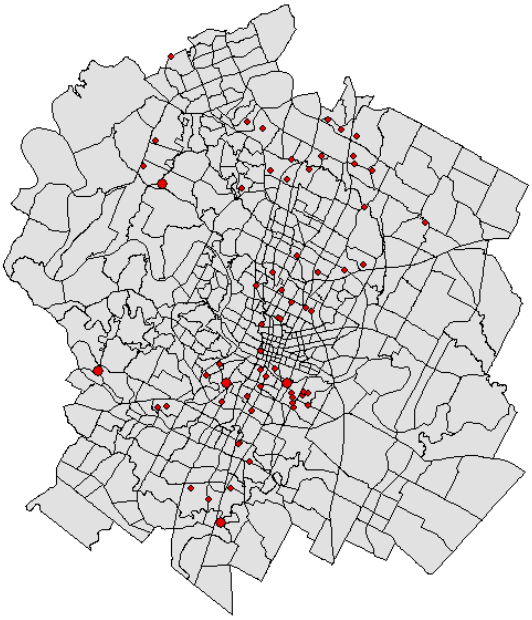
Monday



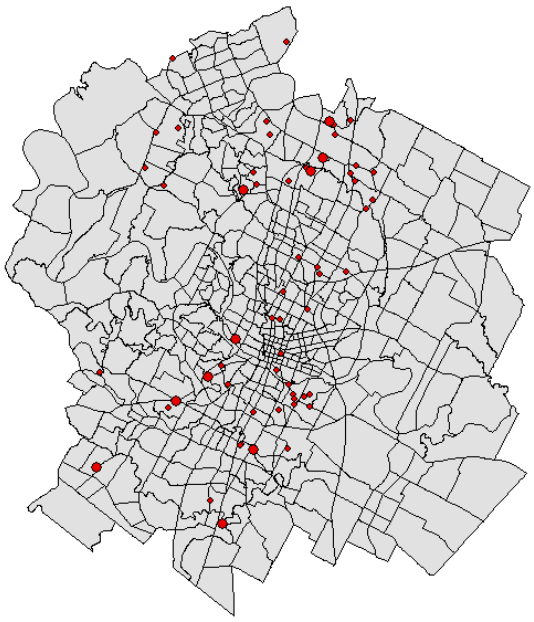
Tuesday



Wednesday



Thursday



Friday

Appendix B

Sample Matlab Code

A sample of the Matlab program code required to perform the analysis of the doubly-constrained gravity and peer-to-peer models are provided in this appendix.

TRIP GENERATION SAMPLE CODE

```
load weekday

weekday = reshape(weekday(:,:),10,1462)';
checkin=weekday(zoneIdx,:);
checkins=reshape(checkin,520,10)';

kp = 0.47334;
ka = 0.66967;
pow = 0.21198;


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,:);

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


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

TRIP DISTRIBUTION SAMPLE CODE

Friction Function Sample Code

```
load centroids
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));
```

…

```
alpha=x(1);
beta=x(2);
alpha1=x(3);
beta1=x(4);
TD = x(5);
```

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

Doubly-Constrained Gravity Model Sample Code

Genetic Optimization Sample Code

```
function doubleGravityOpt
load campo
zoneIdx = csvread('tazid.txt');
totalOD=HBO+HBR+HBUT+HBW+NHBE+NHBO+NHBW+NWAir;
algCells = {@dgravity1};
algNames = {'Doubly1'};


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


lowerBdsCell = {lowerDBds};
upperBdsCell = {upperDBds};


for a=1:1
    alg = algCells{a};
    algName = algNames{a};
    lowerBds = lowerBdsCell{a};
    upperBds = upperBdsCell{a};

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


CR=0;
swapRatio=0;
params=[];


options = gaoptimset('Generations', 100, 'TolFun', 1.000e-03);
    [params,fav,exitflag,output] =
ga(@(x)eva(x,checkins,'CR',totalOD,tripdist,n,alg),nVars,[],[],[],[],lowerBds,upperBds)


[swapRatio,CR,FR,MAE,ME] = eva(params,checkins,'SwapRatio',totalOD,tripdist,n,alg)
    display(CR)


truTSum = sum(sum(totalOD));
predictedTrips1 = dgravity1(params,checkins,tripdist,n,truTSum);

truT = reshape(totalOD,1,[]);
algDT = reshape(predictedTrips1,1,[]);


save(['res_D1_allpurpose.mat'],'params','CR','FR','ME','MAE','swapRatio','totalOD','predi
ctedTrips1');


csvwrite(['res_D1_ProAttHeatMap.csv'],[zoneIdx sum(totalOD,2) sum(predictedTrips1,2)
sum(totalOD)' sum(predictedTrips1)']);

display('optimization done.')


fig1=figure(1)
interval=100;
totalLength=3000;
CR = compareTripLengthDist(totalOD,predictedTrips1,tripdist,interval,totalLength)
```

```
saveas(fig1,'cr1.fig')

end
```

Doubly-Constrained Trip Distribution Sample Code

```
function [Tcur]=dgravity1(x,checkins,tripdist,n,truTSum)
TUpperBd = x(6);
TLowBd = x(7);
adjMid = x(8);
alphaj=ones(1,n);
betai=ones(1,n);


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


AS=alphaj.*attraction;
PS=betai.*production;


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


prevDif = 1;
curDif = 0;
stepCnt = 0;


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


    prevAlphaj = alphaj;
    prevBetai = betai;


    betai=1./((alphaj.*Aj)*friction');
    alphaj=1./((betai.*Pi)*friction);
    AS=alphaj.*attraction;
    PS=betai.*production;
    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));
```

Peer-to-Peer Model Sample Code

```
function [Tcur]=P2P1(x,checkins,tripdist,n,truTSum,venues)
```

```
load venues

k = x(6);
s = x(7);
TUpperBd = x(8);
TLowBd = x(9);
adjMid = x(10);

friction = friction.^s;
delta = ones(n);
friction1 = (delta./friction);

venues = venues*k;
friction1 = venues.*friction1;

AS = attraction;
PS = production;

Tcur =(AS'*PS).*friction1;

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));
```

EVALUATION SAMPLE CODE

```
function [z,CR,FR,MAE,ME] = eva(x,checkins,obj,totalOD,tripdist,n,alg)

truTSum = sum(sum(totalOD));
predictedTrips = alg(x,checkins,tripdist,n,truTSum);

truT = reshape(totalOD,1,[]);
algT = reshape(predictedTrips,1,[]);

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

switch obj

    case 'CR'
            z = -CoincidentRatio(totalOD,predictedTrips,tripdist);

    case 'SwapRatio'
            z = swapRatio(truT,algT);

    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);


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,predictedTripsPercen
t));

end

end



function sr = 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);

end
```

Trip Length Comparison Sample Code

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

```
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)+predictedTripsPercent(nn);
    end
end
subplot(1,2,1)
CR=sum(min(tripsPercent,predictedTripsPercent))/sum(max(tripsPercent,predictedTripsPercen
t));
plot(m,tripsPercent,'o',m,predictedTripsPercent,'*')
xlabel('Trip Length (mile)')
ylabel('Percentage')
set(gca,'XTickLabel',str2double(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',str2double(get(gca,'XTickLabel'))/100);
hleg1 = legend('Survey Trips','Predicted Trips');
title('(b)Cumulative Trip Length Distribution');
```

Intensity Diagram Sample Code

```
function colorDiagram

load res_P1_allpurpose
close all

fig=figure
drawTruOD = log10(totalOD);
drawTruOD(drawTruOD<0) = 0;
```

```
drawAlgOD = log10(predictedTripsP2P_1);
drawAlgOD(drawAlgOD<0) = 0;


sumLog10Tru = sum(sum(log10(totalOD)))
sumLog10Alg = sum(sum(log10(predictedTripsP2P_1)))
hold on
subplot(221)
colormap;
pcolor(drawTruOD);
xlabel('Destination Zone')
ylabel('Origin Zone')
title('CAMPO OD Matrix (Log10)')
axis square
shading flat


subplot(222)
pcolor(drawAlgOD)
xlabel('Destination Zone')
ylabel('Origin Zone')
title('Foursquare OD Matrix (Log10)')
axis square
shading flat


subplot(223)
pcolor(log10(abs(totalOD-predictedTripsP2P_1)))
xlabel('Destination Zone')
ylabel('Origin Zone')
title('OD MAE Matrix (Log10)')
axis square
shading flat
caxis([0 6])
colorbar
hold off
saveas(fig,'ODCompare_P1.fig')


subplot(224)
hold on
[truHist,truC]=hist(reshape(totalOD,1,[]),100);
[algHist,algC] = hist(reshape(predictedTripsP2P_1(drawAlgOD<6000),1,[]),100);
plot(truC,log10(truHist),'-*g',algC,log10(algHist),'-^k')
legend('CAMPO','Foursquare')


xlabel('OD Trip Frequency')
ylabel('Number of OD Pairs (Log10)')
hold off
```

References

"AdMob." Google. Accessed November 19, 2014. www.google.com/ads/admob/

Abdulazim, Tamer, Hossam Abdelgawad, Khandker M. Nurul Habib, and Baher Abdulhai. "A Framework to Automate Travel Activity Inference Using Land-Use Data: The Case of Foursquare in the Greater Toronto and Hamilton Area." In Transportation Research Board 94th Annual Meeting, no. 15-5850. 2015.

Aboltins, Krisjanis, and Baiba Rivza. "The Car Aftersales Market Development Trends in the New Economy." Procedia-Social and Behavioral Sciences 110. 2014: 341-352.

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

AirSage. AirSage, Inc Accessed: August 11, 2013. http://www.airsage.com/

AirSage. "Cast a Wider Planning Net: Using Cost-Effective Technology to Analyze Regional and Rural Locations." www.airsage.com Visited 2014.

Agard, Bruno, Catherine Morency, and Martin Trépanier. "Mining public transport user behaviour from smart card data." In 12th IFAC symposium on information control problems in manufacturing-INCOM, pp. 17-19. 2006.

Alesiani, Francesco, Konstantinos Gkiotsalitis, and Roberto Baldessari. "A Probabilistic Activity Model for Predicting the Mobility Patterns of Homogeneous Social Groups Based on Social Network Data." In The 93rd Annual Meeting of Transportation Research Board. 2014.

Alexa. Alexa Internet, Inc. Accessed: 14 April 2015. http://www.alexa.com/

Amad, Mourad, Ahmed Meddahi, Djamil Aïssani, and Gilles Vanwormhoudt. "GPM: A generic and scalable P2P model that optimizes tree depth for multicast

communications." International Journal of Communication Systems 25, no. 4. 2012: 491-514.

Austin Chamber. The Austin Chamber of Commerce. Accessed May 7, 2015. https://www.austinchamber.com/site-selection/greater-austin-profile/employers.php

Backstrom, L., E. Sun and C. Marlow. “Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity.” Proceedings of the 19th International Conference on World Wide Web, pp. 61-70. ACM, 2010.

Bagchi, Mousumi, and P. R. White. "The potential of public transport smart card data." Transport Policy 12, no. 5 (2005): 464-474.

Ballús-Armet, Ingrid, Susan A. Shaheen, Kelly Clonts, and David Weinzimmer. "Peer-to-Peer (P2P) Carsharing: Exploring Public Perception and Market Characteristics in the San Francisco Bay Area." In Transportation Research Board 93rd Annual Meeting, no. 14-4286. 2014.

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.

Barshan, Maryam, Mahmood Fathy, and Saleh Yousefi. "Improving the availability of P2P-based network management systems by provisioning fault tolerance property." The Journal of Supercomputing 61, no. 3. 2012: 912-934.

Bartle, Caroline, Erel Avineri, and Kiron Chatterjee. "Online information-sharing: A qualitative analysis of community, trust and social influence amongst commuter cyclists in the UK." Transportation Research Part F: Traffic Psychology and Behaviour 16. 2013: 60-72.

Bhat, Chandra R. “Conducting Travel Surveys Using Portable Devices: Technological Challenges.” Presentation at the 10th International Conference on Transport Survey Methods, Leura, Australia, November 16-21, 2014.

Bhat, Chandra R., and Frank S. Koppelman. "Activity-based modeling of travel demand." In Handbook of Transportation Science, pp. 35-61. Springer US, 1999.

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

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.

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.

Bogenberger, Richard, and Timo Kosch. "Ad-hoc Peer-to-peer Communication-webs on the Street." In 9th World Congress on Intelligent Transport Systems. 2002.

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

Boyson, Sandor, Thomas Corsi, and Alexander Verbraeck. "The e-supply chain portal: a core business model." Transportation Research Part E: Logistics and Transportation Review 39, No. 2. 2003: 175-192.

Brenner, Joanna. “Pew Internet: Mobile” Pew Internet & American Life Project.Pew Research Center. June 6, 2013. Accessed: July 8, 2013. http://www.pewinternet.org/

Bricka, Stacey. "Non-response challenges in GPS-based surveys." In Resource Paper Prepared for the International Steering Committee on Travel Survey Conference,

http://ganymede.Nustats.com/nustats_dot_com/templates/yet_again_newmenu/docs/great_reads/Nonresponse_GPS_BasedSurveys.pdf. 2008.

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

Briesemeister, Linda, Lorenz Schafers, and Günter Hommel. "Disseminating Messages Among Highly Mobile Hosts Based on Inter-vehicle Communication." In Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE, pp. 522-527. IEEE, 2000.

Brokke, Glenn E., and William L. Mertz. "Evaluating Trip Forecasting Methods with an Electronic Computer." Highway Research Board Bulletin 203. 1958.

Brooks, Chad. “The Most Productive Day of the Workweek Is…” *Business News Daily*. Accessed June 5, 2015. http://www.businessnewsdaily.com/5637-the-most-productive-day-of-the-workweek-may-surprise-you.html

Caceres, N., J. P. Wideberg, and F. G. Benitez. "Deriving Origin-Destination Data from a Mobile Phone Network." Intelligent Transport Systems no. 1 (1). 2007:15-26.

Cain, J. Bibb, Tom Billhartz, Larry Foore, Edwin Althouse, and John Schlorff. "A Link Scheduling and Ad Hoc Networking Approach using Directional Antennas." In Military Communications Conference, 2003. MILCOM'03. 2003 IEEE, vol. 1, pp. 643-648. IEEE, 2003.

Caizzone, Giuseppe, Walter Erangoli, Paolo Giacomazzi, and Giacomo Verticale. "An Enhanced GPSR Routing Algorithm for TDMA-based Ad-hoc Networks." In Global Telecommunications Conference, 2005. GLOBECOM'05. IEEE, vol. 5, pp. 6-pp. IEEE, 2005.

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.

Cao, Jianping, Ke Zeng, Hui Wang, Jiajun Cheng, Fengcai Qiao, Ding Wen, and Yanqing Gao. "Web-Based Traffic Sentiment Analysis: Methods and Applications." IEEE Transactions on Intelligent Transportation Systems 15, no. 2. 2014.

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.

Castiglione, Joe, Mark Bradley, and John Gliebe. Activity-Based Travel Demand Models: A Primer. No. SHRP 2 Capacity Project C46. 2014.

Cebelak, M. K., "Location-based Social Networking Data: Doubly-Constrained Gravity Model Origin-Destination Estimation of the Urban Travel Demand for Austin, TX." Master's thesis, University of Texas at Austin, 2013.

Cebelak, M. K., P. J. Jin, and C. M. Walton. Location-Based Social Networking: Moving Toward Mode Choice. 10th ITS European Congress, 2014.

Chappell, Brian "2012 Social Netowrk analysis Report." *Ignite Social Media*. Accessed June 23, 2013. http://www.ignitesocialmedia.com/social-media-stats/2012-social-network-analysis-report/#Foursquare

Chen, Roger B., Nathan McNeil, and Jennifer L. Dill. "Exploring Demographic Market Segments for Peer-to-Peer Car-sharing Programs." In Transportation Research Board 93rd Annual Meeting, no. 14-3474. 2014.

Chen, Ying, Andreas Frei, and Hani S. Mahmassani. "From Personal Attitudes to Public Opinion." Transportation Research Record: Journal of the Transportation Research Board 2430, no. 1 (2014): 28-37.

Chen, Ying, Alireza Talebpour, and Hani S. Mahmassani. "Friends Don't Let Friends Drive on Bad Routes: Modeling the Impact of Social Networks on Drivers' Route Choice Behavior." In Transportation Research Board 94th Annual Meeting, no. 15-4974. 2015.

Cheng, Z., J. Caverlee and K. Lee. "You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users." Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759-768. AMC, 2010.

Cheng, Z., J. Caverlee, K. Lee and D. Z. Sui. "Exploring Millions of Footprints in Location Sharing Services." ICWSM 2011. 2011: 81-88.

Chervenak, Ann, and Shishir Bharathi. "Peer-to-Peer Approaches to Grid Resource Discovery." In Making Grids Work, pp. 59-76. Springer US, 2008.

Cho, E., S. A. Myers and J. 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, pp. 1082-1090. ACM, 2011.

"CLRSearch." CLRChoice, Inc. Accessed November 10, 2012. http://www.clrsearch.com/Austin-Demographics/TX/

Çolak, Serdar, Lauren P. Alexander, Bernardo Guatimosim Alvim, Shomik R. Mehndiretta, and Marta C. Gonzalez. "Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities." In Transportation Research Board 94th Annual Meeting, no. 15-5279. 2015.

Collins, Craig, Samiul Hasan, and Satish Ukkusuri. "A Novel Transit Riders' Satisfaction Metric: Riders' Sentiments Measured from Online Social Media Data." Journal of Public Transportation 16, no. 2. 2013.

Coplen, Michael. Clear Signal for Action Program Addresses Locomotive Cab Safety Related to Constraining Signals. No. RR07-08. 2007.

"CV Pilot Deployment Project." Intelligent Transportation Systems Joint Program Office. Accessed November 24, 2014. http://www.its.dot.gov/pilots/cv_pilot_progress.htm

"Demographic Data." The City of Austin. Accessed May16, 2015. http://www.austintexas.gov/page/demographic-data.

Darlagiannis, Vasilios. "21. Hybrid Peer-to-Peer Systems." In Peer-to-Peer Systems and Applications, pp. 353-366. Springer Berlin Heidelberg, 2005.

Davis, Beth R., and John T. Mentzer. "Logistics Service Driven Loyalty: An Exploratory Study." Journal of Business Logistics. Vol. 27, No. 2 (2006): 53-73.

Deroo, Roland, and Jean-Michel Auberlet. "A First Macroscopic-Microscopic Pedestrian Model: Results in the Case of a Unidirectional Flow." Transportation Research Board (TRB) 91st Annual Meeting. 2012.

Devillaine, Flavio, Marcela Munizaga, and Martin Trépanier. "Detection of activities of public transport users by analyzing smart card data." Transportation Research Record: Journal of the Transportation Research Board 2276, no. 1 (2012): 48-55.

Dill, Jennifer, Steven Howland, and Nathan McNeil. "Peer-to-Peer Carsharing: An Preliminary Analysis of Vehicle Owners in Portland, Oregon, and the Potential to

Meet Policy Objectives." In Transportation Research Board 93rd Annual Meeting, no. 14-5576. 2014.

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.

Doran, Derek, Swapna Gokhale, and Karthik C. Konduri. "Participatory Paradigms: Promises and Challenges for Urban Transportation." In Transportation Research Board 93rd Annual Meeting, no. 14-4507. 2014.

Duives, Dorine C., Winnie Daamen, and Serge P. Hoogendoorn. "State-of-the-Art Crowd Motion Simulation Models." Transportation Research Part C: Emerging Technologies 37 (2013): 193-209.

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

Efthymiou, Dimitrios, and Constantinos Antoniou. "Use of social media for transport data collection." Procedia-Social and Behavioral Sciences 48 (2012): 775-785.

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.

“Facebook.” Facebook. Accessed November 19, 2014. www.facebook.com/facebook/info?tab=page_info

Festag, Andreas, H. Füßler, Hannes Hartenstein, Amardeo Sarma, and Ralf Schmitz. "Fleetnet: Bringing car-to-car communication into the real world."Computer 4, no. L15 (2004): 16.

Filali, Imen, Francesco Bongiovanni, Fabrice Huet, and Françoise Baude. "A survey of structured P2P systems for RDF data storage and retrieval." In Transactions on large-

scale data-and knowledge-centered systems III, pp. 20-55. Springer Berlin Heidelberg, 2011.

“Find My Friends.” Apple. Accessed November 19, 2014. www.apple.com/apps/find-my-friends/

Fiorentino, A., C. De Gioia, M. Gaido, G. Conti, D. Magliocchetti, R. De Amicis, and W. Kipp. "Mobile Integration Platform Concept: The Naples Pilot Test Site." Procedia-Social and Behavioral Sciences 48 (2012): 1855-1864.

Fire, Michael, Dima Kagan, Rami Puzis, Lior Rokach, and Yuval Elovici. "Data mining opportunities in geosocial networks for improving road safety." In Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of, pp. 1-4. IEEE, 2012.

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.

“FRED.” Federal Reserve Economic Data. Accessed: May 16, 2015. http://research.stlouisfed.org/fred2/

Friesen, M. R., and R. D. McLeod. "Bluetooth in Intelligent Transportation Systems: A Survey." International Journal of Intelligent Transportation Systems Research (2014): 1-11.

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. Accessed May 17, 2015. https://foursquare.com/

"Foursquare 8 Main Screen" by Source (WP:NFCC#4). Licensed under Fair use via Wikipedia Accessed May 17, 2015. http://en.wikipedia.org/wiki/File:Foursquare_8_Main_Screen.png#/media/File:Foursquare_8_Main_Screen.png

Franke, Markus. "Innovation: The Winning Formula to Regain Profitability in Aviation?" Journal of Air Transport Management. Vol. 13, No. 1 (2007): 23-30.

Füßler, Holger, Martin Mauve, Hannes Hartenstein, Michael Käsemann, and Dieter Vollmer. "A comparison of routing strategies for vehicular ad hoc networks." In Proceedings of MOBICOM. 2002.

Gaille, Brandon. “26 Great Foursquare Demographics.” Brandongaille.com Posted January 13, 2015. Accessed May 17, 2015. http://brandongaille.com/26-great-foursquare-demographics/

Gal-Tzur, Ayelet, Susan M. Grant-Muller, Tsvi Kuflik, Einat Minkov, Silvio Nocera, and Itay Shoor. "The potential of social media in delivering transport policy goals." Transport Policy 32 (2014): 115-123.

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

“Glympse.” Glympse. Accessed November 19, 2014. www.glympse.com/what-is-glympse

“Gnip.” Gnip. Accessed November 25, 2014. http://gnip.com/sources/foursquare/

Goers, Randy. "Who's checking in downtown Tampa?." Planning 79, no. 6 (2013).

“Google Maps About.” Google. Accessed November 19, 2014. www.google.com/maps/about/

Gradowski, Tomasz, Maciej J. Mrowinski, and Robert A. Kosinski. "Cooperation in Peer-to-Peer Networks." Acta Physica Polonica B 41, no. 5 (2010): 1143.

Grant-Muller, Susan M., Ayelet Gal-Tzur, Einat Minkov, Silvio Nocera, Tsvi Kuflik, and Itay Shoor. "The Efficacy of Mining Social Media Data for Transport Policy and Practice." In Transportation Research Board 93rd Annual Meeting, no. 14-1716. 2014.

Gumzej, Roman, and Brigita Gajšek. "Introducing quality of service criteria into supply chain management for excellence." International Journal of Applied Logistics (IJAL) 2, no. 1 (2011): 1-16.

Gumzej, Roman, Panchalee Sukjit, and Herwig Unger. "Modelling Overlay Networks for Autonomous Supply Chain Systems." Logistics & Sustainable Transport 3, no. 2 (2012).

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.

Hampshire, Robert C., and Craig Gaites. "Peer-to-Peer Carsharing."Transportation Research Record: Journal of the Transportation Research Board2217, no. 1 (2011): 119-126.

Hasan, Samiul, and Satish V. Ukkusuri. "Urban activity pattern classification using topic models from online geo-location data." Transportation Research Part C: Emerging Technologies 44 (2014): 363-381.

“Hash Table.” Wikipedia. Accessed November 24, 2014. http://en.wikipedia.org/wiki/Hash_table

Helbing, Dirk, Lubos Buzna, Anders Johansson, and Torsten Werner. "Self-organized Pedestrian Crowd Dynamics: Experiments, Simulations, and Design Solutions." Transportation Science Vol. 39, no. 1 (2005): 1-24.

Helbing, Dirk. "A Mathematical Model for the Behavior of Individuals in a Social Field." Journal of Mathematical Sociology Vol. 19, No. 3 (1994): 189-219.

Helbing, Dirk. "A Mathematical Model for the Behavior of Pedestrians." Behavioral Science Vol. 36, No. 4 (1991): 298-310.

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.

Horn, Christopher, Stefan Klampfl, Michael Cik, and Thomas Reiter. "Detecting Outliers in Cell Phone Data." Transportation Research Record: Journal of the Transportation Research Board 2405, no. 1 (2014): 49-56.

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

Huang, Haosheng, and Georg Gartner. "Collective intelligence-based route recommendation for assisting pedestrian wayfinding in the era of Web 2.0."Journal of location based services 6, no. 1 (2012): 1-21.

“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. Accessed: 9 July 2013.

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

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

Janssen, Cory. “Many-to-Many Relationship.” Techopedia. Accessed July 11, 2014, http://www.techopedia.com/definition/27291/many-to-many-relationship.

Jin, P. J., F. Yang, M. Cebelak, B. Ran and C. M. Walton. Urban Travel Demand Analysis for Austin Tx USA Using Location-Based Social Networking Data. Transportation Research Board 92nd Annual Meeting, 2013.

Jin, P. J., M. Cebelak, F. Yang, B. Ran and C. M. Walton. “Location-Based Social Networking Data: An Exploration in to the Use of a Doubly-Constrained Gravity Model for Origin-Destination Estimation.” Transportation Research Board 93rd Annual Meeting, no14-5314. 2014. Accepted for publication, 2014.

Jin, Xing, and S-H. Gary Chan. "Unstructured Peer-to-Peer Network Architectures." In Handbook of Peer-to-Peer Networking, pp. 117-142. Springer US, 2010.

Jin, Wen-Long, and Wilfred W. Recker. "Instantaneous information propagation in a traffic stream through inter-vehicle communication." Transportation Research Part B: Methodological 40, no. 3 (2006): 230-250.

Karimi, H. A. “Genetic Location-Based Social Networks (G-Lbsn).” Proceedings of the 3rd International Workshop on Location and the Web, pp. 9. AMC, 2010.

Karp, Brad, and Hsiang-Tsung Kung. "GPSR: Greedy perimeter stateless routing for wireless networks." In Proceedings of the 6th annual international conference on Mobile computing and networking, pp. 243-254. ACM, 2000.

Kaufman, Sarah M., and Mitchell L. Moss. "Co-Monitoring for Transit Management: Using Web-Based Rider Input for Transit Management." University Transportation Research Center – Region 2. New York University. (2014).

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.

“Company Info” *Facebook*. Accessed April 14, 2015. http://newsroom.fb.com/company-info/

Kosch, Timo, Christian Schwingenschlogl, and Li Ai. "Information dissemination in multihop inter-vehicle networks." In Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on, pp. 685-690. IEEE, 2002.

Kretz, Tobias, Stefan Hengst, and Peter Vortisch. "Pedestrian Flow at Bottlenecks-Validation and Calibration of Vissim's Social Force Model of Pedestrian Traffic and its Empirical Foundations." arXiv preprint arXiv:0805.1788 (2008).

"Labor Force Statstics from the Current Population Survey." Bureau of Labor Statistics. Accessed November 25, 2014. Last Modified February 26, 2014. http://www.bls.gov/cps/cpsaat03.htm

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.

Lee, Jin-Ki, Soon-Ki Jo, and Ki-Jung Kum. "The Second Step SMART Highway Services In Korea-Use Wave and SNS." In 18th ITS World Congress. 2011.

Leinbach, Thomas R. Globalized Freight Transport: Intermodality, E-Commerce, Logistics and Sustainability. Edward Elgar Publishing, 2007.

Li, Deng, Hui Liu, and Athanasios Vasilakos. An efficient, scalable and robust p2p overlay for autonomic communication. Springer US, 2009.

Li, N. and G. Chen. "Analysis of a Location-Based Social Network." Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4, pp. 263-270. IEEE, 2009.

"Life360." Life360. Accessed November 19, 2014. www.life360.com

Link, Michael W., Joe Murphy, Michael F. Schober, Trent D. Buskirk, Jennifer Hunter Childs, Casey Langer Tesfaye, Mario Callegaro et al. "Mobile Technologies for

Conducting, Augmenting and Potentially Replacing Surveys: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research." (2014).

Marozzo, Fabrizio, Domenico Talia, and Paolo Trunfio. "P2P-MapReduce: Parallel data processing in dynamic Cloud environments." Journal of Computer and System Sciences 78, no. 5 (2012): 1382-1402.

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).

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

Medina, Daniel, Felix Hoffmann, Francesco Rossetto, and C-H. Rokitansky. "A crosslayer geographic routing algorithm for the airborne internet." InCommunications (ICC), 2010 IEEE International Conference on, pp. 1-6. IEEE, 2010.

Medina, Daniel, Felix Hoffmann, Serkan Ayaz, and C-H. Rokitansky. "Topology characterization of high density airspace aeronautical ad hoc networks." InMobile Ad Hoc and Sensor Systems, 2008. MASS 2008. 5th IEEE International Conference on, pp. 295-304. IEEE, 2008.

“Métro.” Métro. Accessed November 19, 2014. www.metro.nanika.net

Millard-Ball, Adam. Car-Sharing: Where and how it succeeds. Vol. 108. Transportation Research Board, 2005.

Min, Hokey. Developing a Model-based Decision Support System for Call-A-Ride Paratransit Service Problems. No. MIOH UTC TS13 20011–Final. 2011.

Mirtaheri, Seyedeh Leili, Ehsan Mousavi Khaneghah, Mohsen Sharifi, Behrouz Minaei-Bidgoli, Bijan Raahemi, Mohammad Norouzi Arab, and Abbas Saleh Ardestani. "Four-dimensional model for describing the status of peers in peer-to-peer distributed systems." Turkish Journal of Electrical Engineering & Computer Sciences 21, no. 6 (2013): 1646-1664.

Misra, Aditi, Aaron Gooze, Kari Watkins, Mariam Asad, and Christopher A. Le Dantec. "Crowdsourcing and Its Application to Transportation Data Collection and Management." In Transportation Research Board 93rd Annual Meeting, no. 14-3358. 2014.

Molin, Eric, Theo Arentze, and Harry JP Timmermans. "Social activities and travel demand: model-based analysis of social network data." Transportation Research Record: Journal of the Transportation Research Board 2082, no. 1 (2008): 168-175.

Morris, William P., Kelly Robertson, and Jeremy Spinks. Utilizing Information Technology in Innovative Marketing Approaches for Public Transportation. No. BD-549-53. 2009.

Munizaga, Marcela A., and Carolina Palma. "Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile." Transportation Research Part C: Emerging Technologies 24 (2012): 9-18.

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.

Naor, Moni, and Udi Wieder. "Novel architectures for P2P applications: the continuous-discrete approach." ACM Transactions on Algorithms (TALG) 3, no. 3 (2007): 34.

“Napster.” Wikipedia. Accessed November 19, 2014. http://en.wikipedia.org/wiki/Napster

Neumayer, Robert, Christos Doulkeridis, and Kjetil Nørvåg. "A hybrid approach for estimating document frequencies in unstructured P2P networks." Information Systems 36, no. 3 (2011): 579-595.

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

O'Flaherty, Coleman, ed. Transport planning and traffic engineering. Elsevier, 1997.

Pagano, Anthony M. "Factors and Trends Affecting the Parcel Delivery Industry." In Journal of the Transportation Research Forum, Vol. 40, No. 3. 2001.

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.

Park, Ji-Young, Jung Ung Min, and Jeong Soo Park. "Analysis of Causal Relationship between Supply Chain Security and Its Performance Using Balanced Scorecard Model." Journal of International Logistics and Trade. Vol. 9, No. 2 (2011): 99-118.

Park, Jin Young, and Dong Jun Kim. "The Potential of Using the Smart Sard Data to Define the Use of Public Transit in Seoul." Journal of the Transportation Research Board 2063, no. 1 (2008): 3-9.

Pelletier, Marie-Pier, Martin Trépanier, and Catherine Morency. "Smart card data use in public transit: A literature review." Transportation Research Part C: Emerging Technologies 19, no. 4 (2011): 557-568.

Popa, L., C. Raiciu, I. Stoica, and D. S. Rosenblum. "Reducing congestion effects by multipath routing in wireless networks." (2006): 96-105.

Rabah, M. Y., and H. S. Mahmassani. Impact of Electronic Commerce on Logistics Operations: A Focus on Vendor Managed Inventory (VMI) Strategies. No. SWUTC/01/167227-1, 2002.

Ranjan, Rajiv, Aaron Harwood, and Rajkumar Buyya. "Peer-to-peer-based resource discovery in global grids: a tutorial." Communications Surveys & Tutorials, IEEE 10, no. 2 (2008): 6-33.

Ranney, J., Wu, S., Austin, C., and Coplen, M. “Positive Safety Outcomes of Clear Signal for Action Program at Union Pacific Yard Operations.” U.S. DOT Federal Railroad Administration [RR08-09] (2008, June). http://www.fra.dot.gov/eLib/details/L03482

Rasouli, Soora, ed. Mobile technologies for activity-travel data collection and analysis. IGI Global, 2014.

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

Rivas, David Antolino, and Manel Guerrero-Zapata. "Simulation of points of interest distribution in vehicular networks." Simulation (2012): 0037549712456440.

Rivasplata, Charles R., Zhan Guo, Richard Lee, David Keyon, and Luis Schloeter. Residential On-Site Carsharing and Off-Street Parking Policy in the San Francisco Bay Area. No. CA-MTI-12-1001-1. 2012.

Rudloff, Christian, Thomas Matyus, Stefan Seer, and Dietmar Bauer. "Can Walking Behavior Be Predicted?" Transportation Research Record: Journal of the Transportation Research Board 2264, No. 1 (2011): 101-109.

SA, Rokib, Md Ahsanul Karim, Tony Z. Qiu, and Amy Kim. "Origin-Destination Trip Estimation from Anonymous Cell Phone and Foursquare Data." In Transportation Research Board 94th Annual Meeting, no. 15-2379. 2015.

Sabra, Ziad A., and Keith A. Riniker. "Maximizing Benefits of Signal Timing Optimization." In ITE 2009 Annual Meeting and Exhibit. 2009.

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, S., A. Noulas, R. Lambiotte and C. 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.

Sener, Ipek N., Nazneen Ferdous, Chandra R. Bhat, and Phillip Reeder. Tour-based model development for TxDOT: evaluation and transition steps. No. FHWA/TX-10/0-6210-2. 2009.

Shaheen, Susan A., and Adam P. Cohen. "Carsharing and personal vehicle services: worldwide market developments and emerging trends." International Journal of Sustainable Transportation 7, no. 1 (2013): 5-34.

Shaheen, Susan A., Mark A. Mallery, and Karla J. Kingsley. "Personal vehicle sharing services in North America." Research in Transportation Business & Management 3 (2012): 71-81.

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

Smith, Aaron. “Usage and Adoption.” Pew Research Internet Project. April 3, 2014. Accessed November 24, 2014. http://www.pewinternet.org/2014/04/03/usage-and-adoption/

Spurr, Tim, Robert Chapleau, and Daniel Piché. "Use of Subway Smart Card Transactions for the Discovery and Partial Correction of Travel Survey Bias." Transportation Research Record: Journal of the Transportation Research Board 2405, no. 1 (2014): 57-67.

Sukjit, Panchalee. Regularising Amorphous Peer-to-peer Networks with Overlay Grids Generated Locally. VDI Verlag, 2011.

"Swarm app checkin screen" by Source. Licensed under Fair use via Wikipedia - http://en.wikipedia.org/wiki/File:Swarm_app_checkin_screen.jpg#/media/File:Swarm_app_checkin_screen.jpg

The Road Information Program (TRIP). Stuck in Traffic: How Increasing Traffic Congestion is Putting the Brakes on Economic Growth. Road Information Program, 2001.

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

Tisdale, Stacey M. U in the Driver Seat—A Peer-to-Peer Pilot Program for Decreasing Car Crashes by College Students. No. SWUTC-14/600451-00015-1. 2013.

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

Toole, Jameson L., Carlos Herrera-Yaqüe, Christian M. Schneider, and Marta C. González. "Coupling human mobility and social ties." Journal of The Royal Society Interface 12, no. 105 (2015): 20141128.

Tornero, Rafael, Javier Martínez, and Joaquín Castelló. "A Multi-Agent System for Obtaining Dynamic Origin/Destination Matrices on Intelligent Road Networks." In Proceedings of the 6th Euro American Conference on Telematics and Information Systems, pp. 157-164. ACM, 2012.

Trépanier, Martin, Catherine Morency, and Bruno Agard. "Calculation of transit performance measures using smartcard data." Journal of Public Transportation 12, no. 1 (2009): 79-96.

"Twitter." Twitter. Accessed on November 19, 2014. www.twitter.com/?lang=en

"U.S. Census Bureau (USCB)." USCB. June 27, 2013. Accessed July 12, 2013. http://www.census.gov/

"U.S. Wireless Quick Facts." CTIA -The Wireless Association. n.d. Accessed July 8, 2013. http://www.ctia.org/your-wireless-life/how-wireless-works/wireless-quick-facts

van den Berg, Pauline, Theo A. Arentze, and Harry JP Timmermans. "Size and composition of ego-centered social networks and their effect on geographic distance and contact frequency." Transportation Research Record: Journal of the Transportation Research Board 2135, no. 1 (2009): 1-9.

Van Zuylen, Henk J., and Luis G. Willumsen. "The most likely trip matrix estimated from traffic counts." Transportation Research Part B: Methodological14.3 (1980): 281-293.

Wall, Thomas A., Gregory S. Macfarlane, and Kari Edison Watkins. "Exploring the Use of Egocentric Online Social Network Data to Characterize Individual Air Travel Behavior." Transportation Research Record: Journal of the Transportation Research Board 2400, no. 1 (2014): 78-86.

Wang, Q., and J. E. Taylor. "Massive Online Geo-Social Networking Platforms and Urban Human Mobility Patterns: A Process Map for Data Collection." In 2014 International Conference on Computing in Civil and Building Engineering. 2014.

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 Board1964.1 (2006): 260-269.

"Waze About Us." Waze. Accessed November 19, 2014. www.waze.com/about.

"WeatherPro." WeatherPro. Accessed November 19, 2014, http://www.weatherpro.eu/home.html

Weiner, Edward. Urban Transportation Planning in the United States. An Historical Overview. No. DOT-I-86-09. 1986.

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

Werberich, Bruno Rocha, Carlos Oliva Pretto, and Helena Cybis. "Pedestrians' Route Choice Based On Friction Forces Assuming Partial And Full Environment Knowledge." In Transportation Research Board 93rd Annual Meeting, No. 14-3067. 2014.

"What Can We Help You With?" *Foursquare Help Center*. Accessed May 17, 2015. https://support.foursquare.com/hc/en-us

Wikipedia contributors, "Foursquare," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Foursquare&oldid=661006465 (accessed May 17, 2015).

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

Winston, Flaura K., and Lela Jacobsohn. "A practical approach for applying best practices in behavioural interventions to injury prevention." Injury prevention 16, no. 2 (2010): 107-112.

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.

Xu, Ke, Meng Shen, Yong Cui, Mingjiang Ye, and Yifeng Zhong. "A Model Approach to the Estimation of Peer-to-Peer Traffic Matrices." IEEE Transactions on Parallel and Distributed Systems 25, no. 5 (2013): 1-1.

Yang, Beverly, and Hector Garcia-Molina. "Comparing hybrid peer-to-peer systems." In Proceedings of the 27th Intl. Conf. on Very Large Data Bases. 2001.

Yang, F., 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.

Yang, Xu, and Wilfred W. Recker. "Modeling dynamic vehicle navigation in a self-organizing, peer-to-peer, distributed traffic information system." Journal of Intelligent Transportation Systems 10, no. 4 (2006): 185-204.

Yang, Xu, and Will Recker. "Evaluation of Information Applications of a Self-Organizing Distributed Traffic Information System for a Large-Scale Real-World Traffic Network." Computer-Aided Civil and Infrastructure Engineering 23, no. 8 (2008): 575-595.

“Yelp.” Yelp. Accessed November 19, 2014. http://www.yelp.com/

Yi, Huiming. "Prototype Development of the Open Mode Integrated Transportation System (OMITS)." University Transportation Research Center – Region 2. Columbia University. (2013).

Yi, Su, Yong Pei, and Shivkumar Kalyanaraman. "On the capacity improvement of ad hoc wireless networks using directional antennas." InProceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing, pp. 108-116. ACM, 2003.

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

Yucel, S., Tuydes-Yaman, H., Altintasi, O., and Ozen, M. "Determination of Vehicular Travel Patterns in an Urban Location Using Bluetooth Technology." Presentation at the ITS America Annual Meeting and Expo, Nashville, TN, April 22-24, 2013.

Zheng, Y., X. Xie and W.-Y. Ma. "Geolife: A Collaborative Social Networking Service among User, Location and Trajectory." IEEE Data Eng. Bull. 33, No. 2 (2010): 32-39.

Ziemke, D., Nagel, K. and Bhat, C. Integrating CEMDAP and MATSim to increase the transferability of transport demand models, Proceedings of the 94. Annual Meeting of the Transportation Research Board, Washington, DC, USA. (2015).

Ziliaskopoulos, Athanasios K., and Jiang Zhang. "A zero public infrastructure vehicle based traffic information system." In Transportation Research Board 82nd Annual Meeting. 2003.

"Zillow." Zillow. Accessed November 19, 2014. http://www.zillow.com/