© 2014 Ryan Robert Smith
LARGE-SCALE TRANSIT SERVICE NETWORK DESIGN UNDER CONTINUOUS HETEROGENEOUS DEMAND

BY

RYAN ROBERT SMITH

THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Civil Engineering in the Graduate College of the University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Associate Professor Yanfeng Ouyang
ABSTRACT

The design of public transit systems typically relies on discrete or continuous models to determine route layout and service frequency. To avoid computational complexity associate with discrete models, continuous models are proposed to formulate the problem in terms of a few key continuous variables. One drawback of many continuous models is that they typically assume a uniform distribution of passenger trips, however we know that trip origins and destinations are characteristically not uniform. The purpose of thesis is to (i) investigate how the design and operations of Daganzo (2010b) hybrid transit system is effected by spatially heterogeneous demand; and (ii) how large is and where to locate a transit network under a spatially heterogeneous demand. The spatially heterogeneity of the passenger demand is captured in the model by transforming a continuous passenger demand density function into zone-to-zone passenger demand. For (i), the optimal hybrid transit structure and operations is analyzed for two distinct spatial demand distributions each with a low and a high passenger demand scenario. The results show the effects on agency and user cost metrics when passenger demand is increased across the region and concentrated (i.e, the origins and destinations of passengers trips are grouped more tightly together in the center of the city). For (ii), automobiles are introduced as an alternative mode of travel and each person who choose this mode incurs a cost associated with not taking transit. This is formulated and people with an origin and/or a destination outside the transit service region are assumed to not take transit. For each transit mode a sensitivity analysis is performed on the penalty for not taking transit. The results indicate that the optimal size of the transit service region (i.e., the number of passengers served by the transit system) is dependent on the not taking transit penalty and the cost and operational characteristics of the transit mode. Finally, multiple future research ideas are presented to illustrate some of the other possibilities that may further enhance transit network design.
To my parents and Jess, for all of their love and support.
I would like to give special thanks to my advisor, Dr. Yanfeng Ouyang, for his continuous support on my work and his advice towards my professional development. His help and guidance made graduate school one of the most valuable experience of my life.
# TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION ........................................... 1

CHAPTER 2 LITERATURE REVIEW ..................................... 4
  2.1 Continuum Approximation ........................................ 4
  2.2 Transit Network Design and Scheduling Problem (TNDSP) Literature .... 4
  2.3 Passenger Demand Literature ...................................... 8
  2.4 Solution Methods .................................................. 11

CHAPTER 3 TRANSIT NETWORK DESIGN WITH SPATIALLY HETERO-
  GENEROUS DEMAND .................................................. 13
  3.1 Formulation ....................................................... 13
  3.2 Numerical Results ................................................ 20
  3.3 Conclusion ....................................................... 26

CHAPTER 4 TRANSIT NETWORK DESIGN WITH SPATIALLY ELASTIC
  DEMAND ............................................................... 28
  4.1 Formulation ....................................................... 29
  4.2 Numerical Results ................................................ 32
  4.3 Conclusion ....................................................... 38

CHAPTER 5 FUTURE RESEARCH ......................................... 41
  5.1 Extensions to the Hybrid Transit Network .......................... 41
  5.2 Multiple Transit Networks ........................................ 44
  5.3 Passenger Demand Extensions ..................................... 48
  5.4 Continuum Approximation Approach ................................ 49
  5.5 Others ............................................................ 50

CHAPTER 6 CONCLUSIONS ............................................... 52

REFERENCES ............................................................ 54

APPENDIX A PROOFS ................................................... 62
  A.1 Transit Network Design with Spatially Heterogeneous Demand .......... 62
As of 2010, more than 50% of the world’s population lived in an urban area, and by 2050 this will increase to more than 70% (World Health Organization, 2013). This growth will greatly increase the demand on a city’s limited infrastructure and land, in particular the transportation infrastructure, resulting in less efficient systems increasing delay to all users. Between 1982 and 2011, automobile drivers in Chicago, Illinois, experienced a 292% increase in delay due to traffic congestion (U.S. Department of Transportation, 2013). On top of this, the price of crude oil has almost quadrupled since the early 2000’s (U.S. Energy Information Administration, 2014). Now, not only are drivers delayed longer, increasing fuel consumption and CO₂ discharge, but they have to pay significantly more for the gasoline. Schrank et al. (2012) estimated that traffic congestion annually: (i) costs the average automobile driver $818, (ii) produces an extra 56 billion lbs of CO₂, and (iii) costs the United States economy $121 billion. As the population continues to increase congestion will continue to plague and degrade urban areas unless sustainable solutions are found.

There are many potential solutions to reduce congestion and/or improve an urban area’s transportation system. The potential solutions can be grouped into the following categories: (i) capacity improvement, (ii) demand management policies, and (iii) modal shifting. Capacity can be increased by either expanding the physical space (e.g., adding an additional lane to a highway) or by improving the efficiency (e.g., adjusting the speed limit to maximize vehicle flow). Braess’s paradox explains how the benefits of increasing capacity can be short lived and may not even benefit the population at large (Braess, 1968). Demand management policies either attempt to shift the time vehicles are on the roadway so as to even out peak versus off-peak demand, or attempt to permanently limit the number of vehicles on the roads. These can be achieved by a variety of methods (e.g., congestion pricing, tax breaks for companies that allow telecommuting). For example, in Beijing, China automobile drivers are limited to driving only on certain days dictated by their license plate number. While these types of polices allow for the current infrastructure to operate more efficiently, there are many societal problems that can arise (e.g., the system disproportionately reduces opportunities for a particular social class). Modal shifting refers to shifting
individual drivers to other modes of travel where similar trips can be aggregating together thus requiring less transportation infrastructure. Public transit has already been shown to decrease congestion and reduce the total cost of congestion by $20.8 billion (Schrank et al., 2012). One relationship that is often ignored is that good public transit systems not only provide good service to transit passengers, but also serve automobile driver’s by reducing traffic congestion (Daganzo et al., 2011).

Public transportation has had its ups and downs in the past century, especially in the United States, due to the increased availability of the automobile to the general population, urban sprawl, limited funding, etc. From 1980 to 2000 there was a 13% increase in the total passenger miles traveled on transit, while highway passenger miles increased 71% (U.S. Department of Transportation, 2013). From 2000 to 2011, due to numerous reasons, highway passenger miles decreased 7% and transit passenger miles has increased 20%, which is more than the rate of population growth in the United States (American Public Transportation Association, 2013). This shows that there has already been an increase in ridership due to people shifting from driving to transit. By improving line capacity, service frequency, coverage, reliability, comfort, safety, and service quality ridership will likely continue to increase (Sinha, 2003; Vuchic, 2005).

The question that arises is how to design and operate an efficient, sustainable, and environmentally friendly public transit system that can provide mobility via a variety of modes (e.g., bus, BRT, light rail, commuter rail, subway, ferry) to a diverse set of users with heterogeneous origins and destinations. To answer this it is important to understand the current steps for planning and operating a public transit system. The public transit planning process typically comprises of strategic and tactical planning, where strategic planning corresponds to long-term decisions, and tactical planning corresponds to relatively short-term decisions. Ceder and Wilson (1986) deconstructs the public transit planning process into 5 steps: (i) network structure and route design, (ii) frequency setting, (iii) timetabling, (iv) vehicle assignment, and (v) crew scheduling. Strategic planning consists of step (i) while tactical planning consists of steps (ii)-(v), although in the literature there is some debate if (ii) should also be considered a strategic decision. Each step by itself can be a multi-objective problem with multiple conflicts of interest between the users, operators, other modes of travel (i.e., auto drivers), and the environment.

With a drastic increase urban population, especially in the developing world where urban growth is estimated to more than double by 2050, there may not be time to adequately zone and plan developments to have a uniform distribution of trip origins and destinations (World Health Organization, 2013). This requires transit agencies to consider the spatially heterogeneous distributions of passenger origins and destinations. Even in urban areas of the
developed world, heterogeneous origins and destinations exist as a result of specialization and pursuit of economies of scale – already built up urban areas typically have a central business districts (CBD), or other high density regions, that attract a majority of automobile and transit trips.

The purpose of this thesis is to (i) investigate how the design and operations of transit systems are effected by spatially heterogeneous demand; and (ii) how large and where to locate transit networks. To study these topics two distinct spatial demand distributions and their transit system’s optimal structure and operations are analyzed for low and high demand scenarios. The results show many insightful results (e.g., users’ total cost for bus and BRT modes decreases more when demand becomes more concentrated, while the users’ total cost for metro decreases more when there is an increase in the number of passengers). Next, the optimal transit structure and size is investigated for a mono-centric and twin cities, where users who’s origin and/or destination is not in the transit service region are assumed to drive an automobile. A sensitivity analysis performed on the cost to drive shows the optimal number of passengers served by the transit system is dependent on the cost to drive and the cost and operational characteristics of the transit mode.

The rest of this thesis is organized as follows. Previous work related to transit network design will be reviewed in Chapter 2. Chapter 3 introduces analytical equations for the users’ cost and agency’s investment considering spatially heterogeneous demand. Chapter 4 extends the previous chapters work by determining how large a transit system should be and where it should be located. Many ideas for future research are provided in Chapter 5, and Chapter 6 concludes this thesis.
CHAPTER 2
LITERATURE REVIEW

This chapter presents an overview of some of the major contributions in transit network design and operation. First, additional background on continuum approximation methods and transit network design is provided. Then, different types of transit related problems are introduced.

2.1 Continuum Approximation

The methodology to optimize network level decisions (e.g., route spacing) analytically using so called continuum models started to appear in the early 1970s (Hurdle, 1973; Newell, 1971); for a comprehensive review see Newell (1973); Clarens and Hurdle (1975); Wirasinghe et al. (1977); Chang and Schonfeld (1991); Daganzo (2005) and Ouyang et al. (2014). Continuum models allowed for network level decisions to be approximated, in most cases very accurately, by a few number of decision variables. For instance, transit routes and stops are specified in terms of their physical separations (e.g., spacing), instead of discretely determining the exact location of each route and stop. Similarly, other decisions, such as network structure, service frequency, and rolling stock, can be modeled as continuous variables. Daganzo et al. (2012) addresses the major benefits of using a model that is parsimonious, of just a few variables, to address big picture questions. This thesis will follow the continuum approximation approach to focus on the big picture question of how spatially heterogeneous passenger demand will effect transit network design and frequency setting.

2.2 Transit Network Design and Scheduling Problem (TNDSP)

Literature

Public transit design and operation requires many decisions to be made. As introduced in Chapter 1, there are considered to be five main stages of transit planning; for additional background on public transit design and operation see Vuchic (2005, 2007); Daganzo (2010a).
According to Van Nes et al. (1988), and focusing on the strategic planning steps only, these can be separated by their design and operation optimization approach: (i) analytical models; (ii) route construction models; (iii) models assigning frequencies to routes; (iv) two-stage models first performing route construction and then assigning frequency; and (v) models for simultaneously determining routes and frequencies. This thesis will focus on approach (v) where the strategic decisions of network structure and route design and tactical decision of route frequency will be determined simultaneously. This approach ensures a transit system that would provide good spatial and temporal coverage to passengers. Adopting the same notation as Guihaire and Hao (2008), this type of problem is called a transit network design and frequency setting problem (TNDFSP). This is a subproblem of the broader transit network design and scheduling problem (TNDSP) which incorporates every transit design planning step; see Figure 2.1. Furthermore, each type of transit network problem is typically an NP-hard combinatorial optimization problem (Israeli and Ceder, 1995). Therefore, the practical approach is to solve each step sequentially. While computationally this is easier, the solutions obtained sequentially may not be the global optimal solution. Still, optimizing sequentially can result in an economical and efficient transit system.

![Figure 2.1: Structure of transit network problems (Guihaire and Hao, 2008).](image)

A common method to optimize TNDSP and its subproblems consists of formulating transit related objectives and constraints as mathematical models that can be solved analytically or by using heuristics. Typical objectives include user benefit maximization, user travel time minimization, operator cost minimization, total welfare maximization, minimization of transfers, capacity maximization, and energy conservation (Fielding, 1987; Van Oudheusden et al., 1987; Black, 1995; Van Nes and Bovy, 2000; Kepaptsoglou and Karlaftis, 2009; Tirachini et al., 2010). Typical constraints include feasible ranges of variables, maximum and minimum occupancy, directness, maximum length, maximum number of routes/corridors, maximum fleet size, maximum budget (Fan and Machemehl, 2004; Zhao and Zeng, 2006; Estrada et al., 2011). Numerous solution approaches that have been proposed in the
literature will be discussed in Section 2.4.

Next, the important literature for TNDP, TNFSP, and TNDFSP will be discussed. The purpose of introducing some TNDP and TNFSP literature is to give the reader a glimpse of how these two subproblems are solved together in a TNDFSP.

2.2.1 Transit Network Design Sub-Problem (TNDP)

The first subproblem, TNDP, consists of designing the route structure (e.g., route spacing, stop spacing) and/or network structure (e.g., hub and spoke, grid, circular) only. The frequency is typically assumed to be given. Additionally, the studies discussed below argue that the total transit route length should play a larger impact on the total costs. This is due to the relative ease of adjusting frequencies as opposed to altering the route spacing. The major works of TNDP can be summarized as (i) corridor analyses and (ii) route improvement heuristics.

The optimal stop spacing along a single corridor with many-to-one passenger demand (i.e., everyone travels to the city center) which could vary by location was investigated by Vuchic and Newell (1968). Vuchic and Newell (1968) found that the optimal stop spacing was highly sensitive to access time. Later Wirasinghe and Ghoneim (1981), expanded this work to include many-to-many passenger demand. Another area of TNDP literature represents the transit network by a set of nodes and links. Mandl (1980) created an initial set of routes (connected links) and assigned passengers to them. Then, Mandl (1980) used a route selection heuristic to improve the combined agency’s and users’ cost. Finally, Van Nes and Bovy (2000) studied the relationship between objectives and stop and route spacing. This work doesn’t actually design transit networks but shows different total cost surface functions for different objectives. This can help decision makers understand the objectives effect on stop and route spacing.

2.2.2 Transit Network Frequency Setting Sub-Problem (TNFSP)

The TNFSP normally determines the optimal frequency for a given route or set of routes to minimize or maximize an objective (e.g., minimize total waiting time, maximize social benefits) and/or to ensure that no bus is over capacity. For instance, Furth and Wilson (1981) calculated optimal frequencies to maximize social benefits. Similarly, Newell (1971) determined the optimal dispatch time of transit vehicles for a continuous passenger arrival rate function that minimizes the total waiting time. This work was a major contribution in two ways: (i) it allowed for the arrival rate to vary with time; and (ii) it extended on
the continuum approximation approach to transit related problems. Another approach to
determine frequencies is to use real passenger count data, collected by workers or technology.
This data can then be analyzed to calculate the necessary frequency on each route (Ceder,
1984). Others have created models that determine the necessary frequency to maintain the
load factors under a specified maximum (Baaj and Mahmassani, 1990). Lastly, LeBlanc
(1988) used an iterative model that considered modal splitting to determine the optimal
frequencies for specified transit lines. With this model the demand was elastic and would
change with respect to the service performance.

2.2.3 Transit Network Design & Frequency Setting Sub-Problem (TNDFSP)

TNDFSP combines the network and route structure decisions of TNDP and the frequency
decisions of TNFSP. The literature discussed in this section falls into one of the following
network structures: (i) rectangular; (ii) radial; (iii) feeder and commuter; and (iv) flexible.
Basic corridor and rectangular transit structures have been covered extensively and are not
thoroughly presented here; see Guihaire and Hao (2008); Kepaptsoglou and Karlaftis (2009);
Daganzo (2010a) for comprehensive reviews. Some of the more intriguing studies relevant
to this work are presented next.

Byrne and Vuchic (1972) derived the optimal route spacing, headways, and vehicle fleet
size for parallel routes in a rectangular area, where passengers traveled to one side of the
rectangle assumed to be the CBD area. A year later, this work was expanded to allow
for each route to have different speeds (Byrne, 1976). Daganzo (2010b) presented a hybrid
transit structure for square cities where there is a central region that receives double coverage
(i.e., grid network) and a peripheral region that consists of single cover (i.e., radial lines).
This work is the basis of this thesis and will be introduced in further detail in Chapter
3. Estrada et al. (2011) continued this work by expanding the formulation to rectangular
cities and providing more detailed formulations for the expected number of transfers. Two
interesting papers, one by Chien and Schonfeld (1997) and the other Ouyang et al. (2014),
created irregular grid networks where route spacing was allowed to change over the area
(e.g., areas of high passenger density had a higher density of routes). Chien and Schonfeld
(1997) did this by dividing the city into zones with near uniform passenger demand. The
zones were constructed in such a way that the shared edge of the neighboring zones had the
same length. Then the optimal number of horizontal and vertical routes were determined
for all the zones in one vertical or horizontal section. The network structure in Ouyang et al.
(2014) was similar, but also allowed for the transit routes to branch and form local lines.
in areas of higher passenger density. Unlike Chien and Schonfeld (1997), this lessened the impact of over designing the entire vertical or horizontal route because one small portion had a much higher passenger density. Additionally, Ouyang et al. (2014) was able to formulate the problem as a continuum approximation model where the optimal solutions depended on the local conditions (e.g., passenger density) only.

Other researchers have studied the optimal spacing and headways for circular cities with radial transit routes (Byrne, 1975; Wirasinghe et al., 1977; Black, 1979) and radial and ring transit routes (Vaughan, 1986; Badia et al., 2014). Byrne (1975) found the optimal radial route positions and headways for a commuter train with many-to-one demand. Wirasinghe et al. (1977) expanded on this by allowing (i) certain areas to be served by direct commuter buses and (ii) determining feeder bus boundaries between competing modes of transit. Both Vaughan (1986) and Badia et al. (2014) network structures allowed for the addition of ring lines; however Badia et al. (2014) took it one step farther and allowed for different designs in the central and peripheral regions similar to Daganzo (2010b).

Hurdle (1973), Kuah and Perl (1988), and Chang and Schonfeld (1993a) each approached the TNDFSP for feeder buses by analyzing parallel feeder bus routes in a rectangular area. The passenger demand in this area wants to access a commuter transit station or CBD area. Hurdle (1973) presented analytical formulations for optimal spacing and dispatch times for parallel feeder routes. Kuah and Perl (1988) expands this by incorporating spatial heterogeneous demand and allowing stop spacing and line spacing to vary by location. In Chang and Schonfeld (1993a) the analysis was done for a rectangular zone, but it included an extra commuting distance to the CBD area. This allowed Chang and Schonfeld (1993a) to build an entire network of multiple feeder bus zones over the city. The zones were given, but Chang and Schonfeld (1993a) stated that they could be designed so that passenger density (or another characteristic) was uniform within each of them.

### 2.3 Passenger Demand Literature

Trip demand can be uniform or heterogeneous (in time and/or space), deterministic or stochastic, and inelastic or elastic, which corresponds to the board categories of: (i) temporal distribution; (ii) spatial distribution; (iii) uncertainty; and (iv) elasticity, respectively. The majority of literature presented thus far has focused on designing transit systems for areas with uniform (both in time and space) deterministic inelastic demand. This is the easiest for analytical optimization of network variables, but there are many more combinations of demand characteristics and their impact needs to be better understood.
2.3.1 Temporally Heterogeneous Demand

Temporally, research has focused on a single time period where it is assumed there is no temporal fluctuation. Newell (1971) presented a continuous trip demand function that varied with time in effort to find the optimal dispatching times of vehicles to minimize the total waiting time of all passengers. However, generating a continuous trip demand function can be quite complex for real transit systems, especially remembering that the strategic and tactical decisions are not made in real time. Additionally, operators and passengers prefer systems that do not change drastically in a short amount of time. Another approach that can be implemented in the strategic planning stage, is considering multiple time periods throughout the day (Chang and Schonfeld, 1991, 1993a,b). With this approach, the day could be split into multiple time periods (e.g., morning peak, midday, evening peak, late night) and some decisions could change to better serve the passengers. For instance, the network structure may not change, but the frequency can be adjusted to the different levels of passenger demand.

2.3.2 Spatially Heterogeneous Demand

There has also been significant research for spatially heterogeneous demand. Byrne (1975), Black (1979), Vaughan (1986), and Badia et al. (2014) investigated spatially heterogeneous demand in a ring and radial network, where the demand varied in the radial direction only. Due to the ring and radial network structure, the demand function could be directly used in their formulations to determine optimal structure and frequency. Results similar to those for ring and radial networks have not been produced for rectangular transit service regions. Byrne (1976) proposed formulations for rectangular cities with spatially heterogeneous demand, however the only results were for a uniform demand scenario.

Others, have allowed for the transit service region to be split up spatially into homogeneous zones that each has different trip generation and termination rates (Chien and Schonfeld, 1997; Chien and Spasovic, 2001; Chien et al., 2003). The zonal demands could be calculated by integrating a continuous spatial demand density function over the boundaries of the zone. Using zonal demands instead of the continuous spatial demand density function allows for easier calculations of cost metrics (e.g., access time, infrastructure length). Ouyang et al. (2014), was able to overcome the computational difficulties of using a continuous spatial demand density and find the optimal route spacing by using continuum approximation techniques (i.e., their formulations of users’ cost and agency’s investment depended on local conditions and were optimized locally).
It is noteworthy to state that Hurdle (1973) and Clarens and Hurdle (1975) presented work in which the demand was allowed to be heterogeneous both temporally and spatially. However, Hurdle (1973) was not able to produce any examples, most likely due to the computational power at the time. Clarens and Hurdle (1975) was able to illustrate an example by designing the transit network structure for Berkeley, California during the peak time period and using a manual heuristic. Both of these studies, and the vehicle dispatch scheduling work by Newell (1971), were the basis for research exploring temporal and spatial heterogeneity.

### 2.3.3 Stochastic Demand

There has not been much research on stochastic demand and transit network design; see Li (2009) and Watling and Cantarella (2013) for comprehensive reviews. It is important to analyze the impact stochastic demand has on the design and operation of transit systems. Chapman et al. (1976) pointed out that stochastic factors may introduce some irregularity into the headway distribution. This could generally cause passengers to experience longer waiting times on average. Yan et al. (2006) contributed the (i) link-based and (ii) path-based heuristics algorithms to solve an inter-city bus routing and timetable design problem under stochastic demand. Yan et al. (2008) also looked at airline scheduling with stochastic demand. Recently, Lo et al. (2013) and An and Lo (2014) developed and implemented a method based on service reliability which required the agency to input a desired service reliability factor. The designed transit networks ensured that a portion of stochastic demand, determined by the service reliability factor, would be covered by regular transit operations and the remaining stochastic demand would be covered by additional flexible service (e.g., on-demand shuttle bus) when needed. Both Lo et al. (2013) and An and Lo (2014) found cost savings compared to when the regular transit was designed to never, or almost never, reach capacity. The regular transit service and the additional flexible service were designed sequentially with a two-stage stochastic program. This method, based on service reliability, allows for more complex network designs to be handled in a stochastic demand environment.

### 2.3.4 Elastic Demand

As mentioned earlier, most transit planning research has focused on inelastic demand; however transit system demand has a degree of elasticity in that an efficient system attracts more modal share. Studies that focused on elastic demand include performance based models, utility based competitive mode models, and proportional travel-time reduction models.
The basic models consider the demand to be a function of the transit system’s performance (e.g., waiting time, access time, in vehicle travel time, and the fare); see Oldfield and Bly (1988); Chang and Schonfeld (1991); Chien and Spasovic (2001); Chen et al. (2014). More complex models consider the transit system’s performance and competition between alternative modes of travel (e.g., car, bike, walking), while some even allow the trip to not occur (Webster and Oldfield, 1972; Kocur and Hendrickson, 1982; LeBlanc, 1988; Oum, 1989; Imam, 1998; Ferrari, 1997, 1999; Lee and Vuchic, 2005; Fan and Machemehl, 2006a; Cipriani et al., 2006, 2012a). Kocur and Hendrickson (1982) adopted a linear modal split model; Ferrari (1997), Ferrari (1999), and Lee and Vuchic (2005) used a logit modal split model; and Imam (1998) used a log linear modal split model; see Oum (1989) for a comprehensive review of modal split models. These modal split models calculate the relative utility (travel time and other costs) of each mode of travel, and use it to determine the probability of a trip occurring on each mode. One of the challenges with this method is calculating the travel time and other costs for competing modes that interact (e.g., as people switch from driving to transit the travel time and other costs for driving are lowered). One solution, presented by Cipriani et al. (2006) and later used in Beltran et al. (2009) and Cipriani et al. (2012b) was to use a linear function to relate the travel time and other costs between modes. This allows an increase in transit demand to generate a proportional reduction of travel time and costs for other modes.

2.4 Solution Methods

To truly appreciate the complexities of transit related problems one must study the solution methods. Considering how many types of transit related problems there are it should not be surprising that there exists numerous solution methods; see Fan and Machemehl (2004); Kepaptsoglou and Karlaftis (2009) for comprehensive overviews of different methodological approaches to solving transit network design programs. Here, a brief introduction is given to familiarize the reader with the terminology and solution methods for TNDP, TNFSP, and TNDFSP. As discussed previously, a common approach to model transit related problems is to formulate a mathematical program with objectives and constraints. Solutions are typically found using the following approaches: (i) analytical; (ii) heuristic; (iii) meta-heuristic; or (iv) simulation. An example of an analytical solution approach can be found in Clarens and Hurdle (1975), as well as the majority of older literature.

There are multiple heuristics approaches, such as genetic algorithms (Pattnaik et al., 1998; Chien et al., 2001; Bielli et al., 2002; Fan and Machemehl, 2006a), and hybrid heuristics (Sil-
man et al., 1974; Baaj and Mahmassani, 1995; Cipriani et al., 2006; Mauttone and Urquhart, 2009; Cipriani et al., 2012a,b). Hybrid heuristics generally use route generation and improvement algorithms to create the transit network. Baaj and Mahmassani (1995) generated and improved routes using a so-called route generation algorithm and route improvement algorithm; while Mauttone and Urquhart (2009) generated and improved routes with a so-called pair insertion algorithm. This algorithm is superior when there is a constraint that a predefined portion of the demand is covered by the routes generated during each iteration of the algorithm. Cipriani et al. (2006, 2012a,b) used a similar approach to generate initial routes, but then used a genetic algorithm to improve the routes. Similarly, there are numerous meta-heuristic approaches such as simulated annealing (Robusté et al., 1990; Fan and Machemehl, 2006b), tabu search (Fan and Machemehl, 2008), and hybrids consisting of multiple meta-heuristics approaches (Zhao and Zeng, 2008). Simulation has had a limited use in transit network design problems. One area that uses simulation is dial-a-ride (DRT) and other similar on-call transit systems. Typically, simulation is used to generate random passenger calls and then the performance of the system is recorded.
CHAPTER 3

TRANSIT NETWORK DESIGN WITH SPATIALLY HETEROGENEOUS DEMAND

Trip origin destination distributions of an urban area are characteristically not uniform. Areas of higher interest (e.g., a central business area, a university, a commercial area) will attract a large percentage of the total transit trips. Furthermore, transit agencies benefit more by providing transit in these high demand areas, as opposed to low demand areas, due to economies of scale. Therefore, when designing the structure of a transit network and its operation, consideration of spatially heterogeneous demand is vital. This chapter strives towards this by developing analytical formulations for the generalized cost of the system’s passengers and to its operating agency in terms of key parameters and decision variables based on Daganzo (2010b) hybrid transit structure, while using zone-to-zone demand calculated from a continuous demand density function, based on Ouyang et al. (2014), to capture the spatial heterogeneity of the demand. This results in a realistic mathematical representation of a city while at the same time, using continuous variables conveniently keeps the number of decision variables low.

The rest this chapter is organized as follows. First, the important characteristics, definitions, and parameters are introduced in Section 3.1.1. Next, Section 3.1.2 presents analytical formulations for the agency investment and passengers cost. Section 3.1.3 presents a constrained optimization model that can optimize the systems decisions variables. Section 3.2 presents the optimal network structure and frequency for two distinct spatial demand distributions. Finally, 3.3 concludes this chapter.

3.1 Formulation

3.1.1 Transit System, Vehicle, and Passenger Characteristics

A short summary of the hybrid network structure in Daganzo (2010b) is provided in the following paragraphs to give background as well as to introduce notation that will be used for the rest of this paper. The transit service region is a square with sides of \( D \) and consist
of a peripheral region and a square central region with sides of \( d \leq D \); see Figure 3.1 for illustration and explanation of N-S and E-W hemispheres. Within this transit service region the street layout is of a grid pattern. The central region is covered by two perpendicular transit lines (double coverage), while the peripheral region is only covered by a single transit line (single coverage). The percentage of the service area that has double coverage is \( \alpha^2 \), where \( \alpha = d/D \).

![Figure 3.1: Hybrid transit structure (Daganzo, 2010b).](image)

The transit agency will operate transit vehicles (e.g., buses, subway cars, trains) throughout the service region with a stop/line spacing of \( s \) and headway of \( H \), but this headway is only guaranteed in the central region. In the peripheral region, a single transit line may split, to keep the stop spacing constant for the whole service region, and form two, or more, branches, each with a larger headway; see Figure 3.2 for an example. In this example, transit line Aa and Ab would each have a headway of \( 2H \) and be coordinated so that in the central region the headway is \( H \). Transit vehicles have a capacity of \( C \) and a cruising speed of \( v_t \), which includes: (i) stops due to traffic and pedestrian interaction, and (ii) a lost time per stop, \( \tau \), due to door operation, deceleration, and acceleration. Finally, the fixed and variable
costs the transit agency will incur by creating and operating a transit system are: (i) transit infrastructure, which costs $L$ per unit length; (ii) vehicle ownerships, which costs $M$ per vehicle; and (iii) vehicle maintenance and operation, which costs $V$ per vehicle per unit distance per unit time.

A passenger’s trip consists of: (i) walking at a speed $v_w$ to the nearest stop, (ii) waiting for a bus to arrive, (iii) traveling as directly, and with the least transfers, as possible, (iv) walking to their destination, again at speed $v_w$, from the nearest stop. Whenever there is a tie among routes or stops it is assumed passengers randomly choose one of the options. For each passenger who boards a vehicle at a stop, the vehicle, and everyone on board, experiences $\tau'$ of additional time added to their trip. There is no additional time for alighting passengers, since it is assumed that alighting occurs during boarding and requires less time than boarding.

Unlike, Daganzo (2010b) and Estrada et al. (2011) this paper does not assume a uniform demand. Instead, a continuous demand density function is borrowed from Ouyang et al. (2014) to describe the transit trip demand from origin ($x_1, y_1$) to destination ($x_2, y_2$):

$$
\delta(x_1, y_1, x_2, y_2) = \prod_{i=1}^{2} \left( a_1 + a_2 \sum_{j=1}^{2} \exp \left[ - \left( a_{3j}x_i - b_{ij} \right)^2 - \left( a_{4j}y_i - \bar{b}_{ij} \right)^2 \right] \right),
$$

(3.1)

where the additional parameters $a_1, a_2, a_{31}, a_{41}, b_{11}, b_{12}, b_{21}, b_{22}, \bar{b}_{11}, \bar{b}_{12}, \bar{b}_{21}$, and $\bar{b}_{22}$ will be addressed in Section 3.2. To capture the spatial heterogeneity of the demand, we do not use this function directly in the formulations; rather, the passenger trips are first separated...
into four distinct trip types based on origin and destination regions (zone-to-zone); (i) $\lambda_{c-c}$: central region to central region; (ii) $\lambda_{c-p}$: central region to peripheral region; (iii) $\lambda_{p-c}$: peripheral region to central region; and (iv) $\lambda_{p-p}$: peripheral region to peripheral region. The total number of passengers for each trip type is calculated as follows:

$$\lambda_{c-c} = \int_{y_2=LB}^{UB} \int_{x_2=LB}^{UB} \int_{y_1=LB}^{UB} \int_{x_1=LB}^{UB} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2;$$  \hspace{1cm} (3.2)

$$\lambda_{c-p} = \int_{y_2=0}^{UB} \int_{x_2=0}^{ UB} \int_{y_1=LB}^{UB} \int_{x_1=LB}^{UB} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c};$$  \hspace{1cm} (3.3)

$$\lambda_{p-c} = \int_{y_2=LB}^{UB} \int_{x_2=LB}^{UB} \int_{y_1=0}^{UB} \int_{x_1=0}^{UB} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c};$$  \hspace{1cm} (3.4)

$$\lambda_{p-p} = \int_{y_2=0}^{UB} \int_{x_2=0}^{UB} \int_{y_1=0}^{UB} \int_{x_1=0}^{UB} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c} - \lambda_{c-p} - \lambda_{p-c}, \text{ where}$$  \hspace{1cm} (3.5)

$$UB = \frac{D}{2} \left(1 + \alpha\right), \text{ and}$$  \hspace{1cm} (3.6)

$$LB = \frac{D}{2} \left(1 - \alpha\right).$$  \hspace{1cm} (3.7)

After calculating each $\lambda$, to simplify the derivation of cost formulas, it is assumed that the origins and destinations are uniformly distributed within their respective origin and destination areas. In summary, this approach takes the continuous demand density function and transforms it into a set of discrete zone-to-zone demands which describe the general shape of the continuous demand density function without overly complicating the formulations. Chien and Spasovic (2001) and Chien et al. (2003) also used zone demands, but only considered trip generation in a region, whereas this thesis considers each origin-destination region pair as a zone-to-zone demand. Note, Byrne (1975), Black (1979), Vaughan (1986), and Badia et al. (2014) were able to directly use a centripetal demand, since their network structure was a ring and radial network. Furthermore, averaging the demand over large areas may pool and cancel out some of the local errors in the continuous density function generated during data collection and/or caused when a city’s boundaries force a non-continuous shape that must be smoothened (e.g., when city’s boundary and demand abruptly end due to a body of water). Finally, $PHF$ and $T_{pax}$ are defined as the peak hour factor and the total number of passengers on the transit system during one hour (not during the peak hour), respectively.
3.1.2 Agency Investment and User Cost

In this section, the generalized cost for the system’s passengers and its operating agency during one hour of operation is derived in terms of the key system parameters and variables presented in the previous section. First, it is important to present the four key decisions that the agency is allowed to make: (i) stop/line spacing, \( s \); (ii) headway in the central region, \( H \); (iii) the size of the central region, \( d \), by controlling \( \alpha = d/D \); and (iv) the vehicle mode (i.e., the vehicle and infrastructure type) along with any operational controls. The vehicle mode decisions should have been captured in \( v_t \), \( \tau \), \( C \), \( $L \), \( $V \), \( $M \), and \( \tau' \) by including an extra subscript. However, for simplicity and since each mode will be optimized individually and then compared, this additional subscript was dropped. Finally, we present three metrics to evaluate the performance of the transit system: (i) the commercial speed, \( v_c \); (ii) the expected number of transfers per passenger, \( e_t \); and (iii) the maximum vehicle occupancy, \( O \). The commercial speed is a measure of how quickly in-vehicle passengers can move throughout the transit system, including stops along the way, and can be used to compare transit to driving. The expected number of transfers per passenger illustrates how direct the transit system is for all passengers. The formulas for the commercial speed and expected number of transfers are as follows; they are derived in Appendix A.1:

\[
\frac{1}{v_c} = \frac{1}{v_t} + \frac{\tau}{s} + \frac{\tau' PHFsH}{D^2 (3\alpha - \alpha^2)} \left( \lambda_{c-c} + \lambda_{c-p} + \lambda_{p-c} + 0.5 (e_t + 1) \lambda_{p-p} \right), \text{ where (3.8)}
\]

\[
e_t = 1 + \frac{1}{2} \left( \frac{\lambda_{p-p}}{T_{pax}} \right). \text{ (3.9)}
\]

The maximum vehicle occupancy allows the agency to determine how close their system is from capacity. When the expected maximum passenger occupancy is close to the vehicles capacity, the system tend to suffer more from operational uncertainties (e.g., turning away passengers, bus bunching), which degrade the systems performance and the users experience. The formulation of vehicle occupancy will be introduced in Section 3.1.3 with other constraints.

As previously mentioned, the agency will incur a combination of fixed and variable costs for providing service to the passengers. Given the hybrid transit structure, the agency costs can be calculated by determining: (i) the total infrastructure length, \( L \); (ii) the total vehicular distance traveled per hour of operation, \( V \); and (iii) the total vehicle hours traveled during the peak hour, \( M \). The formulas for each of these terms are presented next, and they are derived in Appendix A.1:

\[
L = \frac{D^2}{s} \left( 1 + \alpha^2 \right); \text{ (3.10)}
\]
\[ V = \frac{2D^2}{sH} (3\alpha - \alpha^2) \]; and
\[ M = \frac{V}{v_c}. \]  

(3.11)  
(3.12)

Passengers, on the other hand, experience the system differently than the agency. The agency is more focused on costs and passengers are more interested in the time it takes to complete their trip. This is especially true when comparing transit trips to automobile trips. Even if the fare is much lower than the cost to own and operate an automobile, as long as the total travel time on transit is much longer than that via automobile the transit system isn’t likely to attract and retain many passengers. From the passengers perspective the least desirable part of transit is walking to and from the stop and waiting for a transit vehicle to arrive. Once on board, passengers can make more use of their time by doing something productive (e.g., reading) that cannot be completed while driving. The model presented in this thesis does not weight any part of a passengers trip (e.g., in-vehicle versus waiting) differently, although this could be a straightforward extension. More specifically, a passengers trip on transit include: (i) walking to the transit stop, (ii) waiting for the transit vehicle to arrive, (iii) in vehicle travel, (iv) if needed, transferring to another route and waiting for the transfer vehicle to arrive, and (v) walking to the final destination. Combining (i) and (v) as well as (ii) and (iv), the formulations for the expected (i) access time, \( A \); (ii) waiting time, \( W \); and (iii) in vehicle travel time, \( T \), per passenger are as follows; they are derived in Appendix A.1:

\[ A = \frac{s}{v_w}; \]  

(3.13)  

\[ W = \frac{H}{2T_{pax}} \left[ 2\lambda_{c-c} + \lambda_{p-c} + \lambda_{c-p} \frac{\lambda_{p-p}}{2} ight. \\
+ \left. 2 \frac{(2\lambda_{p-p} + \lambda_{p-c} + \lambda_{c-p})}{3} \left( \frac{1}{\alpha (\alpha + 1)} + 1 \right) \right]; \]  

(3.14)  

\[ T = \frac{E}{v_c}, \text{ where} \]  

(3.15)  

\[ E = \frac{\lambda_{p-p}}{T_{pax}} \left[ \frac{D}{2} (2 - 3\alpha + \alpha^3) + \frac{11\alpha D}{12} \right] + \frac{\lambda_{c-c}}{T_{pax}} \left[ \frac{2\alpha D}{3} \right] \\
+ \frac{\lambda_{c-p} + \lambda_{p-c}}{T_{pax}} \left[ \frac{D}{4} (2 - 3\alpha + \alpha^3) + \frac{\alpha D}{12} \left( 11 - \alpha^2 \right) \right]. \]  

(3.16)
3.1.3 Mathematical Model

To optimize the transit system’s design the sum of the agency’s investment and generalized passengers’ costs is minimized. However, the previous formulations for the agency’s investment are in units of money, while the passengers’ costs are in units of time per passenger. To covert the units of money to units of time per passenger the agency’s investments are divided by the passengers’ value of time \( \mu \), and by the total number of passengers, \( T_{pax} \). Daganzo (2010b) suggests \( \mu \) should be similar to the prevailing wage. Additionally, each transfer that a passenger is required to make is inconvenient and increases their trip time. To consider the impact transferring has on a passenger a fixed penalty for transfers \( \delta_t \) is included in terms of an equivalent walking distance. The total agency investment and passenger cost are represented by \( Z_A \) and \( Z_U \), respectively. Therefore, the objective is:

\[
\min_{s,H,\alpha} \{ Z = Z_A + Z_U \}, \quad \text{where}
\]

\[
Z_A = \left[ S_L L + S_V V + S_M M \right] / (\mu T_{pax}), \quad \text{and}
\]

\[
Z_U = A + W + T + \frac{\delta_t e_t}{v_w},
\]

where the last term in \( Z_U \) is the transferring penalty. Objective (3.17) is subject to the following constraints:

\[
s \geq s_{\min}; \quad \text{(3.18)}
\]

\[
H \geq H_{\min}; \quad \text{(3.19)}
\]

\[
s/D \leq \alpha \leq 1; \quad \text{and} \quad \text{(3.20)}
\]

\[
O \leq C, \quad \text{where} \quad \text{(3.21)}
\]

\[
O = \frac{PHF s H}{\alpha D} \max \left\{ \frac{\max \left\{ \lambda_{c-p} + \lambda_{p-c} \right\} + \lambda_{p-p}}{2}, \quad \text{max} \left\{ \frac{\lambda_{c-p} + \lambda_{p-c} + \lambda_{p-p}}{2} - \frac{\lambda_{p-p}}{8} \right\} \right\}.
\]

The first term inside the first max operator is the maximum vehicle occupancy on the critical links entering/leaving the peripheral region, and the second term is the maximum vehicle occupancy on the critical links in the central region. Constraint (3.18) enforces that the stop/line spacing must be greater than a certain lower bound \( s_{\min} \), which could be defined by the city’s street spacing. Constraint (3.19) guarantees that the transit agency operates with a headway larger than \( H_{\min} \). The minimum headway could change for different transit modes especially if one has its own right of way and another shares the roadway with other
vehicle modes. Constraint (3.20) sets the upper and lower bounds on $\alpha$. Finally, Constraint (3.21) ensures that the vehicle occupancy is less than the vehicles capacity. There are many other constraints that could be applied as described in Chapter 1. Each could constrain the problem differently, and potentially produce different optimal network designs and operating frequencies. Next, two distinct spatial demand distributions are introduced and the optimal network design and operating frequencies are determined.

3.2 Numerical Results

In this section, the optimal hybrid transit network design was obtained for a square city with $D = 10$ [km] and for two distinct spatial demand distributions: (i) a uniform city, to make comparisons to Daganzo (2010b); and (ii) a mono-centric city, where both the origins and destinations are clustered in the city center. To make further comparisons both spatial demand distributions were optimized for a low demand scenario (20,000 total passengers/hr) and a high demand scenario (80,000 total passengers/hr). Figure 3.3 illustrates the marginal distributions of trip origins and destinations for the low and high demand scenarios for the mono-centric city. Table 3.1, which is from Ouyang et al. (2014), provides the additional mono-centric city spatial demand density function parameters, previously mentioned and shown in Equation (3.1), for a base demand of 10,000 passengers/hr. To reach the low and high demand scenarios Equation (3.1) was multiplied by 2 and 8, respectively. The uniform demand density function was set to have a constant trip generation density that would equal a total of 20,000 passengers/hr and 80,000 passengers/hr over the whole area.

![Figure 3.3: Marginal distributions of trip origins and destinations.](image-url)
### Table 3.1: Demand distributions parameters.

<table>
<thead>
<tr>
<th>Case</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
</tr>
<tr>
<td>Mono-centric</td>
<td>0.0016</td>
</tr>
<tr>
<td>Twin</td>
<td>0.0030</td>
</tr>
</tbody>
</table>

### Table 3.2: Modal characteristics.

<table>
<thead>
<tr>
<th>Modes</th>
<th>$C$ [pax]</th>
<th>$\tau$ [s]</th>
<th>$\tau'$ [s/pax]</th>
<th>$v_t$ [km h]</th>
<th>$\delta_t$ [km]</th>
<th>$S_L$ [$/km/hr$]</th>
<th>$S_V$ [$$/veh/km]</th>
<th>$S_M$ [$$/veh/hr]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>120</td>
<td>30</td>
<td>1</td>
<td>25</td>
<td>0.03</td>
<td>9</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>BRT</td>
<td>150</td>
<td>30</td>
<td>1</td>
<td>40</td>
<td>0.03</td>
<td>90</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>Metro</td>
<td>1000</td>
<td>45</td>
<td>0</td>
<td>60</td>
<td>0.1</td>
<td>900</td>
<td>6</td>
<td>120</td>
</tr>
</tbody>
</table>
The other modal and passenger parameters that were used are shown in Table 3.2. Additionally, it was assumed that passengers can walk at a speed of \( v_w = 2 \text{[km/hr]} \), and value their time at \( \mu = 20 \text{[$/hr]} \). Finally, \( s_{min} = 0.15 \text{km} \), \( H_{min} = 1 \text{ min} \), and the PHF was assumed to be equal to 2.5. Given these parameters, the model described by Objective (3.17) and Constraints (3.18) - (3.21) was programmed into Matlab and optimized on a desktop computer with a 2.8 GHz CPU and 8 GB memory. The solver, fminsearchcon (D’Errico, 2012), is designed for constrained optimization problems by transforming a bounded constrained problem into an unconstrained one. The unconstrained problem was then solved by fminsearch, part of the Optimization Toolbox provided by Matlab and based on the Nelder-Mead simplex algorithm (Matlab, 2014). Like Daganzo (2010b), we compute the optimal design and operating frequency for each mode type and then make comparisons. These results are shown in Table 3.1.

Comparing the low and high uniform demand scenarios with Daganzo (2010b) there are some noticeable differences in the optimal \( \alpha \), \( s \), and \( H \), for each of the three modes. However, all of the total costs, \( Z \), presented in this thesis are lower than their corresponding scenario in Daganzo (2010b). Additionally, when Daganzo (2010b) optimal values are used to calculate the user cost and agency investment from this papers formulations, they match closely with Daganzo’s results. A potential reason for the difference in optimal solutions is that this paper uses continuous variables to represent \( \alpha \), spacing, and headway, whereas Daganzo (2010b) only uses discrete values.

Tables 3.4 and 3.5 allow for additional comparisons between demand levels and spatial demand patterns. These tables provide two main topics for discussion: (i) how an increase in the total number of passengers, and (ii) how concentrating demand from uniform to mono-centric city (for the same total demand level) effect the network structure, operating frequency, and other important metrics. Table 3.4 shows the percent change in each metric when the spatial demand remains the same and passengers are added to the system. Table 3.5 shows the percent change in each metric when the number of passengers remains the same but the spatial demand becomes more concentrated. In the next section, rows of these tables are referenced and the corresponding percent changes are displayed inside parenthesis.

Starting from the top of Tables 3.4 and 3.5, it is easy to see that increasing/concentrating passengers affects the size of the central region, spacing, and the headway. A few important insights can be brought forth from these relationships: First, higher passenger total demand decreases the headway (-55%, -53%, -14%, -65%, -53%, and -26%) much more than concentrating them (-6%, -3%, 15%, -26%, -4%, and -1%). These large reductions, due to increasing passenger demand, increases the number of transit vehicles significantly as well (218%, 232%, 183%, 284%, 231%, and 199%). This knowledge suggests newer transit sys-
### Table 3.3: Spatially heterogeneous demand results.

<table>
<thead>
<tr>
<th>Demand Scenario</th>
<th>Uniform</th>
<th>Mono-centric</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{pax}$ [pax/hr]</td>
<td>20,000</td>
<td>80,000</td>
</tr>
<tr>
<td>Mode</td>
<td>Bus</td>
<td>BRT</td>
</tr>
<tr>
<td>Decision Variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.724</td>
<td>0.652</td>
</tr>
<tr>
<td>$s$ [km]</td>
<td>0.457</td>
<td>0.515</td>
</tr>
<tr>
<td>$H$ [min]</td>
<td>4.43</td>
<td>3.54</td>
</tr>
<tr>
<td>Demand Metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_{c-c}$ [pax/hr]</td>
<td>5500</td>
<td>3615</td>
</tr>
<tr>
<td>$\lambda_{c-p}$ [pax/hr]</td>
<td>4988</td>
<td>4888</td>
</tr>
<tr>
<td>$\lambda_{p-c}$ [pax/hr]</td>
<td>4988</td>
<td>4888</td>
</tr>
<tr>
<td>$\lambda_{p-p}$ [pax/hr]</td>
<td>4524</td>
<td>6609</td>
</tr>
<tr>
<td>Performance Metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$v_c$ [km/hr]</td>
<td>16.4</td>
<td>22.7</td>
</tr>
<tr>
<td>$e_t$ [tran./pax]</td>
<td>1.11</td>
<td>1.17</td>
</tr>
<tr>
<td>$O$ [pax]</td>
<td>120</td>
<td>127</td>
</tr>
<tr>
<td>Agency Metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L$ [km]</td>
<td>333</td>
<td>277</td>
</tr>
<tr>
<td>$V$ [veh km/hr]</td>
<td>9765</td>
<td>10083</td>
</tr>
<tr>
<td>$M$ [veh]</td>
<td>597</td>
<td>444</td>
</tr>
<tr>
<td>Users Metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$ [min/pax]</td>
<td>13.7</td>
<td>15.4</td>
</tr>
<tr>
<td>$W$ [min/pax]</td>
<td>5.1</td>
<td>4.4</td>
</tr>
<tr>
<td>$T$ [min/pax]</td>
<td>23.7</td>
<td>17.1</td>
</tr>
<tr>
<td>Costs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_A$ [min/pax]</td>
<td>7.0</td>
<td>9.4</td>
</tr>
<tr>
<td>$Z_U$ [min/pax]</td>
<td>43.6</td>
<td>38.1</td>
</tr>
<tr>
<td>$Z$ [min/pax]</td>
<td>50.5</td>
<td>47.5</td>
</tr>
<tr>
<td>Solution time [sec]</td>
<td>43</td>
<td>28</td>
</tr>
</tbody>
</table>

23
Table 3.4: Percent change in metrics between low and high demand.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Uniform</th>
<th>Mono-Centric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>Bus</td>
<td>BRT</td>
</tr>
<tr>
<td>$\alpha$ [km]</td>
<td>18%</td>
<td>16%</td>
</tr>
<tr>
<td>$s$ [km]</td>
<td>-16%</td>
<td>-21%</td>
</tr>
<tr>
<td>$H$ [min]</td>
<td>-55%</td>
<td>-53%</td>
</tr>
<tr>
<td>$v_c$ [km/hr]</td>
<td>-7%</td>
<td>-10%</td>
</tr>
<tr>
<td>$e_t$ [transfers/pax]</td>
<td>-7%</td>
<td>-6%</td>
</tr>
<tr>
<td>$O$ [pax]</td>
<td>0%</td>
<td>18%</td>
</tr>
<tr>
<td>$L$ [km]</td>
<td>36%</td>
<td>39%</td>
</tr>
<tr>
<td>$V$ [veh km/hr]</td>
<td>196%</td>
<td>197%</td>
</tr>
<tr>
<td>$M$ [veh]</td>
<td>218%</td>
<td>232%</td>
</tr>
<tr>
<td>$A$ [min/pax]</td>
<td>-16%</td>
<td>-21%</td>
</tr>
<tr>
<td>$W$ [min/pax]</td>
<td>-59%</td>
<td>-58%</td>
</tr>
<tr>
<td>$T$ [min/pax]</td>
<td>9%</td>
<td>12%</td>
</tr>
<tr>
<td>$e_t$ [min/pax]</td>
<td>-7%</td>
<td>-6%</td>
</tr>
<tr>
<td>$Z_A$ [min/pax]</td>
<td>-26%</td>
<td>-39%</td>
</tr>
<tr>
<td>$Z_U$ [min/pax]</td>
<td>-7%</td>
<td>-10%</td>
</tr>
<tr>
<td>$Z$ [min/pax]</td>
<td>-10%</td>
<td>-16%</td>
</tr>
</tbody>
</table>

Table 3.5: Percent change in metrics between uniform and mono-centric demand.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>Bus</td>
<td>BRT</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>-29%</td>
<td>-25%</td>
</tr>
<tr>
<td>$s$ [km]</td>
<td>-10%</td>
<td>-9%</td>
</tr>
<tr>
<td>$H$ [min]</td>
<td>-6%</td>
<td>-3%</td>
</tr>
<tr>
<td>$v_c$ [km/hr]</td>
<td>-3%</td>
<td>-4%</td>
</tr>
<tr>
<td>$e_t$ [transfers/pax]</td>
<td>-6%</td>
<td>-9%</td>
</tr>
<tr>
<td>$O$ [pax]</td>
<td>-5%</td>
<td>-11%</td>
</tr>
<tr>
<td>$L$ [km]</td>
<td>-8%</td>
<td>-4%</td>
</tr>
<tr>
<td>$V$ [veh km/hr]</td>
<td>-9%</td>
<td>-9%</td>
</tr>
<tr>
<td>$M$ [veh]</td>
<td>-6%</td>
<td>-5%</td>
</tr>
<tr>
<td>$A$ [min/pax]</td>
<td>-10%</td>
<td>-9%</td>
</tr>
<tr>
<td>$W$ [min/pax]</td>
<td>-3%</td>
<td>-5%</td>
</tr>
<tr>
<td>$T$ [min/pax]</td>
<td>-21%</td>
<td>-20%</td>
</tr>
<tr>
<td>$e_t$ [min/pax]</td>
<td>-6%</td>
<td>-9%</td>
</tr>
<tr>
<td>$Z_A$ [min/pax]</td>
<td>-7%</td>
<td>-6%</td>
</tr>
<tr>
<td>$Z_U$ [min/pax]</td>
<td>-15%</td>
<td>-14%</td>
</tr>
<tr>
<td>$Z$ [min/pax]</td>
<td>-14%</td>
<td>-12%</td>
</tr>
</tbody>
</table>
tems should have some excess vehicles if they predict their ridership could drastically increase after initial startup. Second, for a city with higher passenger demand, capital improvement projects should be aimed at increasing the size central area (18%, 16%, 43%, 8%, 7%, and 15%) and decreasing the spacing (-16%, -21%, -33%, -16%, -23%, and -36%). However, for a city where the demand becomes concentrated (e.g., a city with a booming CBD area), the spacing (-10%, -9%, -5%, -9%, -11%, and -9%) still reduces, but the size of the central region (-29%, -25%, -3%, -35%, -31%, and -22%) also reduces. This may seem odd since, the CBD area is flourishing but the central transit region is decreasing. However, the result should be reasonable — as more of the passengers with origins in the peripheral region want to go to a smaller area in the central region, the central region size should be reduced to improve those trips. Finally, concentrating the demand decreases the required number of transit vehicles (-6%, -5%, -9%, 14%, -5%, and -4%), except for the high demand bus scenario due to capacity constraints, while still reducing the headway (-6%, -3%, 15%, -26%, -4%, and -1%).

Another interesting note is that $v_c$ decreases when passenger demand is either increased or concentrated. A possible explanation for this is in both situations the number of boardings, which predominately occurs in the central region, increases faster than the total vehicle-km/hr. Since, $v_c$ is proportional to the inverse of the number of boardings divided by the total vehicle-km/hr, the commercial speed decreases. On a similar note, $T$ increases (9%, 12%, 11%, 7%, 13%, and 20%) when more passengers are added and decreases (-21%, -20%, -25%, -22%, -19%, and -20%) when passengers are concentrated, even with $v_c$ decreasing.

The impact of the capacity constraint should not be overlooked. The capacity constraint is binding in the bus low and high uniform demand scenarios, BRT high uniform demand, and bus and BRT high mono-centric demand scenarios. The occupancy constraint causes a reduction in headway (-5%, -11%, -4%, 0%, 0%, and -11%), when the number of passengers are increased, so that Constraint (3.21) remains satisfied. Similarly, the spacing follows the same trend. Additionally, from Table 3.5 the percent change in $O$ (-5%, -11%, -4%, 0%, 0%, and -11%) shows that, when the capacity constraint is not binding concentrating demand can lower the critical occupancy for both the uniform and mono-centric demand distributions.

Next, the impacts on $Z_A$ and $Z_U$ are discussed when demand increases or concentrates. From the agency’s perspective, increasing the number of passengers reduces $Z_A$ (-26%, -39%, -52%, -11%, -40%, and -53%) significantly more than concentrating passengers (-7%, -6%, 0%, 12%, -8%, and -4%). The formulation of $Z_A$ has the total number of passengers in the denominator so this is not surprising. $Z_U$ has a more unique trend. For the bus and BRT modes, concentrating passengers lowers $Z_U$ (-15%, -14%, -18%, and -15%) more than increasing the number of passengers (-7%, -10%, -11%, and -11%). For the metro, increasing the number of passengers decreases $Z_U$ (-21% and -22%) more than concentrating passengers.
(-12% and -13%). The explanation of the latter could be that when demand is concentrated, there is significantly fewer passengers in the peripheral. However, since every passenger is still required to be served, there is a large amount of infrastructure and vehicles that operate in the peripheral – but they are almost unneeded. Ideally, our formulation should allow the metro to serve the central area and let the peripheral passengers access the system or their destination by other means. For the bus and BRT modes, concentrating passengers is better. This is because both bus and BRT still have relatively low infrastructure, vehicle, and operating costs allowing for the central passengers to dictate the low headway and spacing.

Finally, increasing the number of passengers decreases the total cost, $Z$, more for the BRT and metro modes, while concentrating demand decreases $Z$ more for the bus mode. The explanation of $Z_U$ in the previous paragraph can apply again, but more importantly we shall note the large decreases in $Z_A$ when the number of passengers increases.

The results presented here are only valid for the range of demand levels and spatial demand patterns discussed. More extensive analysis needs to be performed to determine if the current trends continue. For instance one would expect once the demand reaches a certain level, concentrating demand will provide more benefits to the metro users (i.e., a lower $Z_U$) than increasing the number of passengers.

### 3.3 Conclusion

In this chapter, analytical formulas of the generalized cost for the system’s passengers and its operating agency were develop in terms of key parameters and decision variables. From a continuous demand density function the four zone-to-zone demands were calculated to capture spatial heterogeneity of the demand.

Numerical results for the design and operating frequency under low and high demand in uniform and mono-centric cities were presented. The results showed many interesting results and trends such as: (i) an increase in passengers leads to an increase in the size of the central region (i.e., with double coverage); (ii) stop spacing decreases when either the demand increase or becomes more concentrated; (iii) the capacity constraint can quickly become binding in low capacity modes and that can dictate a reduction in headway; (iv) agency investment decreases significantly faster when there is an increase in the number of passengers; (v) users’ total cost for cheaper infrastructure, vehicle, and operating modes (e.g., bus, BRT) decreases more when demand becomes more concentrated; (vi) users’ total cost for expensive infrastructure, vehicle, and operating modes (e.g., metro) decreases more when there is an increase in the number of passengers; (vii) the total cost decreases more
for the BRT and metro modes when the number of passengers increase; and (viii) the total cost decreases more for the bus mode when demand becomes more concentrated.

There are a few limitations with the formulation presented in this chapter: (i) the spatial demand peak, for the spatially heterogeneous city, is assumed to be located exactly at the center of the city; and (ii) the entire city must be served by transit. Regarding the first assumption, the fact that each zone-to-zone demand is uniformly distributed within each pair of zones allows for a small amount of flexibility in the exact location of the peak demand. However, if the peak demand is located farther away from center it would be possible that the design would consist of the central double coverage region being located in an area of much lower demand and the peripheral radial routes to be located in the area of peak demand. This is not a practical design. The second limitation forces the model to determine the optimal transit system for the entire city. It is easy, however, to imagine a city where only a part of the area has transit service coverage. The agency may be able to provide excellent service to the passengers in this area for numerous reasons (e.g., spatial demand pattern, geographical constraints). If the agency is forced to provide transit to the entire city the total agency cost for vehicles and infrastructure is likely to increase. In addition, the passengers who used to receive excellent service may now receive bad service due to limited resources that the agency can afford. The following chapters continue to extend this work by allowing the transit system to be provided as an option; i.e., determining the optimal size and location of transit networks under spatially heterogeneous demand.
CHAPTER 4

TRANSIT NETWORK DESIGN WITH SPATIALLY ELASTIC DEMAND

In the previous chapter, formulations for the generalized cost of the system’s passengers and to its operating agency were developed using zone-to-zone demands calculated from a continuous demand density function to capture the spatially heterogeneity of a city’s demand. The major assumption was that every trip generated within this region would be serviced by the transit system; however we know that many people, especially in the Unites States, continue to drive even if their trip could be via transit. There are numerous reasons for a person to prefer driving over using transit. These reasons can be broken down into two subgroups (i) personal, and (ii) transit performance. Many people prefer to drive for personal reason (e.g., mobility and freedom, thinking transit is for the poor). Additionally, if the transit agency provides a system that has poor temporal or spatial coverage, with many transfers, a high fare, and/or confusing operations, people may not be interested in using it. What is also interesting is that the spiral effects these supply and demand reasons have on each other. If more people decide to drive, the transit agency may receive less funds from fares and possibly from the government. This reduction in funds requires the agency to cut back on the temporal and/or spatial coverage resulting in a less preferred system. After this downgrade even more users will switch to alternative travel modes, further reducing the funds. This cycle could continue until there is no more transit or the agency is able to find some equilibrium. Therefore, it is vital to find the optimal size of the transit region that will provide and maintain the best service possible, reducing the number of transit users who would switch to driving. The goal of this chapter is to develop a model that determines the optimal size of the transit structure when there exists spatial heterogeneous demand. This problem is similar to the TNDFSP with elastic demand since the transit demand depends on the size (i.e., performance) of the transit network. The major difference is that the demand is spatially elastic (i.e., if a person’s origin and destination exist within the transit service region it is assumed they will take transit 100% of the time).

As mentioned in Section 2.3.3, elastic demand was introduced in a number of studies where the area of service is fixed. None, to this authors knowledge, studied how large a transit system should be. Hurdle (1973) did stated that the boundaries of transit service areas are
located where natural (e.g., river) obstacles occur, the population density is deemed too small to justify service, or jurisdictional lines limit operations. In this chapter we look at optimally determining the size of a transit service region by formulating the generalized cost for the transit system’s passengers, its operating agency, and the automobile drivers who cannot complete their trip by transit alone.

4.1 Formulation

4.1.1 Transit System, Vehicle, Passenger, and Automobile Characteristics

In this chapter, the city and the transit service region are concentric squares with side lengths of $\bar{D}$ and $D$, respectively; see Figure 4.1. An additional decision variable, $\alpha_D$, is defined as the ratio of side lengths of the transit service region and the city (i.e., $\alpha_D = D/\bar{D}$). All other notation for the transit system, transit vehicles, and passengers from Chapter 3 remains the same. New parameters for automobile drivers will be introduced shortly, when the driving cost formulation is derived.

4.1.2 Agency Investment, User Cost, and Automobile Cost

Similar to the previous chapter, trip types are defined for each combination of origin and destination region. New for this chapter is a 5th trip type, automobile travel, which include anyone with an origin and/or a destination outside of the transit service region. All the trip types can be calculated with the following zone-to-zone demand formulations:

$$\lambda_{c-c} = \int_{y_2=0}^{ub} \int_{y_1=0}^{ub} \int_{x_2=0}^{ub} \int_{x_1=0}^{ub} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2; \quad (4.1)$$

$$\lambda_{c-p} = \int_{y_2=0}^{\bar{D}} \int_{y_1=0}^{\bar{D}} \int_{x_2=0}^{\bar{D}} \int_{x_1=0}^{\bar{D}} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c}; \quad (4.2)$$

$$\lambda_{p-c} = \int_{y_2=0}^{ub} \int_{y_1=0}^{yb} \int_{x_2=0}^{ub} \int_{x_1=0}^{ub} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c}; \quad (4.3)$$

$$\lambda_{p-p} = \int_{y_2=0}^{ub} \int_{y_1=0}^{ub} \int_{x_2=0}^{ub} \int_{x_1=0}^{ub} \delta(x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c} - \lambda_{c-p} - \lambda_{p-c}, \text{ where} \quad (4.4)$$
Figure 4.1: Hybrid transit structure within city limits; adapted from Daganzo (2010b).

\[ ub = \frac{\bar{D}}{2} (1 + \alpha \alpha_d), \]  
\[ lb = \frac{\bar{D}}{2} (1 - \alpha \alpha_d), \]  
\[ UB = \frac{\bar{D}}{2} (1 + \alpha_D), \]  
\[ LB = \frac{\bar{D}}{2} (1 - \alpha_D). \]  

The total population, \( T_{\text{pop}} \), was calculated by integrating over all the origins and destinations in the entire city. The total number of transit passengers, \( T_{\text{pax}} \) is equal to the summation of every \( \lambda \). Therefore, the total number of drivers is \( T_{\text{auto}} = T_{\text{pop}} - T_{\text{pax}} \). The users’ expected access, waiting, and in-vehicle travel times, from Equations (3.13)-(3.15), and the agency’s expected infrastructure length, vehicle km traveled, and fleet size, from Equations (3.10) - (3.12), will be used.

Typical automobile trips consist of walking to a vehicle (parking lot), driving directly to the desired destination (parking lot), and walking from the vehicle to the final destination. This
is similar to the access and in-vehicle travel time transit passengers experience. The main difference is automobile drivers do not transfer and can drive as direct as possible to their destination. Transit passengers are forced to traverse the transit system, likely increasing their total in-vehicle trip distance. While a driver must endure the time spent accessing and traveling in-vehicle, there is also a monetary cost incurred by the driver (e.g., for gas, maintenance, repairs, parking, insurance, deprecation) and by the government agency who builds and maintains the roads (e.g., roadway maintenance and construction). However, the costs incurred by the government agency are beyond the scope of this thesis, so they are not explicitly included. Similarly, transit passengers and transit agencies also incur a monetary cost to use and operate the transit system, respectively. However, the transit agency’s monetary cost is typically passed along to the transit passenger (either directly via the fare, or indirectly via tax and subsidy). From the description of a driver’s automobile trip and monetary costs, the total automobile cost (in units of time) for a single driver is:

\[
Z_D = \left( \frac{\delta_{auto}}{v_w} + \frac{E_{auto}}{v_{auto}} \right) + \frac{1}{\mu} \left( \$F + \$D E_{auto} \right),
\]

where \( \delta_{auto} \) is an accessing penalty representing the distance to the driver’s vehicle and from the driver’s vehicle to their destination; \( E_{auto} \) is the in-vehicle travel distance; \( v_{auto} \) is the commercial speed of the automobile mode; \( \mu \) is the value of time of the driver; \( \$F \) is the total fixed cost per trip (e.g., tolls, parking); and \( \$D \) is the total variable cost per distance (e.g., gas, insurance, depreciation). The total in-vehicle travel distance of all drivers can be calculated from the continuous density function. The Manhattan distance for a driver located at \((x_1, y_1)\) to travel to \((x_2, y_2)\) on a grid network is \(|x_1 - x_2| + |y_1 - y_2|\). Therefore, the total in-vehicle distance for all drivers is:

\[
\sum_{\forall \text{drivers}} E_{auto} = \int_{y_2=0}^{D} \int_{x_2=0}^{D} \int_{y_1=0}^{D} \int_{x_1=0}^{D} [\delta(x_1, y_1, x_2, y_2) \left( |x_1 - x_2| + |y_1 - y_2| \right)] dx_1 dy_1 dx_2 dy_2 - \int_{y_2=LB}^{UB} \int_{x_2=UB}^{UB} \int_{y_1=LB}^{UB} \int_{x_1=LB}^{UB} [\delta(x_1, y_1, x_2, y_2) \left( |x_1 - x_2| + |y_1 - y_2| \right)] dx_1 dy_1 dx_2 dy_2.
\]

The expected travel distance is calculating by dividing Equation (4.10) by \( T_{auto} \).
4.1.3 Mathematical Model

To find the optimal values of \( \alpha \), \( s \), \( H \), and \( \alpha_D \), the sum of the agency’s investment, generalized passengers’ costs, and automobile drivers’ costs is minimized, again by using Matlab solver fminsearchcon. The objective is formed by converting both the agency and automobile costs into time units (by dividing them by \( \mu \)), and taking the weighted average of the transit and automobile systems:

\[
\min_{s,H,\alpha,D} \left\{ Z = \frac{1}{T_{pap}} \left[ T_{pax} (Z_A + Z_U) + T_{auto} Z_D \right] \right\}, \quad \text{where}
\]

\[
Z_A = \left[ L_L + V_T + M_M \right] / (\mu T_{pax}),
\]

\[
Z_U = A + W + T + \frac{\delta_\text{v}}{v_w}, \quad \text{and}
\]

\[
Z_D = \left( \frac{\delta_{\text{auto}}}{v_w} + \frac{E_{\text{auto}}}{v_{\text{auto}}} \right) + \frac{1}{\mu} \left( F + D E_{\text{auto}} \right).
\]

Subject to the following constraints:

\[
s \geq s_{\text{min}}; \quad \text{(4.12)}
\]

\[
H \geq H_{\text{min}}; \quad \text{(4.13)}
\]

\[
s/D \leq \alpha \leq 1; \quad \text{(4.14)}
\]

\[
O \leq C; \quad \text{and} \quad \text{(4.15)}
\]

\[
s/D \leq \alpha_D \leq 1, \quad \text{(4.16)}
\]

where the \( O \) was provided by Equation (3.22). Constraints (4.12)-(4.15) are the same as those in the previous chapter. Constraint (4.16) ensures that the transit service region cannot be bigger than the city. Next, the numerical results will be presented.

4.2 Numerical Results

In this section, the optimal size, structure, and operation of a mono-centric city with 20,000 passengers/hr and \( D = 10 \) km was studied. This section uses the modal, continuous density function, and other parameter values found in Table 3.1, Table 3.2, and Section 3.2.

It was discovered, after Equations (4.9) and (4.10) were derived, that Matlab was not able to calculate the total automobile driving distance for heterogeneous demand scenarios. Alternate formulations were attempted and none proved successful. In effort to continue the study of optimal transit service region size \( Z_D \) will be approximated by a single fix cost, \( S_{\text{auto}} \),
per driver which can be interpreted as the penalty for not having access to the transit system. To obtain a reasonable estimation of $_{\text{auto}}$, the expected values from literature for each parameter of Equation (4.9) will be used to approximate $_{\text{auto}}$. Daganzo (2010b) assumed a total access time of 10 minutes and an in-vehicle trip distance for drivers of $2\overline{D}/3$, which is the expected Manhattan distance between two random points in a square. The access time included walking to the vehicle, parking, and walking to the destination. Daganzo (2010b) set the automobile commercial speed equal to the bus cruising speed. Therefore, we assume that automobiles have a cruising speed of 25 km/hr. While the expected in-vehicle distance is for uniform demand scenarios, it was used for this approximation. National Parking Association (2010) indicated that $_{F}$ could vary from $5 to $15 per day for monthly permit holders in 34 of the largest cities in the United States. Internal Revenue Service (2014) sets $D$ to a value of 0.348 $/km for gas, insurance, repairs, and maintenance. Given the range of acceptable values for $_{F}$, the range of values for $_{\text{auto}}$ after combing all terms and converting to units of time is $49 to $76 minutes.

A few issues with the above approximation should be mentioned. First and foremost, this approximation assumes a constant travel distance regardless of the size of the transit region and spatial passenger distribution. Secondly, the cruising speed of automobiles is not effected by the number of automobiles on the road. The network structure and operating frequency are being optimized for the peak hour; therefore one could assume that the cruising speed of automobiles could be lower during this time period. The cruising speed of transit vehicles may not be effected as much, because cities may have some exclusive right of way in key locations to help transit vehicles avoid large queues. Thirdly, there is a wide range of acceptable values for $_{F}$. Finally, $\mu$ is assumed to be constant for the entire population, which we already know is not true. It is also very plausible that the average value of $\mu$ will change from city to city; see Sivakumaran et al. (2014). Due to these issues, a sensitivity analysis is conducted for all three modes to show how the decision variables and cost functions vary with $_{\text{auto}}$.

4.2.1 Sensitivity Analysis of $_{\text{auto}}$

Considering the range of $_{\text{auto}}$ that was calculated and the potential issues it was decided that the sensitivity analysis would cover a wider range of values (i.e., \{30 \leq_{\text{auto}} \leq 95\}). Figures 4.2 and 4.3 present the sensitivity analysis of the decision variables and resulting costs for all three transit modes. At first glimpse the results may seem puzzling because the curves for the three transit modes begin at different values of $_{\text{auto}}$. The curves begin at the lowest $_{\text{auto}}$ value, obtain from the numerical results, that produced a transit network size.
larger than $s_{\text{min}}/\bar{D}$, Constraint (4.16). When this constraint was binding, it was assumed that the system optimum was to not have a transit system at all. Considering that the $\alpha_D$ term was so low, the results were not included in the figures. In short, if $s_{\text{auto}}$ was lower than shown for any of the curves the system optimal solution would be not to build or operate a transit system at all. Note, all three curves in Figure 4.2(a) were truncated after the first instance of $\alpha_D = 1$ (i.e., the transit service region spans the entire city) for display purposes.

Figures 4.2 and 4.3 illustrate some interesting results. For one, Figures 4.2(a), 4.2(b), 4.2(d), 4.2(c), and 4.3(d) distinguish the range for each mode of $s_{\text{auto}}$ over which the transit service region exists but is not the entire city (i.e., $0 < \alpha_D < 1$). More importantly these figures show that the ranges and relative values are not the same across modes. The transit service region changes in size for (i) bus mode between 32 and 62 minutes; (ii) BRT mode between 34 and 58 minutes; and (iii) metro mode between 59 and 90 minutes. One of the potential reasons for the various ranges is the different capital costs and operational characteristics (e.g., cruising speed, capacity) of the different modes. Secondly, the curves after $\alpha_D = 1$ produce the same optimal values for $\alpha$, $s$, and $H$ and the resulting costs per person, as in Chapter 3.

The slopes of $\alpha_D$ in Figure 4.2(a), show: (i) the metro mode is the least sensitive to a change in $s_{\text{auto}}$ up until the critical value of $s_{\text{auto}}$ that makes $\alpha_D = 1$; (ii) the bus and BRT modes have a similar sensitivity up until the critical value of $s_{\text{auto}}$. at which point BRT has a higher sensitivity. The bus and BRT react similarly due to their similar capital costs and operation characteristics. While similar, Figure 4.2(b) illustrates that once the entire area is served, the bus mode has a slightly larger central transit region size than BRT due to lower capital costs.

The effects of mono-centric demand are the easiest to see in Figures 4.2(b) and 4.2(d). For bus and BRT the headway, 4.2(d), decreases near the smallest $\alpha_D$ value and increases as the transit service region grows. This is because the mono-centric demand density function clusters passenger demand near the center. Thus a small increase in the transit service region size boosts the number of passengers at a higher rate when the transit service region size is relatively small (as compare to the case when the transit service region size is close to the city size). This same pattern is not viable in the metro mode because the initial size of the transit service region is larger than the bus and BRT modes – the transit service region has already covered the bulk of the centrally located passenger demand. With each increase in $\alpha_D$ the metro system is forced to operate farther and farther away from the high centrally located demand. To compensate for having to provide transit service to the peripheral passengers, the agency will operate with a slightly larger headway.

Next, a small extension is presented to illustrate one of the possibilities that this work
Figure 4.2: Sensitivity of (a) $\alpha_D$, (b) $d = \alpha_D \bar{D}$, (d) $H$, and (c) $s$ to $\$_{auto}$ for bus, BRT, and metro modes.
Figure 4.3: Sensitivity of costs for (a) bus, (b) BRT, and (c) metro modes; and (d) relative total per person costs for all modes.
4.2.2 Optimal Location and Size of Transit Network

Previously the question of how large a transit service region should be was addressed. In this section another important decision for transit planners is addressed: where to locate the transit network. While, the majority of urban areas may have a well defined area that generates and attracts the majority of the trips (e.g., CBD), some urban areas may consist of multiple peaks in demand. Attempting to design the transit network for these areas may come with some difficulty. At this time, the goal is to determine the optimal location of a single transit system that optimally covers one or more of the demand peaks.

Given the number of passengers and automobile drivers, the agency’s investment and users’ costs formulations are unaffected by the demand distribution. Therefore, by updating the zone-to-zone demand functions to include the new decision variables of $x_D$ and $y_D$, which are the x and y coordinates of the center of the single transit service region, we allow for the demand density function to be captured by cost formulations. The zone-to-zone demand formulations become the following:

$$
\lambda_{c-c} = \int_{y_2 = b_y}^{b_y} \int_{x_2 = b_x}^{b_x} \int_{y_1 = b_y}^{b_y} \int_{x_1 = b_x}^{b_x} \delta (x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2; \tag{4.17}
$$

$$
\lambda_{c-p} = \int_{y_2 = 0}^{b_y} \int_{x_2 = 0}^{b_x} \int_{y_1 = b_y}^{b_y} \int_{x_1 = b_x}^{b_x} \delta (x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c}; \tag{4.18}
$$

$$
\lambda_{p-c} = \int_{y_2 = b_y}^{b_y} \int_{x_2 = b_x}^{b_x} \int_{y_1 = 0}^{b_y} \int_{x_1 = 0}^{b_x} \delta (x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c}; \tag{4.19}
$$

$$
\lambda_{p-p} = \int_{y_2 = 0}^{b_y} \int_{x_2 = 0}^{b_x} \int_{y_1 = 0}^{b_y} \int_{x_1 = 0}^{b_x} \delta (x_1, y_1, x_2, y_2) \, dx_1 \, dy_1 \, dx_2 \, dy_2 - \lambda_{c-c} - \lambda_{c-p} - \lambda_{p-c}, \tag{4.20}
$$

where

$$
ub_x = x_D + \frac{\alpha \alpha D \bar{D}}{2}, \tag{4.22}
$$

$$
ub_y = y_D + \frac{\alpha \alpha D \bar{D}}{2}, \tag{4.23}
$$

$$
lb_x = x_D - \frac{\alpha \alpha D \bar{D}}{2}, \tag{4.24}
$$
\[ l_b_y = y_D - \frac{\alpha_D \bar{D}}{2}, \quad (4.25) \]
\[ U_B_x = x_D + \frac{\alpha_D \bar{D}}{2}, \quad (4.26) \]
\[ U_B_y = y_D + \frac{\alpha_D \bar{D}}{2}, \quad (4.27) \]
\[ L_B_x = x_D - \frac{\alpha_D \bar{D}}{2}, \quad \text{and} \quad (4.28) \]
\[ L_B_y = y_D - \frac{\alpha_D \bar{D}}{2}. \quad (4.29) \]

To ensure that the transit service region exists within the city’s boundaries only, the following constraints are added:

\[ x_D + \frac{\alpha_D \bar{D}}{2}, \quad (4.30) \]
\[ y_D + \frac{\alpha_D \bar{D}}{2}, \quad (4.31) \]
\[ x_D - \frac{\alpha_D \bar{D}}{2}, \quad \text{and} \quad (4.32) \]
\[ y_D - \frac{\alpha_D \bar{D}}{2}. \quad (4.33) \]

With these additional constraints and updated bounds for the zonal-to-zonal demand formulations, the agency’s investment, users’ costs, and total penalty for not using transit are now dependent on the location of the transit service region. Next, the results for a mono-centric city with a demand peak located at (3,7) and a twin city with two peaks located at (3.25,5) and (6.75,5) are presented to give a brief glimpse at the possible ways this can be applied; see Figure 4.4 for marginal trip distributions of mono-centric and twin cities.

Considering that BRT had the lowest costs for transit operations in Chapter 3 and Section 4.2, it was selected as the transit mode. To ensure that the transit region would not cover the entire city, negating any insights, $S_{\text{\textit{auto}}}$ was set equal to 42. This is where the transit cost equals penalty for not using transit on average to the population; see Figure 4.3(b). The results for the mono-centric and twin cities are shown in Table 4.1. The discussion of the results will be presented in Section 4.3, because of the limited number of scenarios studied.

4.3 Conclusion

This chapter addressed the optimal transit size and location questions by extending the formulations developed in Chapter 3. While, the cost for driving was simplified many insights
Table 4.1: Optimal transit service region location results.

<table>
<thead>
<tr>
<th>Demand Scenario</th>
<th>Mono-Centric</th>
<th>Twin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak(s) [km]</td>
<td>(3.7)</td>
<td>(3.24,5),(6.75,5)</td>
</tr>
<tr>
<td>$T_{pop}$ [pax/hr]</td>
<td>20,000</td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>BRT</td>
<td></td>
</tr>
<tr>
<td>$s_{auto}$ [min/driver]</td>
<td>42</td>
<td></td>
</tr>
</tbody>
</table>

**Decision Variables**

- $x_D$ [km] 3.00 5.00
- $y_D$ [km] 7.00 5.00
- $D = \alpha D \bar{D}$ [km] 5.69 6.10
- $\alpha$ 0.586 0.563
- $s$ [km] 0.388 0.406
- $H$ [min] 3.20 3.24

**Demand Metrics**

- $T_{auto}$ [drivers/hr] 8093 8650
- $T_{pax}$ [pax/hr] 11331 11349
- $\lambda_{c\rightarrow c}$ [pax/hr] 4062 3535
- $\lambda_{c\rightarrow p}$ [pax/hr] 2722 2799
- $\lambda_{p\rightarrow c}$ [pax/hr] 2722 2799
- $\lambda_{p\rightarrow p}$ [pax/hr] 1824 2216

**Performance Metrics**

- $v_c$ [km/hr] 20.0 20.4
- $e_t$ [tran./pax] 1.08 1.10
- $O$ [pax] 76 81

**Agency Metrics**

- $L$ [km] 112 121
- $V$ [veh km/hr] 4420 4668
- $M$ [veh] 221 228

**Users Metrics**

- $A$ [min/pax] 11.6 12.2
- $W$ [min/pax] 3.8 4.0
- $T$ [min/pax] 9.7 10.3
- $\frac{\lambda_{c\rightarrow c}}{\lambda_{c\rightarrow c}}$ [min/pax] 0.97 0.99

**Costs**

- $Z_A$ [min/pax] 7.35 7.76
- $Z_U$ [min/pax] 26.10 27.43
- $Z_D$ [min/driver] 42.00 42.00
- $Z$ [min/person] 37.01 38.14

**Solution time [sec]** 2245 1248
can still be drawn from the sensitivity analysis. It is also very plausible that there exists other software and/or programming languages more suitable to calculating the cost to drive.

The sensitivity analysis showed that the optimal number of passengers served by the transit system is dependent on the penalty for not using transit and the cost and operational characteristics of the transit mode. Similar to Daganzo (2010b), the metro mode is highly unfavored with the high infrastructure costs, low access speed, and low passenger demand. On top of this, the transit structure presently forces every passenger to access the system by walking to the train stop. In cities that have a commuter rail line, passengers typically access the stations by driving. Therefore, allowing the metro mode to have other access means (e.g., park-and-ride, feeder systems), could result in a more realistic representation of the metro mode and provide for a more fair comparison between modes.

While, a transit planner could have manually decided on the appropriate center location for the transit service region, the developed framework allows for additional extensions on optimal transit operations when demand is spatially heterogeneous; this is discussed in more detail in Chapter 5. Like Estrada et al. (2011), allowing each decision variable to be optimized in each spatial/temporal dimension would be very useful in particular for cities similar to the twin city. The optimal transit region size for the twin city, from Table 4.1, is large enough to cover the majority of the two demand peaks; however, it also covers some areas of very low demand, located in the y-direction. If the decisions were made for each dimension the transit service region could fit the demand better. In the following chapter more detail of these additional research ideas are presented.
This chapter discusses many potential future research topics that came to thought during the completion of this thesis. The ideas are grouped into: (i) Extensions to the hybrid transit network, (ii) Multiple transit networks, (iii) Passenger demand extensions, and (iv) Others not related to the previous groupings.

5.1 Extensions to the Hybrid Transit Network

Efforts from Estrada et al. (2011), Nourbakhsh and Ouyang (2012), Badia et al. (2014) expanded upon the original hybrid transit network structure in Daganzo (2010b), allowing for (i) different line and stop spacing; (ii) flexible-route transit with local pick-ups and drop-offs; and (iii) rectangular and ring-radial transit service regions. It is still possible to further expand this school of work in multiple other directions such as: (i) non-branching transit lines; (ii) offsetting the central core region (i.e., the central region and the transit service region are not forced to be concentric); and (iii) separate central and peripheral line and stop spacing.

5.1.1 Non-Branching Routes

In the works by Daganzo (2010b), Estrada et al. (2011), and Badia et al. (2014) transit lines in the peripheral region branch to insure the same spatial coverage throughout the peripheral region (i.e., boarding and alighting passengers will only walk on average $s/2$ to reach a stop or destination). The temporal coverage and vehicle operations are impacted heavily due to the constant spatial coverage. For instance, if the central region is very small, there will be very few transit lines within the central region. In the peripheral region these transit lines will branch many times to maintain coverage. The passenger located near the edge of the service region could end up having a headway many times what the central region observes. If the headway on the edge is very long passengers would not be inclined to use the service.
The transit agency wouldn’t even be inclined to provide service if they can only have a very large headway, since they wouldn’t be able to compete with automobiles.

Nourbakhsh and Ouyang (2012) briefly presented a fixed route transit structure where transit lines do not branch in the peripheral; see Figure 5.1. However, the formulations presented for the total vehicle distance travel per hour and the passengers waiting and access time do not accurately account for the transit structure in the peripheral. For instance, the access time is for the worst case passenger instead of the average which was used throughout this paper and required to correctly make comparisons. One could improve upon these formulations to more accurately represent the network structure.

5.1.2 Offsetting Central Core Region

Expanding on the theme of spatial heterogeneous passenger demand and optimal location of transit networks, it may be of interest to offset the location of the central region within the transit service area (i.e., remove the implicit assumption that the central double-coverage transit service region is at the center of the transit service region); see Figure 5.2. This would allow for even more flexibility in determining the optimal size and location. Components of the agency investment and user cost are expected to change due to this structure. One would have to investigate ways to keep the headways from becoming too large in the far edges of the peripheries, due to transit line branching. A possible solution to this issue is to
have non-branching transit lines, as previously addressed. Comparing the optimal structure and operation frequencies for a branching versus a non-branching transit line, when the double-coverage region is non-centrally located, could be very interesting.

5.1.3 Separate Line and Stop Spacings for Central and Peripheral Regions

Estrada et al. (2011) allows for different stop and line spacing for the entire transit service region. The formulations derived in this thesis for spatially heterogeneous demand could be expanded to include this. Considering the low demand density in the peripheral region, the potential savings for the agency could be quite large. From a similar thought, if the bus stop and/or line spacing in the peripheral and the central regions were separate decisions, each area could be designed with more attention on the passengers’ traveling, originating, or terminating in each region. There are many complexities that may arise, such as how to maintain transit vehicle flow between the central and peripheral regions. Transit vehicles need to be able to cross the border between central and peripheral regions without conflict (e.g., a transit line terminating, forcing a vehicle to wait to enter the central region until the vehicle achieves the proper central region headway). At the same time the transit system must be easy for the passengers to use and understand.
5.2 Multiple Transit Networks

Most research thus far has focused on optimizing a single transit system without considering impacts on other transit systems operating in the same area. For instance, Chicago’s CBD area has a subway, commuter rail, and two bus networks operated by 3 different transit agencies (CTA, Pace, Metra). There is also a fourth umbrella agency (RTA) whose goal is to facilitate synergies between all three transit agencies. Even with the RTA, each transit system was originally designed without much consideration to one another. Many potential benefits may come to light when all of the area’s transit systems are designed and optimized together (e.g., better connectivity, more economies of scale due to specialization of service). Presented below are three areas of interest.

5.2.1 Separate Central and Peripheral Transit Networks

The first category consists of having different central and peripheral transit systems. Imagine the same grand structure used before, but with one mode of service in the central region and another mode in the peripheral region; see Figure 5.3. For example, a city may wish to have a metro system in the dense urban area, and buses in the peripheral region. The buses would act as a feeder system to the metro network. One of the downsides and complexities is that everyone traveling between the central and peripheral regions would be forced to transfer systems at the boundary (i.e., vehicles for each transit system are only used within their own region). Still, this structure could be beneficial. Ideally, the central and peripheral systems would be synchronized and provide direct connections via stop infrastructure resulting in easier transfers. Within this structure the central region’s transit lines do not need to follow a grid, instead the central region could have its own central and peripheral subregions (i.e., another decision variable would determine the size of the central subregion).

Another possible design within this grand structure of separate central and peripheral transit networks is to have a semi-flexible network, where the central region has fixed routes and the peripheral region has flexible routes (or demand responsive); see Figure 5.4. Flexible and fixed routes have been a heavily researched topic, especially because of the Americans with Disabilities Act of 1991 (Palmer et al., 2008). One group of researchers has made the decision to use fixed or flexible routes a binary decision (Quadrifoglio and Li, 2009; Li and Quadrifoglio, 2010), while others have focused on how flexible networks will operate (Daganzo, 1984; Zhao and Dessouky, 2008). For instance, Daganzo (1984) used a combination of fixed pickup and drop off locations with an operational policy that allowed transit vehicles to skip stops with no demand to lower cost and improve service. The closest research to the
proposed idea was by Aldaihani et al. (2004) and Nourbakhsh and Ouyang (2012), where they designed a semi-flexible and structure flexible network, respectively. Of particular interest is Nourbakhsh and Ouyang (2012) work as this has a similar structural framework as this thesis (i.e., the hybrid network structure). A potential future research topic is adapting Nourbakhsh and Ouyang (2012) to allow for fixed transit in the higher density central region. Another complexity that will need to be addressed is how to keep the headway constant in the central region when the trip length in the peripheral is stochastic.

5.2.2 Feeder and Commuter Transit Networks

Many authors have studied the design and operation of commuter transit (Vuchic and Newell, 1968; Clarens and Hurdle, 1975; Vaughan, 1986; Chien et al., 2003) and feeder routes (Hurdle, 1973; Wirasinghe et al., 1977; Kuah and Perl, 1988; Chang and Schonfeld, 1991, 1993a; Chien and Schonfeld, 1998; Quadrifoglio and Li, 2009; Li and Quadrifoglio, 2010; Deng et al., 2013); for a review see Section 2.2.3. It should be possible to combine feeder routes and commuter routes into the hybrid network structure framework. One idea that stems from this, is to operate the feeder routes in the peripheral region only; see Figure 5.5. Note, Figure 5.5(b)
is just one small area of the peripheral region. This system could greatly benefit the more expensive transit modes (e.g., metro, light rail) due to the increased access speed to the commuting line. Additionally, this would allow for the larger spacing in the peripheral, significantly reducing the infrastructure costs.

5.2.3 Overlapping Multiple Transit Networks

Another structure that could be beneficial is having multiple transit systems overlapping one another; see Figure 5.6. If operated with a high degree of coordination, the overlapping systems could greatly improve the movement of passengers throughout the city. Image a transit system that is optimally designed to move passengers long distances and another for short distances. The benefits are very similar to the feeder and commuter lines previously presented. In Figure 5.6, the top network has a smaller central region that overlaps a portion of the bottom network. One key decision is how to have the two systems operating in unison (e.g., the two systems could be operate on a common headway, have some stops that overlap one another, or have a common central terminal(s) where all transfers would occur). Depending on the coordination and operation procedure transfers could be allowed to only occur at stops with coordination.
Figure 5.5: Feeder and commuter transit network.

Figure 5.6: Overlapping transit structures.
5.3 Passenger Demand Extensions

The main contribution of this thesis was formulating the agency’s investment and users’ costs from a continuous demand density function. As discussed in Section 2.3, temporally heterogeneous, uncertainty, and elasticity are the other defining characteristics of demand. Their individual and combined impacts on transit network design could be investigated.

5.3.1 Spatially and Temporally Heterogeneous Passenger Demand

As previously mentioned in Section 2.3, it is possible to consider spatially and temporally heterogeneous demand together. Hurdle (1973), Clarens and Hurdle (1975), and Newell (1971) already provided for the basis on this topic. With improved solution methods, it should be possible to at least allow the basic hybrid spatially heterogeneous network to be designed considering different levels of spatially heterogeneous demand during multiple time periods. It will be important to investigate the sensitivity of the network design and frequency when the demand, in both time and space dimensions, is not exactly known.

5.3.2 Spatially Heterogeneous Stochastic and Elastic Passenger Demand

Similarly, this research could be further extended if the demand becomes stochastic. Combining the ideas presented in Section 5.2.3 and stochastic demand would allow for a similar service reliability based design as Lo et al. (2013) and An and Lo (2014).

Further work on spatially heterogeneous demand should also be undertaken. In Chapter 4 it was assumed that each driver incurred a fixed cost to drive, covering both their fixed and variable costs. Again, this was assumed due to the computational difficulties in calculating the automobile travel distance. Once this can be calculated, not only will the spatially heterogeneous results improve, but it opens the door to many other exciting possibilities. Of particular interest is applying an elastic modal split demand model to the optimal size and location section. It is anticipated that, depending on the utility function’s coefficients (i.e., how the different parts of their trip are weighted), the size could increase or decrease. Additionally, if the utility functions coefficients could vary spatially (i.e., people who live in the center of a city and those that live on the outskirts likely will value their time/trips differently), the optimal size and location decisions could be further advanced.
5.4 Continuum Approximation Approach

Many of the extensions presented above would indeed be useful, however the complexities resulting from the additional decision variables may be very difficult to solve discretely even with heuristics. Continuum models are a way of reducing the number of variables in the system. The additional needs and purposes for adapting continuum mechanics to continuum models for scheduling, location, and network problems was originally addressed in Newell (1973); see Section 2.1 for further background on continuum models and approximations. Next, two different continuum approximation (CA) approaches are introduced for transit network design and frequency setting problems.

The first CA approach was derived from the area of freight logistics; see Daganzo (2005) for an introduction. There are many types of logistics system problems (e.g., facility location and vehicle routing), and, in general, transit systems can be thought of as specialized logistics systems that move people instead of goods.

Picture a city with a spatially heterogeneous demand. At every location, the optimal stop density could be determined by considering the trade off between passenger access time and agency’s infrastructure and operation costs (i.e., the costs and decisions depend on local conditions only), then discretized in a similar manner as Ouyang and Daganzo (2006). From the discretized bus stop locations, bus routes could be generated that allow for passengers to traverse the network (Van Oudheusden et al., 1987; Fan and Machemehl, 2004). What would be more impressive is to integrate the routing decisions and bus stop location decisions into one mode; which could look like the so called tube model in Wang et al. (2014), except switching the time dimension to be the spatial dimension.

The second CA approach is geared more for a larger metropolitan area that has many small and large peaks in demand (e.g., Los Angeles metropolitan area). Each peak may represent a community’s main street/downtown area to a large city’s CBD area. Placing a grid (Daganzo, 2010b; Estrada et al., 2011) or radial (Badia et al., 2014) hybrid transit structure centered on the largest peak may not adequately serve the communities’ main street/downtown area if they fall within the peripheral region (i.e., single coverage allows for easy access to the central region, but does not allow for easy travel perpendicular to the radial line at other city centers). Therefore, the idea is to place multiple hybrid transit structures within this large metropolitan area and allow local conditions to dictate these size and structure of each; see Figure 5.7. Some areas may not even need a transit system. Then the question becomes how to allow for these subnetwork systems to connect with one another. For the systems that are adjacent to one another it would be possible to have shared routes/vehicles. If instead the area is isolated, additional long-haul transit lines (e.g.,
Figure 5.7: Continuum approximation approach for metropolitan area.

Commuter rail) would need to be added to ensure passengers could reach all destinations from all origins.

5.5 Others

5.5.1 Environmental Impacts

In today’s world, the external impacts of transportation systems on greenhouse gas emissions, energy consumption, other environmental emissions, and pollution (e.g., noise) is well known. However, there has been relatively little literature on operating and design transit networks with environmental considerations. Transit has the potential to make a positive impact on the environment, due to the aggregation of passengers trips compared to when the trips are all made via individual automobiles. However, transit vehicle characteristics and operations (e.g., larger vehicle required for the additional capacity, stopping at traffic lights even during green to pick up passengers) tend to increase the emissions released during operations. Saka (2003) found that the average bus stop spacing nationwide was significantly smaller than the optimal bus stop spacing when the vehicle characteristics and emissions were included. Beltran et al. (2009) presented ways to introduce clean vehicles into the existing transit oper-
ations. One paper, by Griswold et al. (2013), designs the optimal transit system considering a maximum greenhouse gas emission constraint. Combining this modeling framework with other ideas, in particular the multiple transit systems, is of high interest.

5.5.2 Considering Transits Impact on Other Vehicles

As people switch between transit and driving, the vehicular traffic on the respective modal network changes. If everyone abandons transit the additional vehicles on the highway will cause congestion and the speed will drop. The majority of transit network design with elastic demand papers only address the volume changes between the modes; see Section 2.3.4 for a review of elastic demand issues. Only Cipriani et al. (2006), Beltran et al. (2009), and Cipriani et al. (2012b) allowed for each mode to effect the costs (e.g., travel time) of the other modes via a linear function. The research idea given here is to use the macroscopic fundamental diagram (MFD) to determine the cruising speed of each mode, thereby allowing for the vehicle speed and volume to change with demand (Daganzo and Geroliminis, 2008; Geroliminis and Daganzo, 2008). One must be careful to consider the limitations of the MFD, such as accuracy in the congested state (Daganzo et al., 2011).

5.5.3 Service Reliability

The service reliability should be considered during the design of transit networks to ensure steady-state operations in the face of daily traffic changes due to the stochastic nature of commuters (Watling and Cantarella, 2013). Similar to Section 5.3.2, one can use a service reliability based design to keep the transit network reliably operated (Lo et al., 2013; An and Lo, 2014). Another solution would be incorporating operation considerations (e.g., headway-based approach to eliminate bus bunching in Daganzo (2009)) and design decisions (e.g., fare collection system choices as in Tirachini and Hensher (2011)) into the network design model. The additional complexities may prohibit it, but one could adjust/add to the parameters in the given model to represent these decisions. For example, the impact of different fare collection systems on the system design would be quite interesting.
CHAPTER 6

CONCLUSIONS

In this thesis the design, operation, size, and location of the hybrid transit system (Daganzo, 2010b) under continuous heterogeneous demand was investigated. Prior to the analysis, Chapter 2 introduced related literature on (i) continuum approximation methods and their application in transit network design; (ii) TNDP; (iii) TNFSP; (iv) TNDFSP; (v) temporally heterogeneous passenger demand; (vi) spatially heterogeneous passenger demand; (vii) stochastic passenger demand; (viii) elastic passenger demand; and (ix) common solution methods for transit related problems.

Chapter 3 provided the formulations for the agency’s investment and users’ cost for a hybrid transit system with spatially heterogeneous passenger demand. In addition, Chapter 3 analyzed the optimal hybrid transit structure and operations for uniform and mono-centric spatial demand distributions each with a low (20,000 pax/h) and a high (80,000 pax/h) passenger demand scenario. Section 3.3 discusses the effects on agency and user cost metrics when passenger demand is increased and when passenger demand is concentrated. The overlapping theme is that the cost parameters of the infrastructure, vehicle, and operating modes dictate which change (i.e., uniform passenger demand distribution turning into a mono-centric passenger demand distribution, low to high passenger demand) results in the lowest total cost. It may be possible to analytically determine the indifference boundaries that separates the cases which either prefers an increase in passengers or concentrating. For the results presented, bus and BRT are preferred when demand becomes more concentrated while metro is preferred when there is an increase in demand. Because the hybrid transit structure forces all passengers to be served, it is this authors belief that the distribution of passengers in the peripheral region dictates if an increase in concentration or increase of passengers results in a lower total cost per passenger. For instance, due to the expensive infrastructure costs and large passenger capacity of the metro, passengers located in the peripheral can experience long waiting times if the demand becomes more concentrated, due to the excessive branching that can occur and lack of additional lines required to cover demand (i.e., the capacity constraint is not binding so additional lines are not required to cover the demand in the central region).
Chapter 4 studied the optimal size and location for building hybrid transit systems under spatially heterogeneous passenger demand. The population was now allowed to decide between taking transit and driving an automobile. It was assumed for simplicity that everyone within the transit service region would take transit and the rest of the population would drive (i.e., incur a penalty for not taking transit). The penalty for not taking transit was formulated and for each transit mode a sensitivity analysis was performed (due to the uncertainty in the value of the non-transit penalty). The results indicate that the optimal size of the transit service region was dependent on the penalty for not taking transit and the transit mode’s cost and operational characteristics. Mono-centric and twin cities were used to illustrate the application of the model.

Finally, Chapter 5 presented a number of future research ideas: (i) non-branching routes; (ii) offsetting the central core region; (iii) separate line and stop spacings for central and peripheral regions; (iv) separate central and peripheral transit networks; (v) feeder and commuter transit networks; (vi) overlapping multiple transit networks; (vii) spatially and temporally heterogeneous passenger demand; (viii) spatially heterogeneous stochastic and elastic passenger demand; (ix) continuum approximation approach; (x) environmental impacts; (xi) considering transits impact on other vehicles; and (xii) service reliability.
REFERENCES


Ouyang, Y., Nourbakhsh, S. M., Cassidy, M., 2014. Continuum approximation approach to
bus network design under spatially heterogeneous demand. Transportation Research Part
B: Methodological (14-0277), accepted for publication.

Palmer, K., Dessouky, M., Zhou, Z., 2008. Factors influencing productivity and operating
cost of demand responsive transit. Transportation Research Part A: Policy and Practice
42 (3), 503–523.


Quadrifoglio, L., Li, X., 2009. A methodology to derive the critical demand density for design-
ing and operating feeder transit services. Transportation Research Part B: Methodological
43 (10), 922–935.


rep., Texas A&M Transportation Institute, available online at http://mobility.tamu.

Computers & operations research 1 (2), 201–211.

Engineering 129 (4), 331–341.

technology. Transportation Research Part A: Policy and Practice 59, 204–221.

Tirachini, A., Hensher, D. A., 2011. Bus congestion, optimal infrastructure investment and
the choice of a fare collection system in dedicated bus corridors. Transportation Research

with an improved model for public transport analysis. Transportation Research Part E:
Logistics and Transportation Review 46 (6), 1148–1168.


online at http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=f000000_


A.1 Transit Network Design with Spatially Heterogeneous Demand

**Result 1.** The total length of the two-way infrastructure system is given by (3.10):

\[ L = \frac{D^2}{s} (1 + \alpha^2). \]

**Proof.** The same as in Daganzo (2010b).

**Result 2.** The total vehicle-distance traveled per hour is given by (3.11):

\[ V = \frac{2D^2}{sH} (3\alpha - \alpha^2). \]

**Proof.** The same as in Daganzo (2010b).

**Result 3.** The expected number of transfers per trip is given by (3.9):

\[ e_t = 1 + \frac{1}{2} \left( \frac{\lambda_{p-p}}{T_{pax}} \right). \]

**Proof.** Similar to Daganzo (2010b), all trips are assumed to require at least one transfer and at most two transfers. The only transit passengers that require two transfers are those passengers with origins and destinations in the peripheral within the same hemisphere. Ignoring the small number of \( \lambda_{p-p} \) passengers, whose origin and destination are along the same route, the probability of a \( \lambda_{p-p} \) passenger to require two transfer is \( \frac{1}{2} \). This is the percentage of \( \lambda_{p-p} \) passengers who have an origin and destination within the same hemisphere. Therefore, the expected number of transfers for the entire population is, as claimed.

**Result 4.** The combined expected walking time at the origin and destination is given by (3.13):

\[ A = \frac{s}{v_w}. \]

**Proof.** The same as in Daganzo (2010b).
Result 5. The expected waiting time per user including the origin and all transfer stops is given by (3.14):

\[ W = \frac{H}{2T_{pax}} \left[ 2\lambda_{c-c} + \lambda_{p-c} + \lambda_{c-p} + \frac{\lambda_{p-p}}{2} + \frac{2(2\lambda_{p-p} + \lambda_{p-c} + \lambda_{c-p})}{3} \left( \frac{1}{\alpha (\alpha + 1)} + 1 \right) \right]. \]

Proof. Similar to Daganzo (2010b) we assume that the headways are low; so people arrive independent of the schedule. Additionally, the expected wait time has three components: (i) at the origin stop, \( W_O \); (ii) at the last transfer stop prior to your destination, \( W_D \); and (iii) at the intermediate transfer stop, \( W_T \), only for trips requiring more than one transfer. Recall all trips were assumed to have to transfer at least once, and only a portion of \( \lambda_{p-p} \) passengers whose origins and destinations are in the same hemisphere will transfer twice.

The expected wait at the origin can be broken into (a) passengers with origins in the central region, \( \lambda_{c-c} + \lambda_{c-p} \) and (b) passengers with origins in the peripheral, \( \lambda_{p-c} + \lambda_{p-p} \). For (a) the headway is constant and thus the expected wait for these passengers is \( H/2 \). For (b) headway in the peripheral is not constant, however we do know that the flow of buses crossing the \( \alpha \) cordon is constant. Due to the conservation of flow, the headway at cordon \( \beta \), \( H(\beta)/2\alpha \), should be proportional to \( H\beta/2\alpha \). To find the unconditional expectation of \( W_O \), we integrate \( H(\beta)/2\alpha \) over the peripheral region considering the population that has origins in cordon \( \beta \). This results in:

\[ W_O = \frac{H (\lambda_{c-c} + \lambda_{c-p})}{2T_{pax}} + \int_{\beta=\alpha}^{1} \frac{H\beta}{2\alpha} \left( \frac{\lambda_{p-p} + \lambda_{p-c}}{T_{pax} (1 - \alpha^2)} \right) d\beta \]

\[ = \frac{H (\lambda_{c-c} + \lambda_{c-p})}{2T_{pax}} + \frac{H (\lambda_{p-p} + \lambda_{p-c})}{T_{pax} \alpha (1 - \alpha^2)} \left( \frac{1}{3} - \frac{\alpha^3}{3} \right), \]

where the term in parenthesis inside the integral can be considered the population density of cordon \( \beta \). From this it is easy to obtain \( W_D \) which is:

\[ W_D = \frac{H (\lambda_{c-c} + \lambda_{p-c})}{2T_{pax}} + \frac{H (\lambda_{p-p} + \lambda_{p-c})}{T_{pax} \alpha (1 - \alpha^2)} \left( \frac{1}{3} - \frac{\alpha^3}{3} \right). \]

The expected wait at an intermediate transfer stop is the percentage of \( \lambda_{p-p} \) passengers who have an origin and destination within the same hemisphere. Considering that transfers will always occur in the central region,

\[ W_T = (e_t - 1) \frac{H}{2} = \frac{H\lambda_{p-p}}{4T_{pax}}. \]

Combining and substituting \( 1/ (\alpha (\alpha + 1)) + 1 \) in place of \( (1 - \alpha^3) / (\alpha (1 - \alpha^2)) \) will result in (3.14). \( \square \)
Result 6. The expected in-vehicle travel distance per trip is given by (3.16):

\[
E = \frac{\lambda_{p-p}}{T_{pax}} \left[ \frac{D}{2} (2 - 3\alpha + \alpha^3) + \frac{11\alpha D}{12} \right] + \frac{\lambda_{c-c}}{T_{pax}} \left[ \frac{2\alpha D}{3} \right] + \frac{\lambda_{c-p} + \lambda_{p-c}}{T_{pax}} \left[ \frac{D}{4} (2 - 3\alpha + \alpha^3) + \frac{\alpha D}{12} (11 - \alpha^2) \right].
\]

Proof.

The expected trip length for Central-Central and Peripheral-Peripheral trips are defined as \(E_C\) and \(E_P\), respectively. The expected trip length for both Central-Peripheral and Peripheral-Central trips is defined as \(E_{C/P}\), since these two trip types are the inverse of one another. Each of these expectations can consist of a central and peripheral trip segment. Daganzo (2010b) provided the formulations for the expected trip length for a peripheral segment, \((D/4) (2 - 3\alpha + \alpha^3)\), and the expected trip length for a central region segment for trips that have only one end of their trip distributed uniformly over the entire central region, \((\alpha D/12) (11 - \alpha^2)\). For trips that have both their origin and destinations distributed uniformly within the central region the expected trip length is \(2\alpha D/3\). Using these formulations,

\[
E_C = \frac{2\alpha D}{3}, \quad \text{and} \quad E_{C/P} = \frac{D}{4} (2 - 3\alpha + \alpha^3) + \frac{\alpha D}{12} (11 - \alpha^2).
\]

Peripheral-Peripheral trips travel inbound to the central region, traverse the central region in one of three ways based on their destination, and then travel outbound from the central region. The expected distance for the inbound and outbound segments of the trip are the same. The three ways Peripheral-Peripheral trips traverse the central region and their expected length are: (i) enter and exit the central region on the same side, \(\alpha D/3\); (ii) exit on a side that is perpendicular to the side of entry, \(\alpha D\); and (iii) exit on the opposite side of entry, \(4\alpha D/3\). Since Peripheral-Peripheral trips have a uniform distribution for both their origins and destinations and the length of each side of the central region is the same, the probability of each of the three ways to traverse the central region are proportion to 1/4, 1/2, and 1/4, respectively. This results in the expected trip length for Peripheral-Peripheral trips:

\[
E_P = \frac{D}{2} (2 - 3\alpha + \alpha^3) + \frac{11\alpha D}{12}.
\]

Multiplying \(E_C\), \(E_{C/P}\), and \(E_P\) by the probability of each trip results in (3.16). □
Corollary 1. The average in-vehicle travel time, \( T \), obeys (3.15):
\[
T = \frac{E}{v_c}.
\]

Proof. Passengers and buses travel at the same speed, so the result is clear. □

Result 7. The expected commercial speed during the rush hour is given by (3.8):
\[
\frac{1}{v_c} = \frac{1}{v_t} + \frac{\tau}{s} \frac{PHFsH}{D^2 (3\alpha - \alpha^2)} (\lambda_{c-c} + \lambda_{c-p} + \lambda_{p-c} + 0.5 (e_t + 1) \lambda_{p-p}).
\]

Proof. The first two terms are the same as in Daganzo (2010b) and are the time consumed overcoming distance and per stop. The third term, which is denoted as \( \tau \gamma \), is the additional time consumed per boarding passenger per unit distance. The additional time per boarding passenger is given as \( \tau \), so we must only find the average number of boardings per distance, \( \gamma \). From Daganzo (2010b) this can be approximate by the ratio of boarding generated per hour and the total vehicle-km traveled per hour. Note that each trip type is assumed to transfer at least one, which means each trip has at least 2 boardings. There is also a portion of Peripheral-Peripheral trips that require an additional boarding. This results in:
\[
\gamma = PHFsH \frac{(2\lambda_{c-c} + 2\lambda_{c-p} + 2\lambda_{p-c} + 2\lambda_{p-p} + (e_t - 1) \lambda_{p-p})}{2D^2 (3\alpha - \alpha^2)}.
\]

Therefore, the expected commercial speed during the rush hour is given by (3.8), as claimed. □

Corollary 2. The second equality of (3.15) holds. The number of vehicles in operation during the rush hour is given by (3.12):
\[
M = \frac{V}{v_c}.
\]

Proof. The same as in Daganzo (2010b). □

Result 8. The expected vehicle occupancy on the critical load point during the rush hour is approximately given by (3.22):
\[
O = \frac{PHFsH}{\alpha D} \max \left\{ \max \left\{ \frac{\lambda_{c-p} + \lambda_{p-c}}{2} + \lambda_{p-p} \right\}; \right. \\
\frac{\lambda_{p-p} \alpha D}{32s} + \frac{\lambda_{c-c}}{4} + \left[ \frac{\lambda_{c-p} + \lambda_{p-c} + \lambda_{p-p}}{2} - \frac{\lambda_{p-p}}{8} \right] \left. \right\}.
\]

Proof. Similar to Daganzo (2010b), the critical vehicle occupancy occurs either on the transit lines just outside the central square that travel radial to/from the central region, or on the lines that are on the \( \alpha \) cordon. The critical vehicle occupancy in the peripheral is located...
between the edge of the central region and the first stop in the peripheral. At this location, none of the transit lines have branched yet. Additionally, we assume that all passengers from the peripheral reach the central region resulting in the highest inbound \((\lambda_{p-c} + \lambda_{p-p})\) and outbound \((\lambda_{c-p} + \lambda_{p-p})\) passenger flows. There are \(4\alpha D/s\) links that carry these passenger flows. Therefore, the maximum average flow per line is \(PHFs(\max\{\lambda_{p-c}; \lambda_{c-p}\} + \lambda_{p-p}) / (4\alpha D)\). Similar to Daganzo (2010b), we assume that the critical passenger flow is no more than twice the average, which results in \(PHFs(\max\{\lambda_{p-c}; \lambda_{c-p}\} + \lambda_{p-p}) / (2\alpha D)\). The critical occupancy is then a ratio of the critical passenger flow and the vehicle flow, \(1/H\), which is the first term inside the max function in (3.22), as claimed.

The critical vehicle occupancy in the central region occurs along the lines on cordon \(\alpha\). To prove this imagine a line slicing through the entire city dividing the city equally into a Northern and Southern component. There are \(\alpha D/s\) equatorial transit lines that cross between the northern and southern halves, and all are located in the central region. Central-Central trips are distributed uniformly over all of the equatorial transit lines, which results in \(PHFs\lambda_{c-c}/4\alpha D\) passengers per equatorial transit line. Central-Peripheral, Peripheral-Central, and Peripheral-Peripheral trips are likely not to be evenly distributed due to different catchment areas and issues maintaining headways when lines branch in the peripheral region. Therefore, we find the average number of passenger per equatorial transit line and multiple by a safety factor of 2. However, passengers with an origin in the Northern half of the E-W hemisphere and destination in the Southern half of the E-W hemisphere will only use the two outside equatorial transit lines. Therefore, these outside two lines will be the location of the critical occupancy. The total number of these trips is \(PHF\lambda_{p-p}/16\), and these trips are divided between the two outside transit lines. There are \(PHF(\lambda_{c-p}/4 + \lambda_{p-c}/4 + \lambda_{p-p}/4 - \lambda_{p-p}/16)\) trips left from all the Central-Peripheral, Peripheral-Central, and Peripheral-Peripheral trips that cross the equator and use all equatorial transit lines. As mentioned above, these trips may not be perfectly uniformly distributed, and thus we find the average passenger flow and multiple by a safety factor of 2. This results in the critical passenger in the central region:

\[
\frac{PHFs}{\alpha D} \left[ \frac{\lambda_{p-p}\alpha D}{32s} + \frac{\lambda_{c-c}}{4} + 2 \left( \frac{\lambda_{c-p} + \lambda_{p-c} + \lambda_{p-p}}{4} - \frac{\lambda_{p-p}}{16} \right) \right],
\]

which is then divided by the vehicle flow, \(1/H\). Combining the critical occupancy in the peripheral and central regions results in (3.22), as claimed. □