"Truly" Empty Vehicle Repositioning and Fleet-Sizing: Optimal Management of an Autonomous Taxi System in New Jersey on a Typical Weekday

Kyle Weston Douglas

Advisor: Professor Alain L. Kornhauser *71

Submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science in Engineering
Department of Operations Research and Financial Engineering
Princeton University

June 2015

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Kyle Weston Douglas

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Kyle Weston Douglas

Abstract

Transportation is central to our everyday lives and exists to increase the utility of all. Since the turn of the twentieth century, innovators have drastically revolutionized the field. Yet still today, inefficiencies exist in our daily routines in terms of cost, safety, mobility, comfort, convenience, and environmental impact. Already, work has begun to analyze a potential system of autonomous taxis for the state of New Jersey from the perspective of local demand on a typical day. However, the practical issues of sizing a finite fleet of autonomous taxis and repositioning the vehicles to best meet stochastic and realized travel demand need to be addressed. As fully autonomous vehicles more closely approach mainstream use, the gravity of these issues magnifies significantly.

This thesis first defines properties of a stochastic queueing network and outlines the significant characteristics of an autonomous taxi system for New Jersey with an infinite fleet of vehicles under different parameter sets. The understanding of ride sharing, vehicle departures, and vehicle arrivals in the infinite fleet case forms a necessary starting point for work on empty vehicle repositioning and fleet-sizing. The thesis outlines a series of naïve empty vehicle repositioning and fleet-sizing policies before providing a formal and detailed mathematical model for the empty vehicle repositioning problem specific to an autonomous taxi system. It then provides policies for empty vehicle repositioning and fleet-sizing based on the model. However, these policies share a high level of complexity and require extremely large dimensionality. The thesis follows up with a series of simpler, alternative policies that require more manageable levels of complexity, and presents some preliminary results on a small but populous subset of New Jersey. Finally, this thesis presents techniques for convenient data visualization through a detailed web application.

Optimal management of even a small-scale autonomous taxi system on a typical day is extremely challenging. The massive, continuous data streams that come with a larger-scale system present further computational difficulties. Research never stops; I hope that this thesis serves as a foundation for further work of my own and of others. This is one small step among many larger ones. Together, we can make autonomous taxi systems a reality.

Acknowledgements

First and foremost, I would like to thank Professor Kornhauser for being not just a great advisor, but also a great mentor, teacher, and friend. Prof. K's enthusiasm is infectious, and his passion for creating a better future is inspiring. Thank you for being there practically every hour of the day to guide me through this work, as well as my Princeton experience overall. I cannot imagine ORFE without you.

I would also like to thank Prof. Simão for his enduring help and support as I developed my mathematical model. Thank you for going out of your way to meet with me; without your help, my model would have looked far messier than it already is.

I am also in debt to Hill Wyrough '14 for his new and improved data set with transportation demand for New Jersey and surrounding areas, courtesy of his national disaggregate transportation demand model. I would also like to thank him for going out of his way to help generate more data sets with new realizations of this demand for continued work of my own and of others. Moreover, I am in debt to Chenyi Chen for extending help and sharing his code that I could adapt to parse through Hill's newly formatted data sets, generating trips files split by mode.

Furthermore, I would like to thank Daniel Jiang for meeting up to help answer my very specific and unusual LaTeX formatting questions. I would pay Daniel back with a few lessons on the basketball court in Dillon; unfortunately, though, I am a terrible basketball player.

I would also like to thank Jaison Zachariah '13 for encouraging me to take ORF 467 with Prof. K and sparking my interest in the future of autonomous transit systems with his own senior thesis.

Thank you to all of my friends who supported me over the last few months. In particular, I owe big thank yous to Chris McCord and to Amanda Muliawan for going out of their way to help me.

Finally, I would like to thank my parents, who always inspire me to be best I can be.

Every conversation helps, and little actions go a long way.

To my parents, who would do anything for me

To my friends, who would always support me

To the future, which only we can help shape

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The empires of the future are the empires of the mind.

Winston Churchill

In this chapter, I highlight key reasons why an autonomous taxi system is a solution that can replace most existing automobile transportation and meet the travel demands of United States residents. I then more fully introduce the empty vehicle repositioning and fleet-sizing problems inherent in the implementation of an autonomous taxi system. Finally, I review some of the existing research on dynamic fleet management. Note that I further discuss autonomous vehicles and their benefits in Chapter 3.

1.1 Overview

Transportation is a necessity of life. Fundamentally, an individual may choose to move from one location to another if he or she believes that the increase of utility from the move will be greater than the disutility incurred through his or her transportation. Thus, transportation is above all a means to increase the utility of individuals. While many people enjoy driving, racing, running, or walking to different extents, it is rarely the case that transportation itself is "fun." Rather, transportation is more of a means to an end and serves the role of a tool in society.

Over the years, different modes of transportation have evolved significantly. As summarized by Billington (1996), Karl Benz is generally considered to have built the world's first automobile with his Benz Patent-Motorwagen in 1886. With the help of early pioneers like Donald Douglas and dozens of large companies, jets and airplanes have replaced steamboats and large vessels to become the most widely used mode of long-distance transportation. The development of intricate infrastructures of railroad and subway lines have also influenced travel patterns and drawn people around the world to public transportation. Yet the primary mode of transportation around the world—and even more so in the United States—is the automobile.

The *personally-owned* automobile dominates all automobile transportation. According to the U.S. Department of Transportation (2009) in the 2009 National Household Travel Survey, personally-owned vehicles have accounted for almost 90% of worker commutes, whereas mass transit has accounted for lower than 5%. The dominance of the personal vehicle extends well beyond commutes into most daily routines in the United States; undoubtedly, the personal vehicle is ubiquitous in modern society.

However, transportation by *non-personally owned* vehicles is currently growing. Traditional taxi systems and car rental companies prove the success of non-personally owned vehicles in serving travel demand that would generally otherwise be taken by personal vehicles if it were convenient to own one or have it at one's location. This is especially true in urban areas and around airports, train stations, and other areas connecting people to different modes of transportation. Recently, services of non-personally owned vehicles outside of traditional taxi systems and car rental companies have become more widely used. As described by Hart et al. (2003), Zipcar was co-founded by Dr. Antje Danielson and Robin Chase in late 1999. Zipcar, now owned by Avis, is an alternative to traditional car rental. was launched in Boston in 2000 as a provider of on-demand vehicles, existing as a service charging membership fees and fixed and variable usage fees. Since 2000, Zipcar has accumulated over 800 thousand members and over ten thousand vehicles to serve its members. Furthermore, newer non-traditional taxi, bike sharing, car rental, and ride sharing companies providing on-demand service like Lyft and Uber have recently begun to capture significant market share. Uber was launched in San Francisco in 2011. According to CEO Travis Kalanick via Huet (2014), Uber is dispatching on average more than one million trips per day with over 160 thousand drivers as of December 2014. It is clear that regardless of ownership, automobiles deliver freedom and mobility.

With the increased sophistication and development of *autonomous* driving technology, the ability for any individual to purchase a fully-automated vehicle may come soon. Furthermore, with the technology comes the ability to soon provide *non-personally owned autonomous* vehicle services almost entirely on the existing infrastructure of roads and highways in New Jersey or the

United States overall. Soon, there will be no reason that a large-scale public transit network of autonomous vehicles, or SmartDrivingCars and potentially SmartDrivingTrucks, cannot provide automobile service on demand in New Jersey or in other subsets of the entire nation.

1.2 Motivation

The underlying purpose of this thesis is to help bring an autonomous taxi system of some scale closer to reality by addressing the specific problems inherent in managing a realistic fleet of self-driving cars. Already, several classes of Professor Kornhauser's students in ORF 467: Transportation Systems Analysis have analyzed potential ride sharing present under different conditions if an autonomous taxi system were in place in New Jersey, and all trips otherwise taken by automobiles were instead taken by autonomous taxis, or aTaxis. I was one of these students in Fall 2013; my experience drew me to investigate such a system more carefully, given that the technology that makes it possible is just around the corner. However, almost all previous analysis performed by ORF 467 classes in recent years applied to a theoretical system with an autonomous taxi system with an infinite fleet size. I highlight some of the ride sharing and other benefits of such a system for New Jersey in Chapter 3. An infinite fleet size is clearly impractical; outside of mathematics, nothing is countably infinite. There will inevitably be negative effects with respect to service upon changing the fleet from an infinite size to some finite size. Mitigating these effects is a critical component of the management of a realistic fleet of aTaxis. The problems of empty vehicle repositioning and fleet-sizing inherent in optimizing the management of a finite fleet presents unique optimization problems that are immensely interesting yet incredibly challenging.

Specifically, empty vehicle repositioning refers the decisions about the movement, or lack of movement, of vehicles between locations without loads of one or more individuals who otherwise dictate the origin and destination locations of trips. It refers to the problem of optimizing such repositioning. In general, this involves simultaneously maximizing the satisfaction of realized and forecasted stochastic travel demand "on time" and the minimizing the transportation cost of moving empty vehicles between locations. The concept of fleet-sizing is straightforward: identifying an optimal number of vehicles of some maximum capacity or an optimal distribution of vehicles of different capacities according to some well-defined objective. This optimization may be based on historical realized travel demand, models of current travel demand, andor forecasting future travel demand. Both optimization problems may possibly be subject to reasonably defined constraints such as service level requirements.

Many things that I have seen firsthand on the road motivate me further to tackle these problems and help bring an autonomous taxi system closer to reality. Texting and driving is a rampant problem on today's roads. While driving on many local roads and highways, I have been unnerved by the astounding number of people I see with their eyes on their phones or other devices instead of the road—at any speed. While I enjoy driving, I often think about how much more productive I could be if I were able to work, or even just send messages or e-mails, while I drive longer distances. I discuss further motivation for the adoption of autonomous vehicles generally in Section 3.2.

1.3 Dynamic Fleet Management

Optimal transportation control systems in today's world often refer to municipalities attempting to optimize traffic flow via the installation and tuning of traffic control instruments such as speed limits, traffic lights, stop signs, exit ramps, speed bumps, and more, possibly subject to legal or other service constraints. Much of the decision making in such control systems are performed by hired transportation engineers or consultants. However, many other transportation control systems exist that guide the management of fleets of jets, trucks, trains, or cars. Trucking and other shipping companies deal with many empty vehicle repositioning and fleet-sizing problems similar to those for an autonomous taxi system; such companies may have to react efficiently to incoming orders throughout the day while possibly strategically positioning units to prepare to accommodate anticipated demand. Other companies such as Uber and Lyft are confronted with even more similar fleet-sizing and empty vehicle repositioning problems on top of pricing problems to maximize revenue subject to certain pre-determined service constraints.

Through the years, many academics and researchers have attempted to provide optimal control policies for empty vehicle repositioning and fleet-sizing using wildly different assumptions. Few of these theoretical frameworks or similar frameworks can be applied to empty vehicle repositioning and fleet-sizing problems that some companies face; most problems, like those that I address in this thesis, are highly specialized.

1.3.1 The Classic Transshipment Problem

The classic transportation or classic transshipment problem refers the deterministic optimization problem of meeting demand requests in a simplified network of supply and demand nodes. It is a highly simplified version of a empty vehicle repositioning problem. It usually assumes that the fleet size, or total supply, is at least as large as the total demand in the model, which is described below. If the fleet size is not at least as large as the total demand, the problem will be infeasible.

The objective of the classic transshipment problem is to perform a one-time assignment of units of supply (i.e., elements of a fleet) to units of demand on a bipartite network of supply and demand nodes. A bipartite graph is a graph composed of two non-intersecting subsets of nodes that together form the set of all nodes in the graph, where each arc, or link, in the graph connects two nodes belonging to different sets. That is, if the set of supply nodes is defined as $\mathcal{S}$, the set of demand nodes is defined as $\mathcal{D}$, and the set of all nodes is defined as $\mathcal{A}$, then $\mathcal{S} \cap \mathcal{D} = \varnothing$ and $\mathcal{S} \cup \mathcal{D} = \mathcal{N}$, while all arcs connect one element of $\mathcal{S}$ with one element of $\mathcal{D}$. For example, the set of supply nodes, $\mathcal{S}$, may represent distribution centers or warehouses, while the set of demand nodes, $\mathcal{S}$ may represent retail stores. Defining the m elements of $\mathcal{S}$ by their indices in the set $\{1, 2, \ldots, m\}$ and the n elements of $\mathcal{D}$ by their indices in the set $\{1, 2, \ldots, n\}$, the classic transshipment problem can then be formulated as follows:

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{D}} c_{ij}\, x_{ij} \\
\text{subject to} \quad & \sum_{j \in \mathcal{D}} x_{ij} \leq s_i, \qquad \forall\, i \in \mathcal{S} \\
& \sum_{i \in \mathcal{S}} x_{ij} = d_j, \qquad \forall\, j \in \mathcal{D} \\
& x_{ij} \geq 0, \qquad \forall\, i \in \mathcal{S}, \qquad \forall\, j \in \mathcal{D}
\end{aligned}
$$

where, in addition to the definitions presented above:

- each element of the decision variable $\mathbf{x} = (x_{ij})_{i \in \mathcal{S},\, j \in \mathcal{D}}$ is the number of units sent from the supply node with index i to the demand node with index j,
- s_i is the total supply of units, at the supply node with index i, where $s_i \geq 0 \;\forall\, i \in \mathcal{S}$, and

- d_j is the total demand for units at the demand node with index j, where $d_j \geq 0 \; \forall \; j \in \mathcal{D}$.

Figure 1.1 below gives an example bipartite graph for the classic transshipment problem. When identified by their labels on the graph, the set of nodes on the left $\{\text{Source } 1, \text{Source } 2, \ldots, \text{Source } m\} = \mathcal{S}$, and the set of nodes on the right $\{\text{Sink } 1, \text{Sink } 2, \ldots, \text{Sink } n\} = \mathcal{D}$. Given that all nodes in the graph are displayed in the figure, we can see that indeed, each of the 9 total arcs, or links, connects one element of $\mathcal{S}$ to one element of $\mathcal{D}$, while $\mathcal{S} \cap \mathcal{D} = \varnothing$ and $\mathcal{S} \cup \mathcal{D} = \mathcal{N}$.

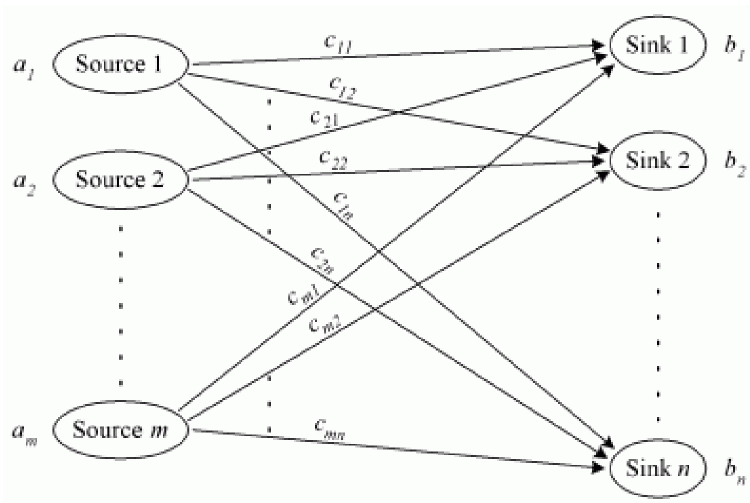


Figure 1.1: Example bipartite graph for the classic transshipment problem.

If units must travel from demand nodes elsewhere and return back to the same or another source node, then the solution to this particular problem will not be optimal because the model does not account for costs between any possible destination locations and supply nodes.

Clearly, the classic transshipment problem is overly simplified with respect to nearly all realistic transportation problems. However, understanding it is fundamental for further research on more complicated empty vehicle repositioning problems as well as fleet-sizing problems with significantly more sophisticated models. I present my own mathematical model in Chapter 5 for the empty vehicle repositioning problem given a finite fleet size specific to the autonomous taxi system that I propose in Chapter 3. My model indirectly serves my fleet-sizing problem through by allowing for tuning of the fixed fleet size assumed in the model.

1.3.2 Deterministic and Stochastic Demand

The empty vehicle repositioning and fleet-sizing problems for an autonomous taxi system tackle both deterministic and stochastic demands. Suppose for simplicity that all trips demanded by individuals are from one node to another node in a large graph where there exist links between each node; I define a relatively similar simple model for an autonomous taxi system for New Jersey more carefully in Chapter 3. These problems require stochastic demand because there is some value in sending one or more empty vehicles from nodes in which I expect relatively fewer "demand requests" to nodes in which I expect relatively more "demand requests" so long as there is sufficient quantitative evidence that the expectation of "effective cost" savings due to such a move is greater than the transportation cost of the move. By "effective cost," I refer to both the cost of transporting empty vehicles and the "cost" associated with failing to provide aTaxi service or failing to provide aTaxi service on time to individuals or groups of individuals. (Tuning this "cost" relative to transportation cost for the empty vehicle repositioning and fleet-sizing problems, and absolutely to some currency for the fleet-sizing problem, is a challenge in and of itself.)

Considerably more important to both problems is reacting to deterministic demand in the form of "demand requests" for aTaxis carrying one or more people, depending on the "demand requests" for aTaxis made at each node over time. This is because one can never accurately predict the future; forecasting demand can be a useful tool, but making use of forecasted demand pales in comparison to making use of demand in the context of these problems. Furthermore, as I will discuss in Chapter 3 and define more precisely in Chapter 5, the total "demand" at each node is the outstanding demand for aTaxis at that node. In the implementation of an autonomous taxi system, there are certain conditions in which individuals will share rides. This is one of most important features of and motivations for such a system; I discuss this fact and outline specific ride sharing conditions further in Chapter 5. One may arguably be able to forecast "demand requests" made by individuals at each node—perhaps including individuals' destination nodes—relatively well, but forecasting "demand requests" for aTaxis is more challenging and less successful due to these constraints. And since what really matters is the demand for aTaxis, forecasting demand for an autonomous taxi system will almost surely be less useful than in most other applications in which the concept of demand is relatively simpler. As I again discuss further in Chapter 3, I assume that there exists a parameter called departure delay that represents the difference between the scheduled time of departure of an aTaxi and the arrival time of the first-arriving individual who will depart in that particular aTaxi. Deterministically, one knows that there must be at least one empty vehicle to assign to a "demand request" at a given node by this scheduled time of departure for an aTaxi in order to serve the demand on time, and thereby avoid some sort of lateness "penalty," if any exists. Even if one cannot have an aTaxi ready by a scheduled departure time at a given node, a deterministic demand remains at the scheduled departure time and thereafter until an empty vehicle arrives at this node and can be assigned to the one or more individuals waiting for a ride.

1.3.3 Existing Research and Literature

There exists significant literature on empty vehicle repositioning and fleet-sizing for systems with both deterministic and stochastic demands. However, no significant theoretical frameworks seem to exist that quite fit the implementation of an autonomous taxi system, which I describe further in Chatper 3. I present a survey below of what I believe to be important papers addressing empty vehicle repositioning and fleet-sizing problems involving some element of stochastic demand, with implementation properties that are relatively similar to those I present in this thesis. I write with as much of the language that I use in the remainder of this thesis as possible to describe equivalent terms or concepts in the papers below.

In their paper entitled "Optimal Empty Vehicle Repositioning and Fleet-Sizing for Two-Depot Service Systems," Song and Earl (2004) consider the problem of determining optimal empty vehicle repositioning and fleet-sizing policies for a two-depot service system with two primary sources of randomness:

1. uncertain loaded vehicle "demand request" times, where the arrival process is a Poisson process, and
2. the times required to reposition empty vehicles, where the times are independent and identically distributed exponential random variables.

A two-depot service system is a simple system of two nodes, in which all "demand requests" for loaded vehicles are between one node and the other. Such a system may represent a port-depot and an inland-depot, between which trucks may transport goods. When a traveling loaded vehicle reaches its destination node, it becomes an available empty vehicle instantaneously. Song and Earl make three key assumptions:

1. Source of randomness (1) above. Poisson processes are very reasonable processes for "demand requests," as is discussed further in Section 4.3.1.

2. Source of randomness (2) above. Travel time between the two nodes is mostly known; the reasoning for the exponential distribution for the time required to reposition each vehicle is due primarily to the availability of drivers, plus any possible required maintenance.

3. If there is no available empty vehicle at a given node when a loaded vehicle "demand request" arrives at that node, an empty vehicle is leased from another company at a location different than that of either node. After delivery from the given node to the other node, this vehicle is returned to the location from which it was leased.

Song and Earl go on to minimize the sum of costs incurred by vehicle maintenance, empty vehicle repositioning, and the leasing of vehicles as described above. Note that the cost of satisfying loaded demands is not included because such costs will always be "required"—at least under the assumption that loaded "demand requests" are always satisfied, which is inherent in the third assumption listed above. Song and Earl identify an optimal empty vehicle repositioning policy using threshold controls. This means that for each node at any given decision time, one vehicle is sent to the other node if a particular event involving the cost function is true, and no vehicles are sent otherwise. Their increments between decision times are closely bunched together relative to the rate of the Poisson process of "demand requests" into each node, such that a binary decision for each node is reasonable. Not surprisingly, Song and Earl classify this type of policy of the "bang-bang type." They then identify an explicit form of the cost function described above, from which they explicitly define the threshold events that determine whether an empty vehicle is sent. Their empty vehicle repositioning policy necessarily assumes a fixed fleet size; they identify an optimal fleet size after identifying the optimal empty vehicle repositioning policy by minimizing the expected total cost, or average simulated total cost, over the set of policies of the same form as that above with different fixed fleet sizes.

Like Song and Earl, I assume that the "demand requests" into each node in my stochastic demand network form Poisson processes. However, the required number of nodes in such a network for an autonomous taxi system will be far greater than two, and likely in the magnitude of tens or hundreds of thousands. In fact, according to the ORF 467 data set that I describe in Section 3.4, there are more than twenty thousand nodes that service at least one "demand request." Additionally, I assume that the rates of the "demand request" processes into each node vary over time. This is a simple realistic necessity, which somewhat complicates my empty vehicle repositioning and fleet-sizing policies.

Like Song and Earl, I assume that the *loaded* vehicle travel time between any two nodes is deterministic. But unlike Song and Earl, I assume that the *empty* vehicle travel time between any two nodes is also deterministic; since the empty autonomous vehicles in my system are "truly empty," I do not need to account for the variance that Song and Earl assume due primarily to the availability of drivers. The objective functions that I identify for empty vehicle repositioning in Chapter 5 and for the Expected Annual Cost fleet-sizing policy that I discuss Section 6.2.2 are similar to Song and Earl's cost function in that they do not included travel costs due to loaded trips. However, I include a penalty for "lateness" in servicing "demand requests" for aTaxis at each node; following their third primary assumption above, Song and Earl avoid such a penalty and instead lease vehicles if there is no available empty vehicle at a node upon the arrival of a loaded "demand request."

It seems necessary to identify an optimal fleet size after identifying an optimal empty vehicle repositioning policy, since a reasonable empty vehicle repositioning policy requires a fixed fleet size. Thus, like Song and Earl, I identify fleet-sizing policies based on the minimization of a cost function by varying the fixed fleet size in a particular class of empty vehicle repositioning policies, possibly subject to service level constraints.

Beaujon and Turnquist (1991) present a methodology that is somewhat closer to a possible application to an autonomous taxi system in their paper entitled "A Model for Fleet Sizing and Vehicle Allocation." Beaujon and Turnquist identify interactions between fleet-sizing and levels of fleet utilization throughout the time period over which empty vehicle repositioning decisions are made. Unlike Earl and Song, who identified a "bang-bang" policy for empty vehicle repositioning, Beaujon and Turnquist form a non-linear program in order to make decisions at each decision time. All of the empty vehicle repositioning policies that I discuss in Chapter 6 consist of solving a linear program at each decision time. Thus, my empty vehicle repositioning policies themselves are more similar to that of Beaujon and Turnquist. In their objective function, Beaujon and Turnquist attribute a "cost" to unsatisfied demand; I include a similar lateness penalty. However, Beaujon and Turnquist maximize profit, which includes a revenue component that is based on their service. I assume that any revenues collected, if any, are not affected by fleet size nor the performance of my empty vehicle repositioning policy, and thus every objective function that I identify in this thesis excludes any expression for revenue or expected revenue.

There are many other papers that investigate optimal policies for empty vehicle repositioning and fleet-sizing for networks involving stochastic demand. Of all of the papers I found, the two papers discussed above appear to be the most relevant to the application of an autonomous taxi system. Further discussion of these other papers is out of the scope and focus of this thesis, but I list several others in my Bibliography for further reference. These papers include those by:

1. Powell and Cheung (1994), which contains a model similar to my own outlined in Chapter 5 in which not all repositioning must occur before realizations of "demand requests" and all travel times are deterministic, but the objective is to maximize a function for profit and all unsatisfied demands are lost to competitors,

2. Powell (1986), which illustrates an optimization problem applied to an empty vehicle repositioning problem with different stochastic demands at different nodes modeled as random variables with known means and variances determined offline, and

3. White (1972), which solves a highly simplified version of an empty vehicle repositioning problem on a network with stochastic demands using an inductive algorithm.

The wide variety of methodologies discussed in each of these papers demonstrates the tremendous impact of different assumptions in each model and the magnitude by which a model generally must change upon switching applications.

Chapter 2

Stochastic Queueing Networks

The future starts today, not tomorrow.

Pope John Paul II

This chapter presents a brief overview of what I define as stochastic queueing networks. I apply this stochastic queueing network framework to an autonomous taxi system for New Jersey, which I discuss in depth in Chapter 3.

2.1 Definition and Properties

I define a stochastic queueing network as a network of nodes into which units arrive in the form of stochastic processes. The graph containing all nodes may or may not be connected, and the stochastic arrival processes into each node need not be identical. Each arriving unit at every node undergoes a service. At each node, there may be a stochastic service time distribution, a deterministic service time, or a service time that is a function of a past, present, or future state or states. In the mathematical model that I provide in Chatper 5 on a stochastic queueing network of nodes representing locations, I assume that the service time at each node is a function of the present and future state of the system. I call this "service" the "assignment" of an available empty aTaxi to a "departure unit," which I later discuss in depth in Section 5.1. Overall, each node represents its own queue with an arrival process, one or more servers with service distributions or service time functions, a maximum capacity for units waiting in the queue, and a service type, such as first-in, first-out. When I apply a stochastic queueing network to an autonomous taxi system, I assume that at each node there is a Poisson arrival process with a time-varying rate, one server with a complicated service time function, an infinite capacity, and a first-in, first-out service type.

Mathematically, Poisson arrival processes are arguably the easiest stochastic arrival processes to work with due to the memorylessness of the exponential interarrival time distributions. However, Poisson processes are not the only possible arrival processes into nodes in a stochastic queueing network. An unknown arrival process or an arrival process that does not fit basic properties of a Poisson distribution may simply be classified as a "general" stochastic process.

2.2 Poisson Arrivals

As discussed by Massey (2014), the canonical customer traffic model is a Poisson process. Çinlar (2012) also explains that Poisson processes reasonably model customer travel demand at fixed locations such as bus stops. Mathematically, a Poisson process with a constant arrival rate is a counting process in which:

- every sample path is an increasing, integer-valued sample path,
- the nth arrival, or "jump," time is a gamma-distributed random variable with shape parameter n and rate parameter equal to the arrival rate, and
- the interarrival times form a family of independent and identically distributed exponential random variables with mean equal to the reciprocal of the arrival rate.

Due to their simple mathematical properties, working with Poisson arrival processes is favorable for complicated models and large-scale stochastic queueing networks. Due to their tendency to reasonably model travel demand requests, transportation engineers and researchers may often take advantage of their simplicity. In Section 4.3.1, I provide evidence reasonably modeling demand for an autonomous taxi system for New Jersey with time-varying Poisson processes at each node in a network.

If there is sufficient evidence for Poisson arrival processes into nodes with fixed rates over short periods of time, I provide three methods below for identifying likely rates as functions of time. I limit my discussions to simple descriptions for each because such theory is not the focus of this thesis.

2.2.1 Arrival Rates Over Uniform Periods

Given sufficient evidence for a Poisson arrival process into a node, one may simply select reasonable uniform time intervals and assign rates matching historical arrival rates within each time interval. The maximum likelihood estimator for the rate of an exponential distribution given a series of interarrival times is equal to the number of interarrival times divided by the sum of all interarrival times. This is equivalent to saying that the maximum likelihood estimator of the rate of a Poisson process over an interval ending with an arrival is the empirical arrival rate over this interval. Thus, assigning arrival rates according to historical rates over uniform periods is a reasonable approach to estimating the arrival rate as a function of time.

If the length of such uniform time periods is too short relative to the interarrival times of the process, the resulting rate as a function of time may contain long periods of assigned rates that are too low, and spikes that overestimate the rate over shorter periods of time. On the other hand, if the length of the uniform time periods is too long, then it may be difficult to capture important trends in the arrival rate.

2.2.2 Offline Change Point Detection

One may assume that the Poisson arrival rate into a node jumps at certain times identified as "change points" from historical data. One possible policy for assigning change points and constant arrival rates between change points is to apply an algorithm minimizing the Bayesian Information Criterion, or BIC. According to Schwarz et al. (1978), BIC is defined as follows:

$$BIC = k \ln{(n)} - 2 \ln L$$

where k is the number of parameters, n is the total number of observations—here, arrivals—over the full interval in which we are interested, and L is the likelihood of the historical realization of arrivals given the rate as a function of time.

The point of minimizing BIC is to maximize likelihood with penalties for increasing complexity and "overfitting" to historical realizations. Such an algorithm may successively minimize BIC for increasing fixed values of k and compare the minimum values for each value of k. I do not perform this analysis in this thesis because using uniform periods of a length that is neither too long nor too short seems to do a better job of predicting “demand requests.” Besides, using uniform periods that are divisible by the uniform length of intervals between decision times is computationally efficient in applying empty vehicle repositioning policies.

2.2.3 Online Change Point Detection

Finally, one may apply the sequential probability ratio test or other techniques of optimal stopping to identify change points online. One motivation for identifying change points online is to handle cases in which arrivals significantly exceed expectations at a node. However, such a policy of assigning arrival rates online may be weak at identifying slowdowns in arrivals, since many algorithms may always require waiting for an arrival before making a decision. Thus, with such a policy, one may overestimate the arrival rate during long periods of inactivity. Indeed, online change point detection is likely not the best policy for forecasting arrival rates in an autonomous taxi system that depends so heavily on historical data. There is a lot to be learned from the past, and to adopt a completely online policy to determine estimated “demand request” rates is to throw away this knowledge.

Chapter 3

Autonomous Taxi System for New Jersey

My interest is in the future because I am going to spend the rest of my life there.

Charles F. Kettering

Here I talk about the Autonomous Taxi System for New Jersey that we currently have in place. It is a system or model in and of itself.

3.1 Overview of Personal Rapid Transit Systems

A personal rapid transit (PRT) system is a public transportation system that operates on a network of stations, but allows individuals or small groups to bypass intermediate stations before reaching their final destination. Examples of successful personal rapid transit systems in place include West Virginia University's PRT system in Morgantown, West Virginia, and the Ultra PRT system at London Heathrow Airport.

A crucial defining characteristic of a PRT system is the need for a separate, independent, and specially built guideway. According to Kornhauser and in ORF 467 (2011), constructing a full-scale PRT system providing on-demand service with small, fully-automated vehicles may require more than ten thousand miles of guideway and upwards of nine thousand stations. Furthermore, building such a system may take up to twenty years and cost over $100 billion. If we could instead use the existing infrastructure of roads and highways to build a public transportation system in New Jersey, or even larger subsets of the United States, constructing a new system becomes significantly more attractive—especially in context of politics. As we continue to make technological advances in the space of autonomous vehicles, an autonomous taxi system using the existing infrastructure becomes closer and closer to possible implementation.

3.2 Autonomous Vehicles: Motivation

As reported by Howard (2015), a Delphi self-driving car drove coast-to-coast just weeks ago over a distance of 3,400 miles, almost entirely under full automation. With significant technological advances as of late, fully autonomous vehicles are coming closer and closer to mainstream use. And even outside of an autonomous taxi system, the motivation to adopt autonomous vehicles over manually-operated vehicles is extensive. Most of this discussion is outside the scope of this paper; I highlight some key facts below.

Autonomous vehicles provide significant overall safety benefits over manually-controlled vehicles. According to Hendricks et al. (2001), driver behavioral errors caused or contributed to 717 out of 723—more than 99% of—studied accidents. As noted in Table 3.1 below, a wide variety of human behaviors can lead to severe accidents. As referred to in the table, perceptual errors include looking for but not seeing other vehicles; decision errors include turning despite an obstructed view; incapacitation includes falling asleep and medical emergencies.

Table 3.1: Distribution of driver behavioral errors in accidents studied by Hendricks et al. (2001).

Behavioral Error	Percentage of Studied Accidents
Driver Inattention	22.7%
Vehicle Speed	18.7%
Alcohol Impairment	18.2%
Perceptual Errors	15.1%
Decision Errors	10.1%
Incapacitation	6.4%

Clearly, human error threatens our safety. Furthermore, human error on the roads calls for high automobile insurance rates and accounts for significant insurance payments. Widespread adoption of autonomous vehicles is likely to be financially beneficial for both insurers and insurees. Autonomous vehicles also allow for increased comfort and productivity during transportation due to the elimination of all or most of the responsibilities associated with driving.

3.3 System Design: New Jersey

3.3.1 Pixilation and Autonomous Taxi Stands

Several classes of ORF 467: Transportation Systems Analysis have defined a coordinate system that converts latitudes and longitudes to square "Pixels" one half of a mile in length on each side. The coordinate system sets the origin at $(-75.6^{o}\,\mathrm{E}, 38.9^{o}\,\mathrm{N})$ and defines Pixel coordinate pairs (X, Y) as follows:

$$X = \text{floor}\,(108.907\,(long + 75.6))$$

$$Y = \text{floor}\,(138.2\,(lat - 38.9))$$

Figure 3.1 displays a visual representation of the grid system overlaid on a map.

We assume that one autonomous taxi stand, or aTaxiStand, is placed in every Pixel throughout the state. Each aTaxiStand serves as a docking station to which individuals travel in order to use an aTaxi. We also assume that individuals who would otherwise demand an automobile trip originating within a given Pixel walk to that Pixel's aTaxiStand in order to fulfill their demand.

"Superpixilation" refers to "superpixilizing" the state by treating 3 x 3 grids of Pixels as described above as one "common location" for trip destinations, trip origins, or both. I do not discuss "superpixilizing" further in this thesis because such treatment of Pixels severely complicates the model that I present in Chapter 5 and any empty vehicle repositioning or fleet-sizing policies that depend on the model.

3.3.2 Ride Sharing

Arguably the most important feature of an autonomous taxi system is its ability to accommodate ride sharing that would otherwise not exist with the use of personally-owned vehicles. The benefits of ride sharing are astounding, ranging from lower emissions, cost savings on fuel or electricity,

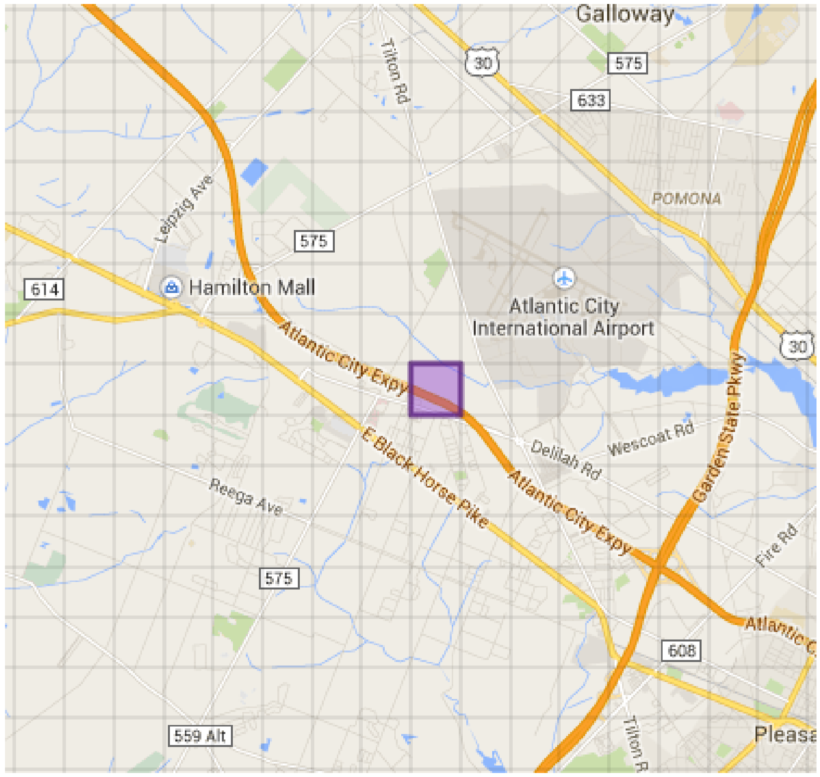


Figure 3.1: Pixilation of New Jersey: a close up of the Atlantic City airport.

decreased congestion, and increased mobility for all.

In the formulation of an autonomous taxi system for New Jersey, I accommodate ride sharing by assigning "departure units" to arriving individuals. If an individual arrives at an aTaxiStand and does not join an existing "departure unit," he or she forms his or her own "departure unit." Immediately, a departure time is scheduled for him or her, plus any individuals who ultimately join his or her "departure unit" prior to this scheduled time. The difference between the scheduled time and the arrival time of the first-arriving individual in each "departure unit" is defined as a tunable parameter called *Departure Delay*, or DD. Until a "departure unit's" scheduled departure time, future arriving individuals will join an existing "departure unit" if they share the destination Pixel of at least one TripNode in that "departure unit." A TripNode consists of a destination Pixel and the number of individuals who share that destination Pixel. Up to CD, or *Common Destination*, TripNodes may be added to any existing "departure unit." An arriving individual will create a new TripNode for a "departure unit" if the "departure unit" currently contains less than CD TripNodes, the addition of the TripNode causes no individual in the "departure unit" to travel more than *maxCircuity* miles over his or her straight-line origin-destination distance, and the conditions in Section 3.3.2.1 below hold. Note that the addition of any TripNode initiates reordering of all TripNodes in the "departure unit" to the sequence that minimizes the total miles traveled by the aTaxi, or VehicleMiles.

3.3.2.1 Forced Efficiency Improvement Conditions

Before checking for the following conditions, the system would occasionally assign shared rides when it was not optimal. Despite increasing the departure occupancy of an aTaxi, sharing rides in such cases would lead to greater overall vehicleMiles traveled. This translates to a lower *Average Vehicle Occupancy*, or AVO, which is defined as the total tripMiles to total vehicleMiles for any particular car, set of origin Pixels, or the entire state. Here, tripMiles refers to the sum of all fixed straight-line origin-destination distances of individuals, which does not change based on ride sharing. I use the updated ride constraints below to determine whether a second, third, or even higher-level TripNode should be included in the path of a given aTaxi. Thus, I use each updated ride constraint for the number of TripNodes in the set $\{2, 3, \ldots, CD\}$ for a given run whenever I investigate adding a second, third, $\ldots, CD$th TripNode, respectively, for a given aTaxi:

– 2nd TripNode:

  - $L_{0\to2} > L_{1\to2}$

– 3rd TripNode:

  - $L_{0\to2} > L_{1\to2}$
  - $L_{0\to3} > L_{2\to3}$
  - $L_{1\to3} + L_{0\to2} > L_{1\to2} + L_{2\to3}$

where the notation $L_{i\to j}$ represents the distance from the ith closest TripNode to the origin pixel to the jth closest TripNode to the origin pixel, and the 0th closest TripNode to the origin pixel is defined as the origin pixel itself. These conditions ensure that the AVO improves upon each new ride-share. Similar constraints follow for TripNodes beyond the 2nd and 3rd TripNodes.

Note also that under the current formulation of the autonomous taxi system for New Jersey, individuals whose origin pixels are identical to Pixels corresponding to intermediate TripNodes cannot be "picked up" by aTaxis visiting these intermediate TripNodes even if these individuals are waiting to depart at the times that any aTaxis arrive at the Pixels. In practice, aTaxis should have no difficulty dropping off each individual at his or her more specific destination; thus, rerouting to aTaxi stands mid-trip only when "fitting" both compromises the time of remaining passengers and is computationally very difficult. Additionally, such pickups could prevent further beneficial ride sharing at these Pixels.

3.4 Data Sets

I have two data sets of files filled with demands for taxi trips for all residents of New Jersey. The first is the previous existing data set from this fall's ORF 467 class. These demands for taxi trips were generated by processing former ORFE graduate student Talal Mufti's Module 7 NN files illustrating travel demand for each of close to nine million residents of New Jersey, plus close to half a million non-New Jersey residents who work in NJ, on a typical weekday. Mufti (2013) sampled from distributions created by 2010 U.S. Census, school, employment, patronage, and other data via population and trip synthesizers. Current ORFE graduate student Chenyi Chen wrote scripts that split all trips onto separate lines, then split trips by mode using reasonable assumptions and information on train station locations into train trips, walking (intrapixel) trips, "bicycling" trips for trips between adjacent Pixels, and finally, taxi trips. These taxi trips served as the foundation for my analysis of ride sharing in the infinite fleet case, which I discuss further in Section 3.5 below. They also served as an example of historical demands to inform my forecasts for "demand requests" in the future. Thus, I occasionally refer to this data set in the remainder of this thesis as "yesterday's" data set.

For my second data set—the "today" data set—I parsed through Module 6 NN files created by Wyrough (2014). The objective of Hill Wyrough's thesis was to extend the New Jersey state trip synthesizer model developed originally by Talal Mufti and the 2011 ORF 467 class, and enhanced by Jingkang Gao in 2013, to apply to any state within the U.S.

Mufti's old Module 7 NN files contained up to eight nodes for each individual indexed from 0 to 7, where, for Tour Types 19-20 as classified by Wyrough (2014), Node 7 departure times were simply set to -1 since Node 7 would refer to "Home." Hill's Module 6 NN files contained Nodes indexed from 1 to 7. For Tour Types 19-20, I calculated arrival times back "Home" implied by the final destination time. Parsing through Hill's Module 6 NN files was slightly more difficult than parsing through Mufti's Module 7 NN files due to the existence of strings such `NA`, `dTime`, and `oTime`.

Occasionally, I found negative departure times. I found that if a departure time were negative, then the trip is not so long such that the arrival time could possibly be ≥ 86400, or onto the next "day." I also assumed that if a departure time is ≥ 86400 (day), neither that departure time nor the corresponding arrival time at the next destination is $\geq 2 * 86400$ (2 days).

For "yesterday's," data set we had assigned 8 non-NJ "counties," namely, NYC (New York, New York), PHL (Philadelphia, Pennsylvania), BUC (Bucks County, Pennsylvania), SOU (some other locations south of NJ), NOR (some other locations north of NJ), WES (some other locations west of NJ), ROC (Rockland County, New York), and INTL (International). Before, we had 52 oTrips files from a list of 55 possible, the first 47 of which were NJ entries (of NewFIPS, splitting up some counties' oTrips files into multiple files due to file length for casual browsing so 21 became 47) and the rest were for these 8 non-NJ "counties". Now, the rest of the counties after the 47 NJ "sub-counties" are dynamically added (into NewFIPS).

I parsed individual trip demands from all non-NJ counties bordering NJ. The counties in PA that border NJ, from the south to the north, are: Delaware, Philadelphia, Bucks, Northampton, Monroe, and Pike. I took New Castle County to be the only bordering county in Delaware, since it is the only one you can drive directly across, though there is a ferry between Sussex County, DE and Cape May County, NJ. Finally, I parsed trip demands from the New York City coexisting counties Kings and Queens, as well as from bordering counties Orange, Rockland, Westchester, Bronx (Bronx), New York Manhatttan), and Richmond (Staten Island). I finally appended all files containing to trips originating in the same counties from all 4 states.

I only used "today's" data set for testing empty vehicle repositioning policies on a small 10 x 10 subset Pixel grid, which I define more precisely in Chapter 7, due to issues with complexity and dimensionality. This grid is entirely contained within New Jersey; thus, I did not yet deal with trips originating from out of state for Hill's data. However, if I were to use Hill's data as test data for empty vehicle repositioning policies that forecasted "demand requests" using assumed "demand request" rates from the ORF 467 data, I would assign the same set of "Parking Lot Pixels" that were assigned in the ORF 467 data set. The alternative of having any aTaxi that originates in NJ whose destination Pixel is out of state wait there until it is to come back to NJ on its next trip does not match the ORF 467 data. Furthermore, such a policy is incredibly complex. This is because finding the trip coming back requires looking in a taxi trips file for the county containing that particular Pixel, of which there are an enormous amount of such files. Figuring the ride sharing on that ride back would require running ride share analysis, too. In fact, if an out-of-state resident makes such a trip from NJ to an out-of-state destination, he or she may not even return to NJ within the day.

3.5 Analysis with an Infinite Fleet

Below I present a summary of results from ride share analysis in the infinite fleet case. Note that in the results below, not only is the fleet size assumed to be infinite, but the maximum number of passengers in any one vehicle is assumed to be infinite. Note that by definition, the total number of tripMiles always remains the same: 475,057,343 miles throughout the day. Figure 3.2 gives the total vehicleMiles traveled for each parameter set, and Figure 3.3 gives the AVO for each parameter set.

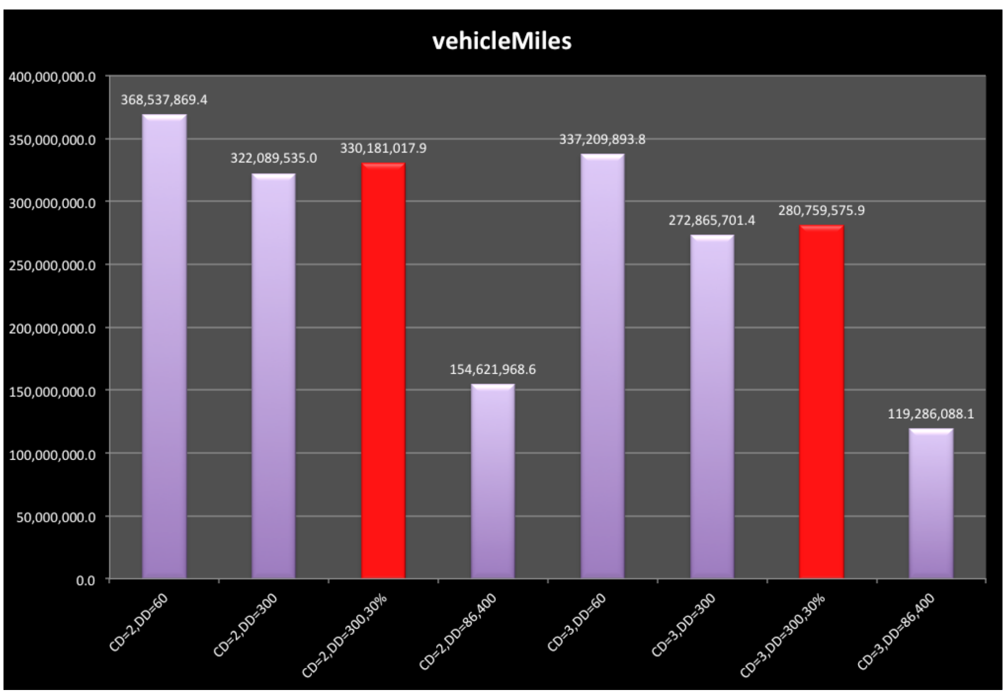


Figure 3.2: Total vehicleMiles in New Jersey for "yesterday's" data set in the infinite fleet case under various parameter sets. All parameter sets listed without a value of maxCircuity have an implied value of 20%.

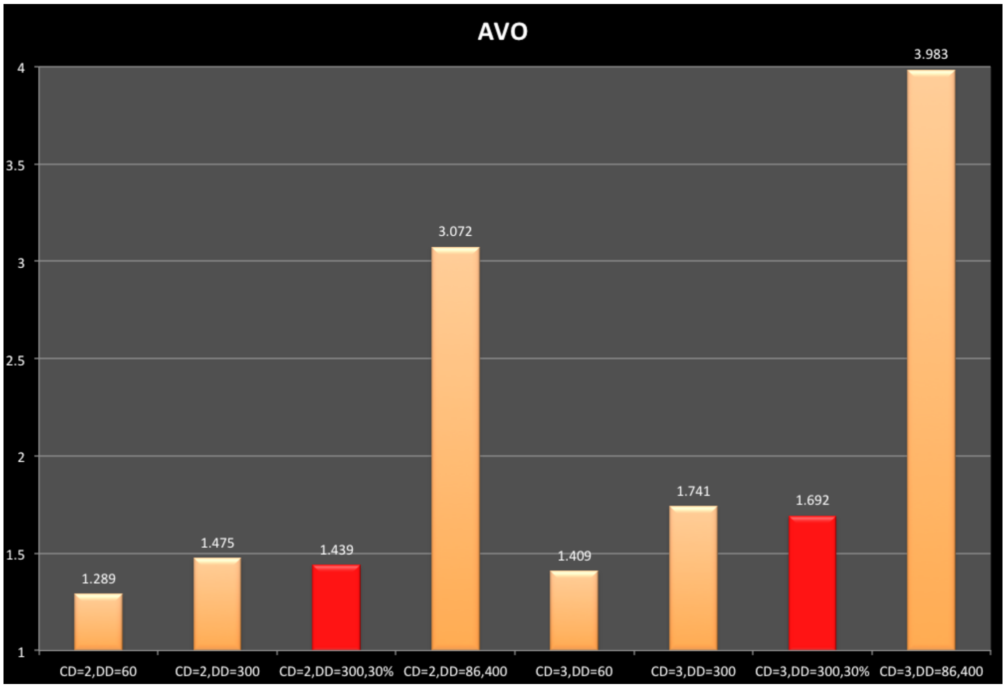


Figure 3.3: Average vehicle occupancy (AVO) in New Jersey for "yesterday's" data set in the infinite fleet case under various parameter sets. All parameter sets listed without a value of maxCircuity have an implied value of 20%.

Interestingly, the AVO for the increased allowable circuity, maxCircuity, to 30% actually decreased despite an increase in the average departure occupancy (ADO):

P2P; CD=3; DD=300

maxCircuity = 20%: AVO=1.741, ADO=2.892

maxCircuity = 30%: AVO=1.718, ADO=3.000

While this seems bizarre at first, the number of departures with occupancy of three and the tripMiles served for the travelers leaving in those aTaxis are both larger for the 30% run, but the vehicleMiles traveled by those departed aTaxis is also significantly larger, so much larger in fact that the AVO for the 30% run is actually lower. This makes sense; we allow trips to go "further from a direct path," and it turns out that an effect of this increased circuity is that the vehicle miles increase overall. This is not true for every single Pixel, and in fact, it is not even true for most pixels (11,658/21,642, or 54% of Pixels have AVO higher for the 30% run, to be exact). However, the 46% of pixels for which the AVO is higher for the 20% run make a larger impact, as they carry more weight.

Chapter 4

Departures, Arrivals, and Naïve Empty Vehicle Repositioning and Fleet-Sizing Policies

Every new beginning comes from some other beginning's end.

SEMISONIC

In this chapter I discuss the generation of Departures files, Arrivals files, and several naïve fleet-sizing policies. The fleet-sizing policies introduced in this section serve as starting points for further optimization in conjunction with empty vehicle repositioning policies, which are discussed in depth in Chapter 5.

4.1 Departures and Arrivals Files

During ride share analysis in the infinite fleet case for the parameter set $\{CD = 3, DD = 300 \text{ seconds}, \text{maxCircuity} = 20\%\}$, I created Departures files that list all aTaxi dispatches that depart from Pixels in a given county with complete details for each departure. I also created and sorted Arrivals files listing all aTaxi dispatches that arrive at final destination Pixels in a given county with the same relevant details as in the Departures files. These Departures files and Arrivals files contain the following header:

```
oCounty OriginX OriginY DepartTime TripNodeCount TripNode1:  DestX1 DestY1
Riders1 TripNode2:  DestX2 DestY2 Riders2 TripNode3:  DestX3 DestY3 Riders3
TotalRiders vehicleMiles tripMiles AVO VehTourEndCounty VehTourEnd_xPixel
VehTourEnd_yPixel VehTourEnd_Time
```

To calculate the arrival times, I assume that all individuals are dropped off at the centers of their destination Pixels, because "yesterday's" data set has origins and destinations stored only as Pixels, without latitude and longitude. However, aTaxis should have no difficulty dropping off each individual at his or her specific destination in a real-world implementation. I also assume that each aTaxi's average speed is 30 miles per hour. In reality, the average speed of a given aTaxi is random, though also likely positively correlated with vehicleMiles due to speed limits on highways versus more local roads. Furthermore, these arrival times are based on listed departure times, which are subject to delay due to lateness. Overall, the arrival times calculated in these files serve as reasonable estimates—but they are no more than just that.

4.2 Individuals and aTaxis in Transit

In order to find an absolute minimum fleet size and to understand the magnitude of reasonable fleet sizes for an autonomous taxi system, it is useful to plot the total number of aTaxis in transit versus time on historical data in the infinite fleet case. The maximum number of aTaxis on the move at any given time for the infinite fleet case—in which all "demand requests" are satisfied on time—may serve as an absolute minimum fleet size for the finite fleet case. Fleet sizes that are significantly larger than this absolute minimum are much more reasonable, especially given the uncertainty of "demand requests" in any given day. Note that the plots for individuals and aTaxis in transit in NJ below are based on "yesterday's" data set for the parameter set $\{CD = 3, DD = 300\,\text{seconds}, \text{maxCircuity} = 20\%\}$.

First, I assume that the maximum capacity of every aTaxi is six. Investigating other distributions of maximum aTaxi capacity is a logical next step for future research. Such distributions severely complicate the model that I describe in Chapter 5 and any policies for empty vehicle repositioning or fleet-sizing based on the model. Besides, processing each of the number of individuals and the number of aTaxis from the 10,479,382 departures from ORF 467 data set in the infinite fleet case, generated after processed for ride sharing, takes a significant amount of time.

Next, I make a simplifying assumption that all aTaxis that serve one "departure unit"—as listed in the Departures files described above—of greater than six individuals follow the same route. This is always true if there is only one TripNode. However, I assume that this is true even when there is more than one TripNode. Furthermore, I assume all of the individuals in such "departure units" take the "departure unit's" originally scheduled path. This means that all of the individuals whose destination Pixel is that of the original second TripNode ride in an aTaxi that visits the original first TripNode, followed by the original second TripNode. Similarly, all of the individuals whose destination Pixel is that of the original third TripNode ride in an aTaxi that visits the original first TripNode, followed by the original second TripNode, followed by the original third TripNode. In practice, it may be the case that after splitting up "departure units" that are too large for one aTaxi, one or more of the resultant "departure units" may be able to visit fewer TripNodes than they were originally scheduled to visit. I ignore such cases, which occur minimally and would likely have a negligible impact on results. My assumptions from the generation of the Departures and Arrivals files continue to hold about average aTaxi speed and dropping all individuals off at the center of—or at least the same relative position in—their destination Pixels.

To generate the aTaxi plots below, I create a two-column matrix with twice as many rows as departures. The first column contains the departure times or arrival times; the second column contains a value of $+1$ if the time to its left corresponds to a departure—which sends an aTaxi in transit—and value of -1 if the time to its left corresponds to an arrival. I sort this matrix by its first column, then overwrite the second column with cumulative sums of the $+1$ and -1 elements in time-ascending order. A plot of the first column versus the second column would produce the same plot as that below after adding in aTaxis in transit at midnight, as I discuss next. I simply cut down the number of elements at given threshold times.

Before plotting, I must also add in the number of aTaxis in transit at midnight, as the column of cumulative sums begins at 0.

In order to find the number of aTaxis in transit at midnight, I modified a script to count the number of aTaxis departing and arriving pre-midnight (i.e., before the time of 0 seconds) and post-midnight (i.e., after the time of 86,400 seconds) in the data. I "wrap" all departure times and arrival times to ensure that all times are within one full day. The numbers of aTaxis departing pre-midnight are likely due to either small errors in the original data set or extremely unlikely realizations of a person's first departure time of the day. Logically, the number of aTaxis in transit at midnight is the number of aTaxis arriving post-midnight, plus the number of aTaxis departing pre-midnight, minus the number of aTaxis departing post-midnight.

I perform similar procedures on the Departures files using the numbers of passengers in each "departure unit" in order to plot the number of *individuals* in transit over time. First, I make an *intentionally bad* assumption that all individuals in each "departure unit" travel from the origin Pixel to the final destination Pixel of the "departure unit. The maximum such "number" of individuals in transit in NJ occurs at 8:23:36 PM, as shown in Figure 4.1.

The stark difference between this plot and the corrected plot in Figure 4.2, in which individuals in different TripNodes arrive at their own destination Pixel, highlights how misleading *average departure occupancy (ADO)* can be. While the number of individuals set to depart in one aTaxi may be large, there is a good chance that a majority of these individuals travel a very short distance. Afterwards, perhaps only one or two passengers continue on a much longer journey that we assumed more passengers were on with the intentionally bad assumption. In fact, it turns out that working straight from the taxi trip files and shifting the straight-line origin-destination departure and arrival times forward by DD would be more accurate.

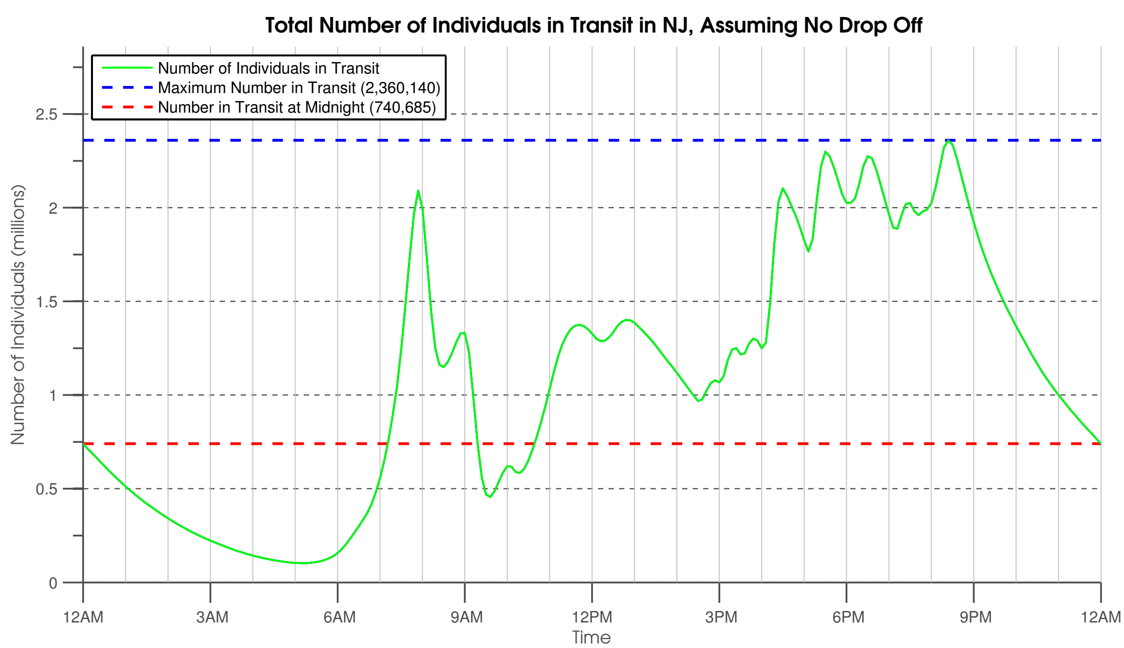


Figure 4.1: Number of individuals in transit in NJ if one were to assume no drop offs.

Figure 4.2 shows the corrected version of the number of individuals in transit in over time. In fact, in this plot, the maximum number of individuals in transit occurs at exactly 5:30 PM. Generating this plot takes an especially long time due to the extra processing required to handle three cases for the number of TripNodes for each aTaxi.

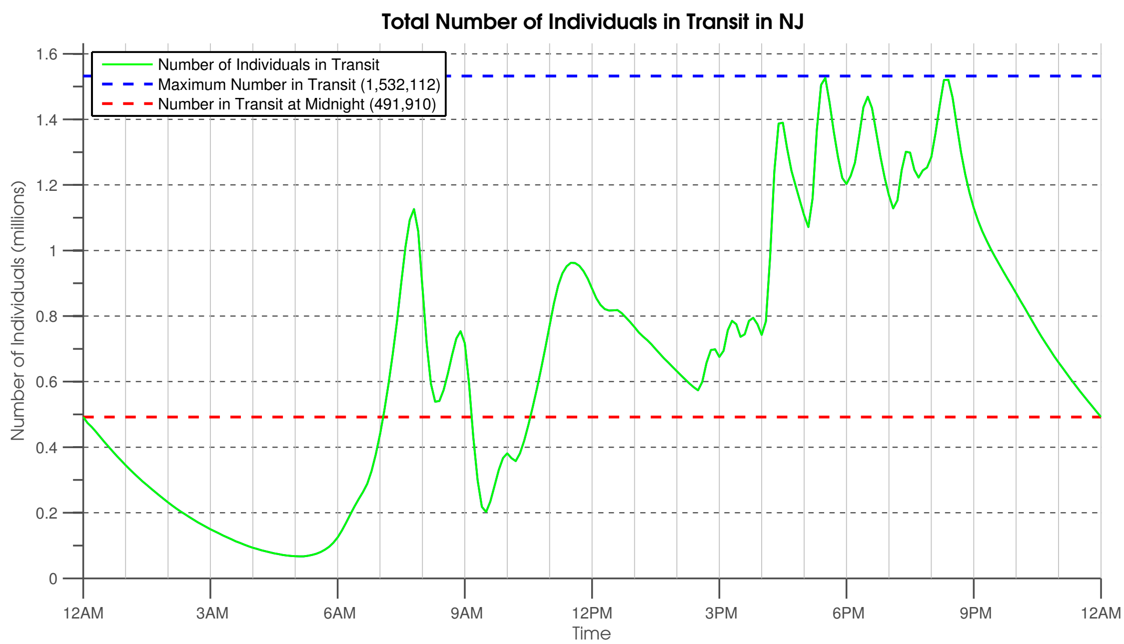


Figure 4.2: Corrected number of individuals in transit in NJ.

Finally, Figure 4.3 presents the number of aTaxis in transit over time; the maximum value occurs at 8:23:20 PM.

4.2.1 Instantaneous Average Vehicle Occupancy

The following instantaneous AVO values were generated by dividing the surplus of persons at a given time by the surplus of aTaxis at a given time, after adding in the number of persons and number of aTaxis in transit at midnight as described in the previous section.

Figure 4.4 presents the instantaneous AVO plot on the number of individuals in transit given the intentionally bad assumption. The point of the graph is to further show the extent of the problem with that assumption, and, thus, just how misleading ADO can be.

Figure 4.5 presents the corrected instantaneous AVO plot.

Compare this plot to the overall AVO of 1.741 under the parameter set

$\{CD = 3, DD = 300 \text{ seconds}, \text{maxCircuity} = 20\%\}$.

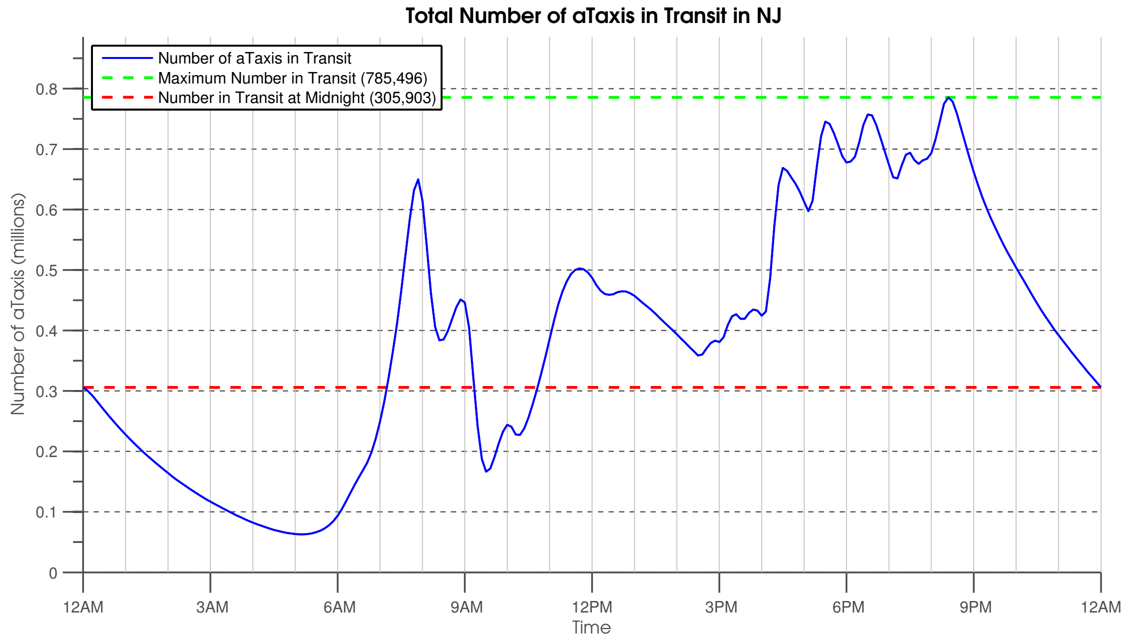


Figure 4.3: Number of aTaxis in Transit in NJ.

4.3 Use of Departures for Prediction of "Arrivals"

The prediction of future "demand requests" at each Pixel is, in fact, a prediction of future "departure units" of passengers who all depart as a unit in the same aTaxi because one "departure unit"—consists of no more than six passengers. In the mathematical model underlying these policies, I equate one "demand request" at each Pixel with the arrival of the first-arriving passenger in each "departure unit" that departs together in the same taxi. This simplifies the model by associating the demand for one available empty aTaxi with one "demand request." Furthermore, this eliminates the need for any ride sharing decisions in the model.

4.3.1 Poisson Processes for "Demand Requests"

There is strong evidence justifying the modeling of the "arrival" processes into Pixels as Poisson processes with time-varying, but discrete, arrival rates.

4.3.1.1 Empirical Distribution Functions

The distribution of arrival times within a fixed interval for any Poisson process is a uniform distribution, which gives a linear cumulative distribution function. Thus, identifying at least piecewise-

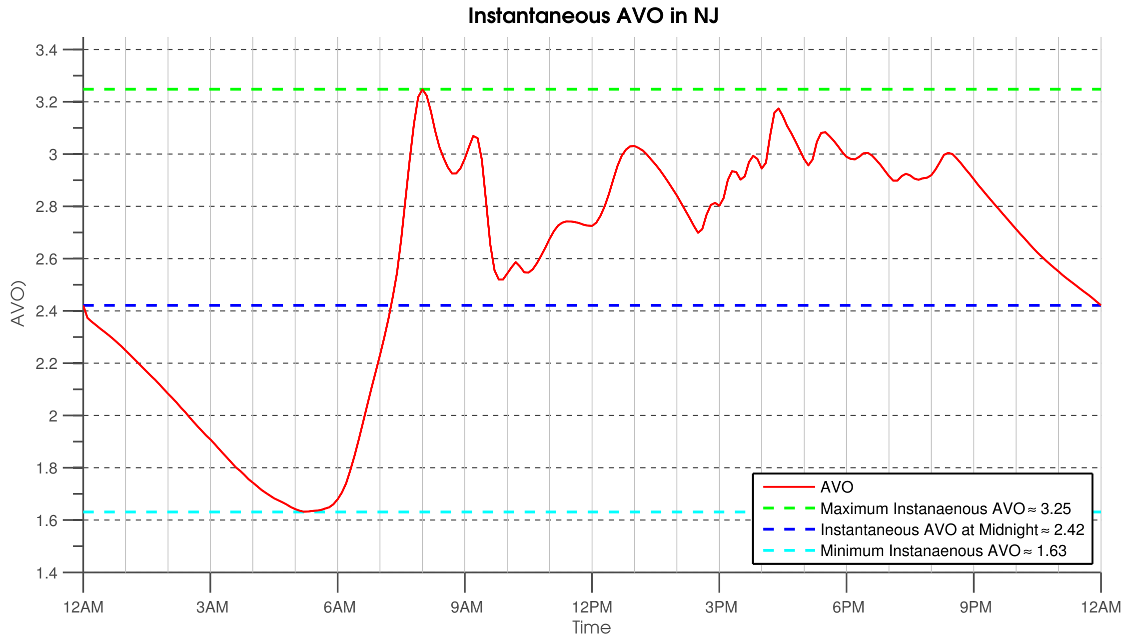


Figure 4.4: Instantaneous AVO if one were to assume no drop offs.

linear empirical distribution functions, or empirical CDFs, of "demand requests" at individual Pixels over the course of "yesterday" provides strong evidence that the arrivals of"demand requests" can be modeled as Poisson processes with constant rates over periods for which the empirical CDF is roughly linear.

By definition, the empirical CDF $\hat{F}_n(t) = \frac{\text{number of elements in sample} \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}}$.

Figure 4.6 displays sample empirical CDFs of "demand requests" at four Pixels in western Atlantic County. Even when there are fewer total demand requests, as there are in the bottom-right plot representing Pixel (74, 84), it is easy to see these empirical CDFs at most Pixels in the state indeed demonstrate piecewise-linear fits.

Theoretically, one could perform chi-square goodness-of-fit tests on exponential distributions for the interarrival times of "demand requests," but I will not present these tests. This is because in order to perform goodness-of-fit tests, one would need to identify all of the break points so that a single constant rate can be estimated for each piecewise-linear component. This is unnecessary. Besides, doing so would likely be overkill; transportation engineers and researchers already use Poisson processes extensively to model demand, as is discussed by Çinlar (2012).

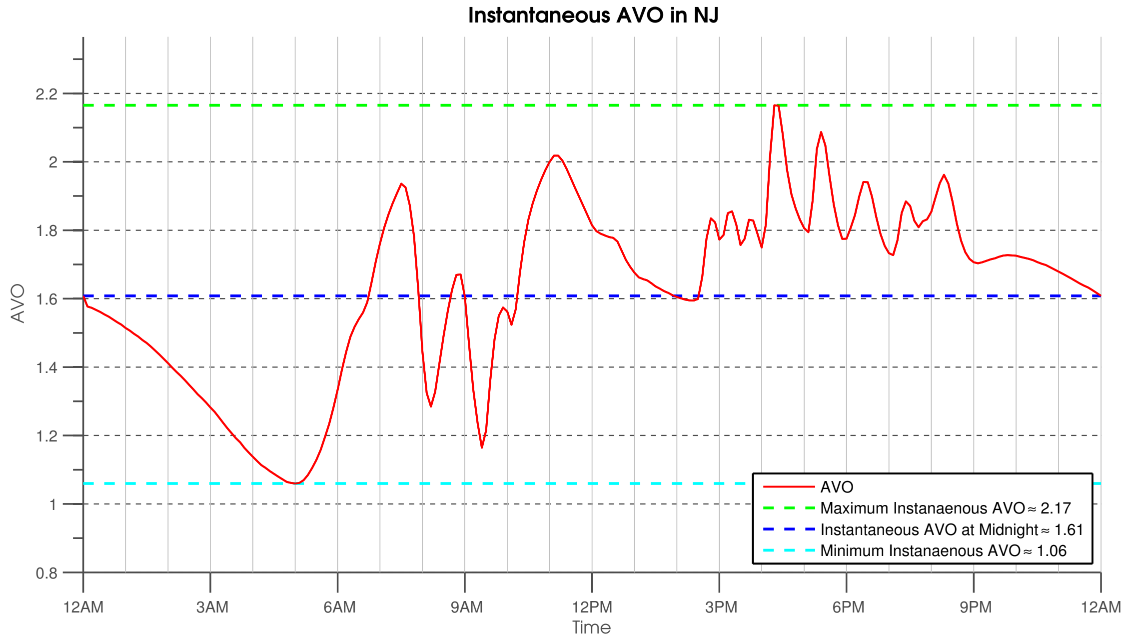


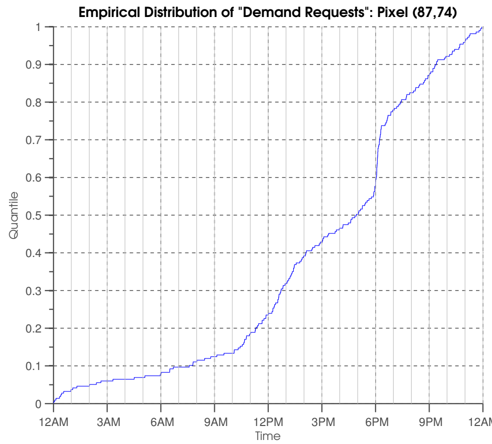
Figure 4.5: Corrected Instantaneous AVO.

4.3.1.2 "Demand Request" Rates Over Uniform Periods

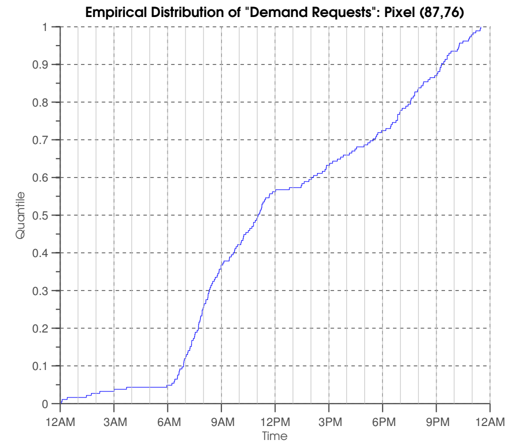
As discussed in Chapter 2, this policy for assigning is basic, but useful for this project because it has relatively significant predictive power, and I will discretize time anyway in order to make empty vehicle repositioning decisions. I found that a uniform period length of ten minutes is ideal. Figure 4.7 displays the number of "demand requests" in each ten minute period over the course of the day for the Pixel (120, 60) in Atlantic City. Each plot in the figure is based on "yesterday's" ORF 467 data set in the infinite fleet case for the parameter set $\{CD = 3, DD = 300\text{ seconds}, \text{maxCircuity} = 20\%\}$.

4.4 Naïve Empty Vehicle Repositioning Policy: "Everyone Goes Home"

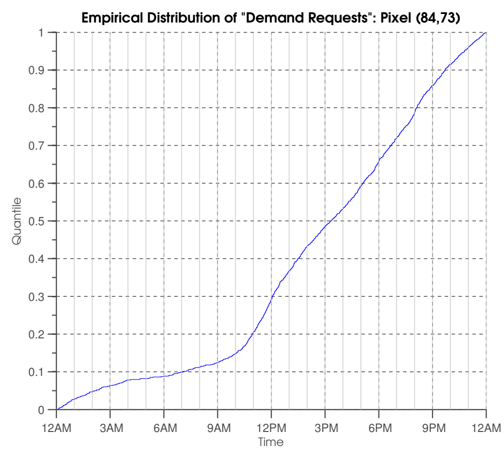
All empty vehicle repositioning policies discussed in this thesis, naïve or not, rely on assuming a fixed fleet size. Fleet-sizing policies can either be independent of these empty vehicle repositioning policies, or can use the results of the policies directly, as my fleet-sizing policy that I discuss in Section 6.2.



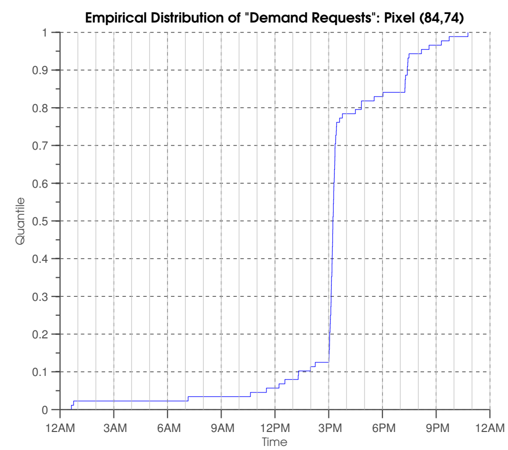
(a) $(74, 87)$



(b) $(76, 87)$



(c) $(73, 84)$



(d) $(74, 84)$

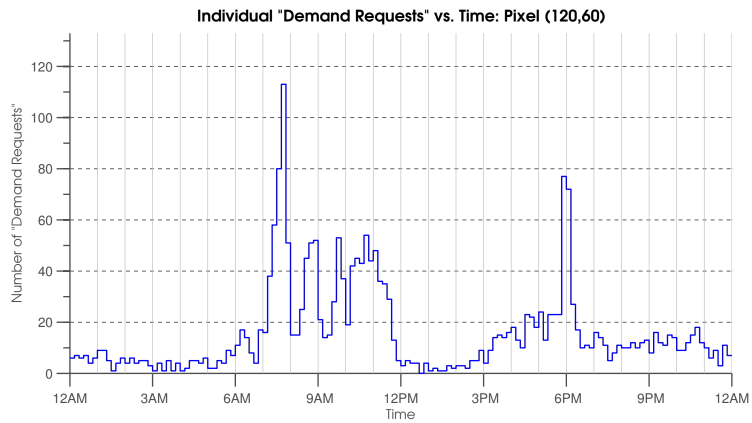
Figure 4.6: Empirical distribution functions (empirical CDFs) of “departure unit” “demand requests.” These four Pixels are located in the westernmost portion of Atlantic County, NJ, in the eastern suburbs of Vineland.

This naïve empty vehicle repositioning policy relies on the assumption inherent in our data sets that the “tour” taken by each individual—if any tour is taken—begins and ends at the same location, and thus the same Pixel, corresponding to his or her assigned home. That is, if any individual takes a strictly positive number of trips throughout the day, the origin Pixel of his or her first trip is his or her home Pixel, and the destination Pixel of his or her final trip is also his or her home Pixel.

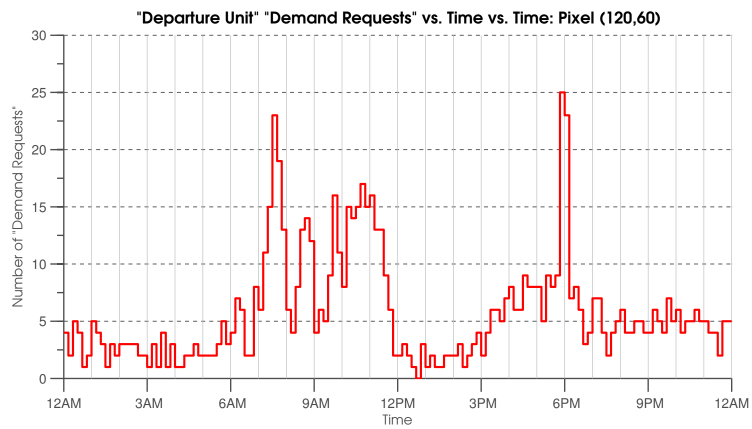
As a result, this policy attempts to distribute the empty vehicles to each Pixel late at night, or overnight, in order to maintain some level of aTaxis reasonably expected for that Pixel in the morning—say, the “departure units” due to 80% of the first individual trips demanded by all residents of that Pixel who demand at least one trip in “yesterday’s” data set.

4.5 Naïve Fleet-Sizing Policies

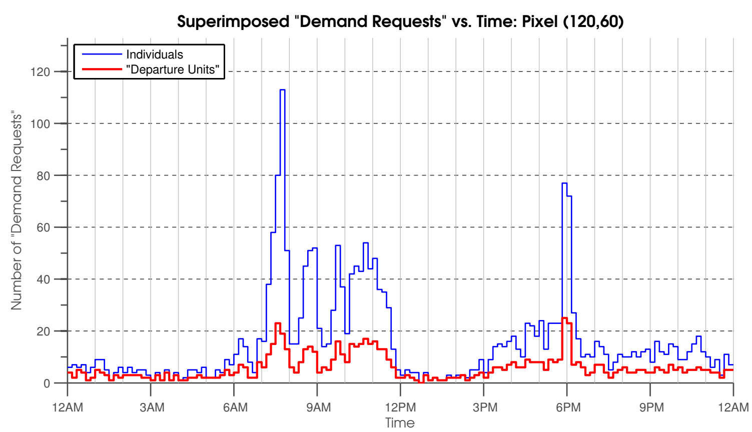
My final fleet-sizing policies are presented in Chapter 6. However, naïve fleet-sizing policies based on this chapter alone may include taking fleets at, say, 150%, 200%, 250%, and 300% of the “absolute minimum” fleet size described above.



(a) Individual "Demand Requests"



(b) "Departure Unit" "Demand Requests"



(c) Overlaid "Demand Requests"

Figure 4.7: "Demand requests" over uniform ten minute periods throughout the course of the day for the Pixel (120, 60) in Atlantic City. The plots are generated from "yesterday's" ORF 467 data set in the infinite fleet case.

Chapter 5

Empty Vehicle Repositioning and Fleet-Sizing: Mathematical Model

Change is the end result of all true learning.

Leo Buscaglia

Below I provide my complete mathematical model for the empty vehicle positioning problem specific to an autonomous taxi system like that discussed in Chapter 3, along with my corresponding assumptions and modeling decisions. This model is necessary for my Model-Based High-Complexity EVR Policy that I discuss further in Section 6.1.1, and any fleet-sizing policies that use the Model-Based High-Complexity EVR Policy as the class of policies over which the fleet size is varied. The model also serves as a theoretical foundation for all other empty vehicle repositioning policies applied to a general autonomous taxi system.

5.1 Assumptions and Modeling Decisions

For the purpose of notational convenience, I assume that time in my model, as detailed in Section 5.2 below, is split into discrete intervals of one second each. That is, I assume that the smallest granularity of time is seconds, such that any time referenced in my model is an integral number of seconds. This causes all indices of each vector in the model to be integral, which simplifies my notation as well as the mathematics overall. Using seconds as the smallest granularity is certainly reasonable; any further precision would likely have negligible effects on the performance of any policies based on the model.

Furthermore, I assume that the arrival of each loaded aTaxi at the aTaxiStand at its final destination Pixel occurs after a fixed "buffering" time, b, past its otherwise calculated arrival time. (Recall that the calculated arrival time of each aTaxi at its final destination is its departure time plus its vehicleMiles divided by its assumed average speed of 30 miles per hour, with units scaled appropriately. If an aTaxi departs "on time," i.e., at the time equal to the arrival time of the first-arriving customer in the "departure unit" at the aTaxiStand plus the fixed departure delay, then the arrival time of the loaded aTaxi at the aTaxiStand at its final destination Pixel will be b plus the calculated arrival time that would appear on an Arrivals file, as discussed in Chapter 4.) This attempts to account for the reality of aTaxis dropping off individuals at specific destinations. This is understandably imperfect because the number of specific destinations at intermediate or final destination Pixels—and these specific destinations themselves—vary randomly. However, this assumption is reasonable because each data set includes destinations only in the form of Pixels, and because it simplifies the model. Note that empty vehicles do not incur any such fixed "buffering" time.

I define the "assignment" of a given available empty vehicle to a given "departure unit" as an action that causes that empty vehicle to no longer be available, and ensures that this "departure unit" will dispatch in that particular empty vehicle. I assume that upon the arrival of the first-arriving individual in any "departure unit" at a given Pixel, that "departure unit" is instantaneously "assigned" to an empty vehicle at that Pixel if there exists one or more available empty vehicles at that Pixel. Otherwise, if there are no available empty vehicles at that Pixel upon this arrival, this "departure unit" is "assigned" to an empty vehicle at that Pixel as soon as one arrives and is not "assigned" to another "departure unit." "Departure units" accumulate a first-in, first-out queue at a Pixel if there are not available empty vehicles at the Pixel upon the arrivals of their respective first-arriving customers. Whenever there is a non-empty queue at a given Pixel, empty vehicles that arrive at that Pixel are "assigned" to "departure units" in ascending order of the arrival times of each "departure unit's" first-arriving customer. If a "departure unit's" time of "assignment" is prior to its scheduled departure time, then the empty vehicle waits at the origin Pixel until it dispatches at its scheduled departure time. Otherwise, if this time of "assignment" only occurs after its scheduled departure time, then the newly loaded vehicle is dispatched simultaneously.

I also assume that the difference between any two consecutive discrete decision times is no greater than the departure delay. If the difference between any two consecutive decision times were greater than the departure delay, then it would be possible for a relatively large number of "departure units" to appear and wait past their scheduled departure times before their "demand requests" are even recognized in the model, and thus before any decisions are made to reposition empty vehicles. This may occur at any Pixel with an insufficient number of empty vehicles stationed at that Pixel to which no "departure units" have been "assigned." Furthermore, this assumption somewhat simplifies the transition function provided in Section 5.2.4 below. Note that I hold constant the difference between any two consecutive discrete decision times in order to simplify the notation in the model. While in reality it may be useful to shorten the difference between consecutive discrete decision times during busier hours like rush hours—without drastically affecting the mathematics of the model—holding these differences constant is nevertheless a natural choice. I further discuss non-constant differences between consecutive discrete decision times as a possible next step of further research in Section 9.1.

Additionally, I assume that there is never any abandonment of individuals at any Pixel. I further discuss investigating the implications of time-varying abandonment rates on a similar model as another possible next step of further research in Section 9.1. Time-varying abandonment rates are an important facet of queueing theory, as discussed by Mandelbaum and Zeltyn (2004) and Brown et al. (2005); they are also an important facet of consumer psychology.

As previously mentioned in Sections 4.2 and 4.5, I assume that the maximum capacity of each aTaxi is fixed at six. This is due to the complexity of my model, empty vehicle repositioning policies, fleet-sizing policies, and implementation. Using alternate distributions of maximum capacities of different vehicles is an excellent candidate for further research, as I discuss in Section 9.1. However, I do not handle this maximum capacity directly in the model. As I discuss in Section 5.2.3 below, I consider all information from individual "demand requests" to be exogenous information; if an individual "demand request" meets ride sharing requirements for an existing "departure unit" consisting of six individuals, he or she is simply treated as the first-arriving member of a new "departure unit."

Finally, I assume that the deterministic travel times for each "departure unit" neither increase nor decrease at any time during the day due to accidents or congestion. In general, an autonomous taxi system should minimize accidents and congestions in practice.

5.2 Model

I present my mathematical model below in the form of a state variable, a decision variable, exogenous information, a transition function, and an objective function. Corresponding policies are discussed in Section 6.1. I use the the model framework proposed by Powell (2011) and supplemented neatly by Powell and Simão (2014), with a minor adjustment of the state variable requirements.

5.2.1 State Variable

The state variable includes the physical state of the network at a given time and any other information necessary to guide the evolution of the system:

$$\mathbf{S}_t = \Big(\left(x^e_{t'tij}\right)_{t'<t,\ i,j\in\mathcal{P}},\ \left(x^\ell_{tt''j}\right)_{t''>t,\ j\in\mathcal{P}},\ \left(s_{ti}\right)_{i\in\mathcal{P}},$$
$$\left(d_{t'tinjk}\right)_{t'<t,\ i,j\in\mathcal{P},\ n,k\in\mathbb{N}^+},\ \left(a_{t'tinjk}\right)_{t'<t,\ i,j\in\mathcal{P},\ n,k\in\mathbb{N}^+}\Big),\quad t\in\mathcal{T}$$

where:

- each element of $\mathbf{x}^{\mathbf{e}}_t$ is the number of empty vehicles that departed at time t' and are in transit from origin Pixel i to destination Pixel j at the given time t,
- each element of $\mathbf{x}^\ell_t$ is the number of loaded vehicles that are in transit at the given time t and will arrive at the aTaxi stand in destination Pixel j at time t'', i.e., including the additional fixed "buffering" time b,
- each element of $\mathbf{s}_t$ (the "supply") is the number of empty vehicles stationed at Pixel i to which no "departure units" have been "assigned" at the given time t,
- each element of $\mathbf{d}_t$, (the "demand") is an indicator variable taking on the value:
  - ∗ 1 if there exists a "departure unit" that arrived at time t' at origin Pixel i with ultimate (i.e., final) destination Pixel j and total travel time k, excluding the fixed "buffering" time b, that has neither been "assigned" to an aTaxi nor departed at the given time t, and its arrival time t' is the nth earliest of all existing "departure" units at Pixel i at

time t, the number of such existing "departure units" being the sum of the elements of $\mathbf{d}_{ti}$ and of $\mathbf{a}_{ti}$, i.e., $\|\mathbf{d}_{ti}\|_1 + \|\mathbf{a}_{ti}\|_1$, and

* 0 otherwise, and

- each element of $\mathbf{a}_t$ is an indicator variable taking on the value:

  * 1 if there exists a "departure unit" that arrived at time t' at origin Pixel i with ultimate (i.e., final) destination Pixel j and total travel time k, excluding the fixed "buffering" time b, that has already been "assigned" to an aTaxi but has not departed at the given time t, and its arrival time t' is the nth earliest of all existing "departure" units at Pixel i at time t, the number of such existing "departure units" being the sum of the elements of $\mathbf{d}_{ti}$ and of $\mathbf{a}_{ti}$, i.e., $\|\mathbf{d}_{ti}\|_1 + \|\mathbf{a}_{ti}\|_1$, and

  * 0 otherwise;

  all such "departure units" must have arrived at time $t - \tau$ at the earliest because they have not yet been dispatched.

$\mathcal{P}$ is the set of Pixels, $\mathcal{T}$ is the set of all discrete times in which decisions are made, i.e., the set of all discrete times spanning the day for which decisions are made, τ is the fixed departure delay, and the vectors $\mathbf{d}_{ti}$ and $\mathbf{a}_{ti}$ contain all elements of $\mathbf{d}_t$ and of $\mathbf{a}_t$, respectively, each with its third index equal to i representing the origin Pixel as defined above. For completeness, $\mathbb{N}^+$ is the set of all strictly positive integers (or natural numbers), i.e., $\mathbb{N}^+ = \{1, 2, 3, \ldots\}$. Furthermore, as discussed by Ahmadi (2014), the 1-norm of a vector $\mathbf{z} \in \mathbb{R}^n$, also known as the taxicab norm or the Manhattan norm, is defined as $\|\mathbf{z}\|_1 \coloneqq \sum_{i=1}^{n} |z_i|$, where z_i is the ith element of $\mathbf{z}$. All elements of $\mathbf{d}_{ti}$ and of $\mathbf{a}_{ti}$ are non-negative; thus, $|z_j| = z_j,\ \forall\, j \in \{1, 2, \ldots, m\}$ for $\mathbf{z} = \mathbf{d}_{ti} \in \mathbb{R}^m$ and $|z_j| = z_j,\ \forall\, j \in \{1, 2, \ldots, q\}$ for $\mathbf{z} = \mathbf{a}_{ti} \in \mathbb{R}^q$. Hence, the 1-norm serves as a useful shorthand notation here for the sum of all elements of each vector without summing over four other indices. Note that in general, the p-norm of a vector $\mathbf{z} \in \mathbb{R}^n$ is defined as $\|\mathbf{z}\|_p \coloneqq \left(\sum_{i=1}^{n} |z_i|^p\right)^{1/p}$, where z_i is similarly defined. As further discussed by Ahmadi (2014), a vector norm is a real-valued function $\|\cdot\| : \mathbb{R}^n \mapsto \mathbb{R}$ that satisfies the following properties:

- Positivity: $\|\mathbf{z}\| \geq 0\ \forall\, \mathbf{z} \in \mathbb{R}^n$ and $\|\mathbf{z}\| = 0 \iff \mathbf{z} = \mathbf{0}$,

- Homogeneity: $\|r\mathbf{z}\| = |r|\,\|\mathbf{z}\|\ \ \forall\, r \in \mathbb{R}, \mathbf{z} \in \mathbb{R}^n$, and

- Triangle Inequality: $\|\mathbf{y}+\mathbf{z}\| \leq \|\mathbf{y}\|+\|\mathbf{z}\| \ \forall\, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.

It is easy to check that, indeed, the 1-norm $\|\cdot\|_1 : \mathbb{R}^n \mapsto \mathbb{R}$ satisfies all three properties.

Recall that I assume that all times in the model are discretized with the smallest granularity being seconds. This includes all arrival times and travel times above. This assumption makes my notation more intuitive and the mathematics in the remainder of this chapter simpler to represent.

According to the model framework proposed by Powell (2011) and supplemented by Powell and Simão (2014) In general, the state variable is required to be "minimally dimensioned." This means that no components of the state variable may carry information that is not necessary for the evolution of the system—that is, for the transition function, objective function, or policies. Mathematically, the final index n on each of the vectors $\mathbf{d}_t, \mathbf{a}_t \ \forall\ t \in \mathcal{T}$ is not necessary. This is because each of these indices n can be determined by sorting all elements of both vectors $\mathbf{d}_{ti}, \mathbf{a}_{ti} \ \forall\ t \in \mathcal{T},\ \forall\ i \in \mathcal{P}$ by their indices t' as defined in the state variable above. However, without including these indices n, the transition function provided in Section 5.2.4 was even longer and messier than it is now.

5.2.2 Decision Variable

At a given time, the decision is how many empty vehicles to send from each Pixel to each other Pixel:

$$\mathbf{x}^{\mathbf{e}}_{tt} = \left(x^{e}_{ttij}\right)_{i,j\in\mathcal{P}}, \ \ t \in \mathcal{T}$$

where each element of $\mathbf{x}^{\mathbf{e}}_{tt}$ is, following the notation in the state variable, the number of empty vehicles chosen to depart at the given time t from origin Pixel i to destination Pixel j. The two indices t in the subscript exist to match the subscript structure necessary in the state variable's component of historical decisions, listed first in the definition of the state variable presented above in Section 5.2.1.

5.2.3 Exogenous Information

The exogenous information includes all random values necessary to guide the evolution of the system that are realized between each pair of discrete times. I assume that I know exactly how long it takes any aTaxi to move from a Pixel to any other Pixel. Thus, the only information-bearing events that can occur between two consecutive discrete times in $\mathcal{T}$, i.e., in (t_1, t_2), where $t_1, t_2 \in \mathcal{T}$ and $t_1 < t_2$, are arrivals, or "demand requests," of new individuals at Pixels. Each "demand request" of a new individual at a Pixel is of exactly one of four types:

`Type 1.` The individual *is the first-arriving member of a new "departure unit."* His or her "departure unit" may or may not be instantaneously "assigned" to an aTaxi at the individual's time of arrival.

`Type 2.` The individual *joins an existing "departure unit"* at the Pixel that has not yet departed and *requires an additional TripNode whose Pixel does not become the final destination.* Thus, the individual changes the total travel time for this "departure unit" but not its final destination. This "departure unit" may or may not have already been "assigned" to an aTaxi at the individual's time of arrival.

`Type 3.` The individual *joins an existing "departure unit"* at the Pixel that has not yet departed and *requires an additional TripNode whose Pixel becomes the final destination.* Thus, the individual changes both the total travel time for this "departure unit" and its final destination. This "departure unit" may or may not have already been "assigned" to an aTaxi at the individual's time of arrival.

`Type 4.` The individual *joins an existing "departure unit"* at the Pixel that has not yet departed and *does not require an additional TripNode.* Thus, the individual changes neither the total travel time for this "departure unit" nor its final destination. This "departure unit" may or may not have already been "assigned" to an aTaxi at the individual's time of arrival.

Individuals of `Type 4` do not affect the state variable nor any decisions dictated by any of the policies that I discuss in Section 6.1. I capture the realized information provided by the arrivals of individuals of the remaining three types through two components of the random vector of exogenous information:

$$\widehat{\mathbf{W}}_t = \left(\left(\tilde{d}^{\,1}_{t'tinjk} \right)_{t'<t,\ i,j\in\mathcal{P}},\ \left(\tilde{d}^{\,2}_{tinjk} \right)_{i\in\mathcal{P},\ n\in\mathbb{N}^+,\ k\in\mathbb{R}^+} \right),\ \ t \in \mathcal{T}$$

where:

- each element of $\tilde{\mathbf{d}}_t^{\mathbf{1}}$ is an indicator variable taking on the value:

  * 1 if at least one individual of `Type 1` belonging to one "departure unit" that did not exist at time $t - \Delta t$ arrived at any time $t' \in (t - \Delta t, t]$ at origin Pixel i with destination Pixel j, and the total travel time of this "departure unit" at time t is k at time t, and the arrival time of this "departure unit" is the nth earliest arrival time of all existing "departure units" at Pixel i at time t, the number of such existing "departure units" being the sum of the elements of $\mathbf{d}_{ti}$ and of $\mathbf{a}_{ti}$, i.e., $\|\mathbf{d}_{ti}\|_1 + \|\mathbf{a}_{ti}\|_1$, and
  * 0 otherwise, and

- each element of $\tilde{\mathbf{d}}_t^{\mathbf{2}}$ is an indicator variable taking on the value:

  * 1 if at least one individual of `Type 2` or of `Type 3` belonging to one "departure unit" that existed at time $t - \Delta t$ at the origin Pixel i, and exists there at time t, arrived at any time in the interval $(t - \Delta t, t]$, the last of which changes the total travel time to k of its "departure unit," and the possibly new destination Pixel of this "departure unit" due to the arrival of the last individual of `Type 3`, if any, is j, and the arrival time of this "departure unit" is the nth earliest arrival time of all existing "departure units" at Pixel i at time t, the number of such existing "departure units" being the sum of the elements of $\mathbf{d}_{ti}$ and of $\mathbf{a}_{ti}$, i.e., $\|\mathbf{d}_{ti}\|_1 + \|\mathbf{a}_{ti}\|_1$, and
  * 0 otherwise.

Note that the time increment Δt is the constant difference between any two consecutive decision times; in order to simplify the model, I assume that Δt is an integral number of seconds.

5.2.4 Transition Function

Let $\tilde{t}$ be the maximal element of $\mathcal{T}$, i.e., $\tilde{t} \in \mathcal{T}$ and $\tilde{t} \leq t \implies \tilde{t} = t, \forall\, t \in \mathcal{T}$. The components of the transition function:

$$\mathbf{S}_{t+\Delta t} = \mathbf{S}^M\left(\mathbf{S}_t,\ \mathbf{x}_t^{\mathbf{e}},\ \widehat{\mathbf{W}}_{t+\Delta t}\right),\ \ t \in \mathcal{T} \setminus \{\tilde{t}\}$$

are:

$$x^e_{t'(t+\Delta t)ij} = x^e_{t'tij}\, \mathbb{1}_{\left\{\left\lceil \frac{\delta_{ij}}{a} \right\rceil > t-t'+\Delta t\right\}}, \qquad \forall\, t' < t+\Delta t,\ \forall\, i,j \in \mathcal{P} \tag{5.1}$$

$$\begin{aligned} x^\ell_{(t+\Delta t)t''j} = &\left(x^\ell_{tt''j} + \sum_{t' \le t-\tau+\Delta t} \sum_{i \in \mathcal{P}} \left(\sum_{\left\{n:\, \left(E^2_{tin}\right)'\right\}} \left(a_{t'tinj\underline{t}} + d_{t'tinj\underline{t}} \mathbb{1}_{E^1_{t'ti}} \right) \right.\right. \\ &\left.\left. + \sum_{\left\{n:\, E^2_{tin},\, E^3_{t'tin}\right\}} \left(\tilde{d}^2_{(t+\Delta t)inj\underline{t}} \right) \right) \mathbb{1}_{\{t''>t+\Delta t\}} \right), \ \forall\, t'' > t,\ \forall\, j \in \mathcal{P} \end{aligned} \tag{5.2}$$

$$\begin{aligned} s_{(t+\Delta t)i} = &\left(s_{ti} + \sum_{t' \in \mathcal{T}\,'} \sum_{j \in \mathcal{P}} x^e_{t'tji} + \sum_{t'' \in \mathcal{T}\,''} x^\ell_{tt''i} - \sum_{t' < t+\Delta t} \sum_{n \in \mathbb{N}^+} \sum_{j \in \mathcal{P}} \sum_{k \in \mathbb{N}^+} \tilde{d}^1_{t'(t+\Delta t)injk} \right. \\ &\left. - \sum_{t' < t} \sum_{n \in \mathbb{N}^+} \sum_{j \in \mathcal{P}} \sum_{k \in \mathbb{N}^+} d_{t'tinjk} - \sum_{j \in \mathcal{P}} x^e_{ttij} \right)^+, \qquad \forall\, i \in \mathcal{P} \end{aligned} \tag{5.3}$$

$$\begin{aligned} d_{t'(t+\Delta t)injk} = &\left(d_{t'tin'jk}\, \mathbb{1}_{\left(E^2_{tin}\right)'} \mathbb{1}_{\{n'=n+D_{ti}\}} + \left(\tilde{d}^1_{t'(t+\Delta t)injk} + \tilde{d}^2_{(t+\Delta t)injk} \right) \right) \mathbb{1}_{\left(E^1_{t'ti}\right)'}, \\ &\forall\, t' < t+\Delta t,\ \forall\, i,j \in \mathcal{P},\ \forall\, n,k \in \mathbb{N}^+ \end{aligned} \tag{5.4}$$

$$\begin{aligned} a_{t'(t+\Delta t)injk} = &\left(a_{t'tin'jk} + d_{t'tin'jk} \mathbb{1}_{E^1_{t'ti}} \right) \mathbb{1}_{\left(E^2_{tin}\right)'} \mathbb{1}_{\{n'=n+D_{ti}\}} \\ &+ \left(\tilde{d}^1_{t'(t+\Delta t)injk} + \tilde{d}^2_{(t+\Delta t)injk} \right) \mathbb{1}_{E^1_{t'ti}}, \quad \forall\, t' < t+\Delta t,\ \forall\, i,j \in \mathcal{P},\ \forall\, n,k \in \mathbb{N}^+ \end{aligned} \tag{5.5}$$

$$\text{each} \quad \forall\, t \in \mathcal{T} \setminus \left\{ \tilde{t} \right\}$$

where:

- M represents my model,
- δ_{ij} is the (deterministic) distance in miles from Pixel i to Pixel j,
- a is the average aTaxi speed in miles per unit time, which all aTaxis are assumed to have,
- $\lceil u \rceil$ is the smallest integer not less than u, $\forall\ u \in \mathbb{R}$, as used in Equation 5.1 and in the definition of the set $\mathcal{T}\,'$ below,
- $\underline{t} = t'' - t' - \tau - b$, as used in Equation 5.2,
- the set $\mathcal{T}\,' = \left\{ t - \left\lceil \frac{\delta_{ij}}{a} \right\rceil + 1,\, t - \left\lceil \frac{\delta_{ij}}{a} \right\rceil + 2,\, \ldots,\, t - \left\lceil \frac{\delta_{ij}}{a} \right\rceil + \Delta t \right\}$ as used in Equations 5.2, 5.3, and 5.4,
- the set $\mathcal{T}\,'' = \{t+1,\, t+2,\, \ldots,\, t+\Delta t\}$ as used in Equations 5.2, 5.3, and 5.4,

- $(u)^+ := \max\{0, u\}$, $\forall\, u \in \mathbb{R}$, as used in Equation 5.3,
- the event $E^1_{t'ti} = \left\{ s_{ti} + \sum_{t''' \in \mathcal{T}'} \sum_{j \in \mathcal{P}} x^e_{t'''tji} + \sum_{t''' \in \mathcal{T}''} x^\ell_{tt'''i} - \sum_{j \in \mathcal{P}} x^e_{ttij} \geq \sum_{t'' \leq t'} \sum_{n \in \mathbb{N}^+} \sum_{j \in \mathcal{P}} \sum_{k \in \mathbb{N}^+} d_{t''tinjk} \right\}$, which is the event that there are enough arrivals of vehicles into Pixel i at times in the set $\{t+1, t+2, \ldots, t+\Delta t\}$ that any "departure units" with initial arrival times of their first passengers no greater than t' that were not assigned at time t have been assigned at time $t + \Delta t$, whether or not they have been dispatched, as used in Equations 5.2 and 5.4,
- the event $E^2_{tin} = \left\{ \sum_{j' \in \mathcal{P}} \sum_{k \in \mathbb{N}^+} \tilde{d}^{\,2}_{(t+\Delta t)inj'k} > 0 \right\}$ as used in Equations 5.2 and 5.4,
- the event $E^3_{t'tin} = \left\{ \sum_{j' \in \mathcal{P}} \sum_{k \in \mathbb{N}^+} \left(a_{t'tinj'k} + d_{t'tinj'k} \mathbb{1}_{E^1_{t'ti}} \right) > 0 \right\}$ as used in Equation 5.2, and
- the quantity $D_{ti} = \min \left\{ \sum_{t'' \leq t - \tau + \Delta t} \sum_{n \in \mathbb{N}^+} \sum_{j \in \mathcal{P}} \sum_{k \in \mathbb{N}^+} \left(d_{t''tinjk} + a_{t''tinjk} \right),\ s_{ti} + \sum_{t'' \in \mathcal{T}'} \sum_{j \in \mathcal{P}} x^e_{t''tji} + \sum_{t'' \in \mathcal{T}''} x^\ell_{tt''i} - \sum_{j \in \mathcal{P}} x^e_{ttij} \right\}$, as used in Equations 5.4 and 5.5. This is the total number of loaded vehicle departures from origin Pixel i at any time $t''' \in \mathcal{T}''$. Here, $s_{ti} + \sum_{t'' \in \mathcal{T}'} \sum_{j \in \mathcal{P}} x^e_{t''tji} + \sum_{t'' \in \mathcal{T}''} x^\ell_{tt''i} - \sum_{j \in \mathcal{P}} x^e_{ttij} \geq 0$ and thus does not require surrounding $(\cdot)^+$ notation because all terms on the left-hand side are non-negative except for $-\sum_{j \in \mathcal{P}} x^e_{ttij}$, and $\sum_{j \in \mathcal{P}} x^e_{ttij} \leq s_{ti}$ because the number of dispatched empty vehicles at any Pixel at any given time cannot be any greater than the number of empty vehicles stationed at that Pixel to which no "departure units" have been "assigned" at the given time (i.e., the "supply"). This fact is further elucidated in each of my empty vehicle repositioning policies that I discuss further in Section 6.1.

Recall that I assume that the (constant) difference between consecutive discretized decision times is no greater than the departure delay. Thus, no loaded vehicles will be dispatched at any given time $t \in \mathcal{T}$ that were not identified in $\mathbf{d}_{t-\Delta t}$ nor in $\mathbf{a}_{t-\Delta t}$ (which are given at the start for the minimal element $\check{t}$ of $\mathcal{T}$, i.e., $\check{t} \in \mathcal{T}$ and $\check{t} \geq t \implies \check{t} = t$, $\forall\, t \in \mathcal{T}$).

5.2.5 Objective Function

Let the cost function, with a given state $\mathbf{S}_t$ and given policy π at a given time t be defined as:

$$C_t \left(\mathbf{S}_t,\, X^\pi \left(\mathbf{S}_t \right) \right) = \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{P}} \left(c\, \delta_{ij}\, x^e_{ttij} + z \sum_{t' \leq t - \tau} \sum_{n \in \mathbb{N}^+} \sum_{k \in \mathbb{N}^+} d_{t'tinjk} \right), \quad t \in \mathcal{T}$$

where:

- $X^{\pi}(\mathbf{S}_t)$ is the decision function at time t, i.e., $X^{\pi}(\mathbf{S}_t) = \mathbf{x}^{\mathbf{e}}_{tt}$, the decision variable at time t,

- c is a tunable parameter representing (expected) cost per mile, which may be estimated by the (expected) cost of fuel per gallon of gasoline divided by (expected) fuel efficiency in miles per gallon of gasoline for gasoline-powered aTaxis, or by the (expected) cost of electricity per kilowatt-hour divided by (expected) fuel efficiency in miles per kilowatt-hour for electric vehicle aTaxis, or, for example, by some reasonable combination of the two for plug-in hybrid electric vehicle aTaxis—all at the (expected) time of implementation of the system,

- z is a tunable parameter representing the estimated "cost" per unit time of lateness per "departure unit" in the same currency units as that for the cost per mile c as described above

- the elements of $\mathbf{x}^{\mathbf{e}}_{tt}$ and of $\mathbf{d}_t$ $\forall\, t > \tilde{t}$, where $\tilde{t}$ is the maximal element of $\mathcal{T}$, i.e., t'', $\tilde{t} \in \mathcal{T}$ and $\tilde{t} \leq t'' \implies \tilde{t} = t''$, $\forall\, t'' \in \mathcal{T}$, and $\mathcal{T}$ contains discrete decision times spanning exactly one full day, correspond to values taken on at the equivalent times of $t' - T$ of the following day, where T is one day expressed in the same units of time as t', and

Now let the objective function, J, to be minimized for a given policy π, be defined as:

$$\begin{aligned} J &= \mathbb{E}\left\{\sum_{t\in\mathcal{T}} C_t\Big(\mathbf{S}_t,\, X^{\pi}(\mathbf{S}_t)\Big)\right\} \\ &= \mathbb{E}\left\{\sum_{t\in\mathcal{T}}\left(\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{P}}\Big(c\,\delta_{ij}\,x^{e}_{ttij} + z\sum_{t'\leq t-\tau}\sum_{n\in\mathbb{N}^{+}}\sum_{k\in\mathbb{N}^{+}} d_{t'tinjk}\Big)\right)\right\} \end{aligned}$$

Presented as a minimization over feasible policies, then, the objective function is:

$$\begin{aligned} \min_{\pi} J &= \min_{\pi}\ \mathbb{E}\left\{\sum_{t\in\mathcal{T}} C_t\Big(\mathbf{S}_t,\, X^{\pi}(\mathbf{S}_t)\Big)\right\} \\ &= \min_{\pi}\ \mathbb{E}\left\{\sum_{t\in\mathcal{T}}\left(\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{P}}\Big(c\,\delta_{ij}\,x^{e}_{ttij} + z\sum_{t'\leq t-\tau}\sum_{n\in\mathbb{N}^{+}}\sum_{k\in\mathbb{N}^{+}} d_{t'tinjk}\Big)\right)\right\} \end{aligned}$$

Put simply, the objective is to find the policy π that minimizes the expectation of the sum of the cost incurred at all decision times. The cost incurred at each decision time is the sum of the travel costs of sending all dispatched empty vehicles to their destinations and a "lateness penalty" for all "departure units" that were scheduled to dispatch before the decision time, but have not yet departed. The policy refers to an empty vehicle repositioning policy that defines the decision function $X^{\pi}(\mathbf{S}_t)$ with a fixed fleet size, which may be determined separately.

Note that in this objective function, lateness penalties are assigned to "departure groups" proportional to time but independently of the number of passengers in each "departure unit." Assigning lateness penalty independently of the number of passengers in each "departure unit" significantly simplifies the model. It is also reasonable for several reasons. Firstly, the instantaneous average departure occupancy at even the busiest times of the day in relatively urban areas is rarely high. Secondly, lateness penalties begin after the departure delay past only the first-arriving passenger's arrival time of a given "departure unit," and it is often the case that the arrival time at the origin Pixel of an aTaxi that is late for a given "departure unit" would not be late for some or all individuals who arrive after the first-arriving passenger in the "departure unit," given that any exist. Finally, in practice, the maximum capacity of any aTaxi—and thus, equivalently, of any "departure unit"—is finite and likely low; in the policies that I implement in this thesis, I set this maximum capacity to six for all aTaxis and thus all "departure units." Thus, even in the rare worst cases—i.e., even when over a hundred individuals would otherwise qualify to share a ride in no more than the maximum number of TripNodes for an aTaxi according to the conditions outlined in Chapter 3—no more individuals than this maximum capacity may be waiting for any single late aTaxi.

Chapter 6

Empty Vehicle Repositioning and Fleet-Sizing: Policies

Do not wait; the time will never be 'just right.'

NAPOLEON HILL

In this chapter, I present different policies for empty vehicle repositioning and fleet-sizing.

6.1 Empty Vehicle Repositioning (EVR) Policies

I previously provided support for time-dependent Poisson processes of "demand requests" at each Pixel via empirical cumulative distribution functions. With these expected rates for forecasting demands, I present empty vehicle repositioning policies below.

6.1.1 Model-Based High-Complexity EVR Policy

This is a lookahead policy based directly on my mathematical model. There is a motivation to limit the length of the lookahead horizon due to computational complexity and uncertainty in arrivals. However, it is important that the choice of the length of the lookahead horizon forward to look is related to travel time. Perhaps, it is a good choice is to take a smallest lookahead horizon that is greater than the average travel time. A myopic policy based on my model would be trivial and would send little or no empty vehicles according to the two-part cost function; it would not be robust at all. In order to feel the effects of the lateness penalty, one must look ahead.

In general, lookahead policies are simple and usually provide better approximations than pure myopic policies. However, lookahead policies are"brute force" policies that effectively require searching the tree of all possible outcomes, causing decisions to be relatively expensive. In fact, it is possible that computation times grow exponentially with the length of the lookahead horizon.

I present my Model-Based High-Complexity EVR Policy for empty vehicle repositioning below. The policy assumes a fixed finite fleet size, relies on the successive minimization of the expectation of the same objective function with the same mathematical model through an LP. However, each policy uses different assumptions for the arrival rates of demand requests into each Pixel. This objective function is a new objective function defined similarly to that in Section 5.2.5, but uses a different contribution function that replaces the cost function defined above.

Let a contribution function $C_{t,t+l}$ be defined as the sum of $l+1$ individual cost functions evaluated at $l+1$ times:

$$
\begin{aligned}
C_{t,t+l} &= \sum_{t'=t}^{t+l} C_{t'}\left(\mathbf{S}_{t'},\, X^{\pi}\left(\mathbf{S}_{t'}\right)\right) \\
&= \sum_{t'=t}^{t+l}\left(\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{P}}\left(c\,\delta_{ij}\,x^{e}_{t't'ij} + z\sum_{t''\le t'-\tau}\sum_{n\in\mathbb{N}^{+}}\sum_{k\in\mathbb{N}^{+}} d_{t''t'injk}\right)\right), \quad t\in\mathcal{T}
\end{aligned}
$$

where l is the (tunable) number of discrete time periods designated to *look ahead* in a given policy.

Simply put, the policy is to minimize the expectation of the sum of these contribution functions at all decision times, i.e., to minimize $\mathbb{E}\sum_{t\in\mathcal{T}} C_{t,t+l}$ subject to all 5 components of the transition function presented in Section 5.2.4.

Note that the fleet size is fixed, and it is not explicitly defined. Rather, it is implicitly defined by the given state at the first decision time. The state identifies the number of empty vehicles currently in transit, the number of loaded vehicles currently in transit, the number of available empty vehicles at each Pixel, and the number of "departure units" to which aTaxis have been assigned but have not yet departed. Logically, the total fleet size is the sum of each of these values because every aTaxi in the system at any given time must either be in transit, available, or assigned but not yet departed. If we define the fixed fleet size as F, then:

$$
F = \sum_{t'<v}\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{P}} x^{e}_{t'vij} + \sum_{t''>t}\sum_{j\in\mathcal{P}} x^{\ell}_{vt''j} + \sum_{i\in\mathcal{P}} s_{vi} + \|\mathbf{a}_v\|_1
$$

where v is the first decision time. Furthermore, the components of the transition function—Equations 5.1, 5.2, 5.3, 5.4, and 5.5—ensure that:

$$\sum_{t' < t+\Delta t} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{P}} x^{e}_{t'(t+\Delta t)ij} + \sum_{t'' > t+\Delta t} \sum_{j \in \mathcal{P}} x^{\ell}_{vt''j} + \sum_{i \in \mathcal{P}} s_{(t+\Delta t)i} + \|\mathbf{a}_{(\Delta t)}\|_1$$
$$= \sum_{t' < t} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{P}} x^{e}_{t'tij} + \sum_{t'' > t} \sum_{j \in \mathcal{P}} x^{\ell}_{tt''j} + \sum_{i \in \mathcal{P}} s_{ti} + \|\mathbf{a}_t\|_1, \ \forall\, t \in \mathcal{T} \setminus \{\tilde{t}\}$$

By a simple inductive argument, it is easy to see that the the fleet size implied by the system state is maintained over any time period in which decisions are made.

6.1.2 Simplified EVR Policy: Norm-Based Penalty

This policy is a myopic policy in which the objective is to minimize the sum of transporting empty vehicles and some tuning parameter z times the norm of the difference between a vector of the distribution of empty vehicles plus the effects on that vector by the decisions and a vector representing an "ideal" distribution of empty vehicles, scaled to the number of empty vehicles currently in the system.

As you can imagine, this policy is incredibly sensitive to z (relative to c). One may tune it by fixing c and scaling z relative to c such that the magnitude of the penalty term is reasonably on the same order as that of the direct cost of empty vehicle trips. One may look at the value of the z times the norm (post-decision) relative to the optimal value at typical decision times that are sufficiently far from the first decision time, and attempt to maximize the ratio of this value to the optimal value at an average time step. In general, increasing z decreases the ratio of z times the norm to the optimal value, because a larger value of z magnifies the incentive to send empty vehicles to minimize otherwise severe penalties due to z times the norm. But this is only true up to a certain point; past a certain point, increasing z causes the "cost" due to the norm grow faster than the empty vehicle repositioning costs can "catch" it. At the first decision time, it is likely that a significantly greater number of empty vehicles will be sent compared to at similar decision times around it, during which similar demand request rates are assumed.

There are two takes on this policy: one that uses the 1-norm and one that uses the 2-norm, in which one must scale appropriately between the two norms. In my attempt to implement this policy on a 10 x 10 subset Pixel grid in New Jersey, I fixed z for the 1-norm and scaled z up appropriately according to the value of z for the 2-norm that would set the initial 2-norm equal to the initial 1-norm pre-decisions (i.e., with all decisions set to 0) according to the initial distribution that I outlined above and the demand requests arriving in the first 60 seconds for my 10 x 10 Pixel grid.

The 1-norm was defined earlier in Section 5.2.1. As discussed by Ahmadi (2014), the 2-norm, or Euclidean norm, of a vector $\mathbf{x} \in \mathbb{R}^n, n \in \mathbb{N}^+$ is defined as the square root of the Euclidean inner product of the vector $\mathbf{x}$ with itself. That is,

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i}$$

where x_i is the ith element of $\mathbf{x}$.

Note that $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2 \; \forall \; \mathbf{x} \in \mathbb{R}^n, \forall \; n \in \mathbb{N}^+$. This is why z must be scaled up, versus down, appropriately for the convex optimization problem at each decision time using 2-norm. A simple, informal proof of the equivalent claim $\|\mathbf{x}\|_1^2 \geq \|\mathbf{x}\|_2^2 \; \forall \; \mathbf{x} \in \mathbb{R}^n$ for any given value of $n \in \mathbb{N}^+$ is as follows:

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Then, by the definitions of the 1-norm and of the 2-norm:

$$\begin{aligned}
\|\mathbf{x}\|_1^2 &= \Big(|x_1| + |x_2| + \ldots + |x_n|\Big)^2 \\
&= x_1^2 + x_2^2 + \ldots + x_n^2 + C \geq x_1^2 + x_2^2 + \ldots + x_n^2 = \left(\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}\right)^2 = \|\mathbf{x}\|_2^2
\end{aligned}$$

since the term $C \geq 0$, where C represents all terms $|x_i||x_j|$, $i \neq j$, $\forall \; i, j \in \{1, 2, \ldots, n\}$. This is because all such terms are non-negative due to the absolute values surrounding both factors of each term.

6.1.2.1 1-Norm vs. 2-Norm

I prefer the 1-norm over the 2-norm in the context of the general empty vehicle repositioning problem. This is because I did not want to "overpenalize" being far off from the "ideal" distribution compared to "underpenalized" distributions closer to the ideal. I also felt the 1-norm was more appropriate because I have the data already for the realized arrivals of "departure units" and incorporate this.

6.1.3 Simplified EVR Policy: Basic Threshold Penalty

This policy is a myopic policy that attempts to minimize the sum of transporting empty vehicles and a penalty term scaled such that only empty vehicles within 4 minutes and 30 seconds of a newly unsatisfied demand will ship. This policy is very simple, but may leave some individuals stranded.

6.2 Fleet-Sizing (FS) Policies

I present two fleet-sizing policies below. Both fleet-sizing policies seek to optimize a tunable parameter—the fixed fleet size—within a particular policy, possibly subject to certain constraints.

6.2.1 Operational Requirement

This policy aims to minimize the fleet size, F subject to meeting a minimum level of service, s. This level of service refers to the proportion of "departure units" that are serviced, and depart, on time. Such This is the simpler of my two policies; upon extensive testing, it may serve as a straightforward policy for catering to an operational requirement from the state legislature or other legal entities. A rough estimate of a reasonable value of s may be $s \approx 0.98 = 98\%$. However, this parameter itself is tunable.

6.2.2 Expected Annual Cost

This policy aims to minimize an overall expected annual cost for the manager of an autonomous taxi system. Following assumptions about demand requests and possible revenue that I discuss below, this optimization problem becomes equivalent to maximizing expected annual profit for the manager, given that the manager collects revenue strictly from providing its service.

As in the operational requirement policy discussed above, I use the fleet sizes suggested by the naï ve fleet-sizing policies discussed in Section 4.5 as a starting point for an optimization problem. Next, I obtain a minimum fleet size that is determined by serving 98% of all trips on time. Next, I try simulations using the same data set using 100%, 110%, and 125% of this minimum fleet size with the Model-Based High-Complexity EVR Policy, collecting the total realized cost for each simulation along the way. This total realized cost is exactly the realized version of the objective function discussed in Section 5.2.5 above. That is, this realized cost R_{π_F} for a given policy π_F for a particular simulation is:

$$\begin{aligned}R_{\pi_F} &= \sum_{t \in \mathcal{T}} C_t\left(\mathbf{S}_t,\, X^{\pi}\left(\mathbf{S}_t\right)\right)\\ &= \sum_{t \in \mathcal{T}}\left(\sum_{i \in \mathcal{P}}\sum_{j \in \mathcal{P}}\left(c\,\delta_{ij}\,x^{e}_{ttij} + z \sum_{t' \leq t-\tau}\sum_{n \in \mathbb{N}^+}\sum_{k \in \mathbb{N}^+} d_{t'tinjk}\right)\right)\end{aligned}$$

Here, the policy specified by π_F refers to the Model-Based High-Complexity EVR Policy given one of the specific fleet sizes, i.e., 100%, 110%, or 125% of the minimum fleet size described above.

Before I proceed, I present two critical assumptions that make this policy possible:

1. Loaded trips taken by individuals are not affected by the fleet size. Thus, variable costs from non-empty trips remain the same across variable fleet sizes.

2. Any revenue generated from these non-empty trips, if any such revenue is sought, is not affected by the fleet size. Thus, overall revenue—of which possible revenue generated from these non-empty trips is the only source—is not affected by the fleet size.

Indeed, different fleet sizes will cause different degrees of lateness of service of "departure units" at different Pixels throughout the day. In general, smaller fleet sizes may induce more lateness than larger fleet sizes. This may inevitably allow marginally more ride sharing, which may actually decrease the total VehicleMiles traveled in a given day. However, in order to simplify this policy from a maximization of profit, if any revenue is generated, to a minimization of cost, I assume that any such difference in the total VehicleMiles traveled in a given day is negligible. This is relatively reasonable, especially given the inevitably relatively low degree of ride sharing overall as described in Chapter 3 due to the sparseness of trip demand, and given the relatively high service levels we hope to provide.

Thus, the only two components in a simple representation of profit from managing this system of aTaxis that vary with fleet size belong to cost and are:

1. realized costs from empty vehicle repositioning, represented for a given day by R_π as above, and

2. fixed costs per vehicle, including:

- expected purchase cost per vehicle, or expected manufacturing cost per vehicle if vehicles are not bought from a third party, and
- expected cost of operations above gas or charge per vehicle over its lifespan, including maintenance, garaging, plus the cost of money due to interest rates.

This is because revenue and variable costs do not vary with fleet size. I assume that the expected annual cost of operations above gas or charge is a proportion $a > 0$ of the expected purchase cost or expected manufacturing cost per year. A rough estimate of a in April 2015 may be $a \approx 0.05 = 5\%$; this parameter itself is tunable. I further assume that the lifespan of each vehicle, L, is a function of the fleet size, F, because a larger fleet size translates to fewer overall TripMiles traveled by each aTaxi as a result of my first critical assumption presented above. I assume that the lifespan of each vehicle translates to a fixed number of total miles traveled, m. A rough estimate of m in April 2015 may be $m \approx 250{,}000$; this parameter is also tunable. If there were no empty vehicle trips, then the lifespan of each vehicle would be strictly a function of the fleet size, F, as another result of my first critical assumption presented above. Furthermore, L would be simply be linear in and directly proportional to F over a feasible region, since a fixed number of (loaded) VehicleMiles would be split roughly evenly among all aTaxis. However, there are empty vehicle trips—and the total VehicleMiles traveled by empty vehicles depends on the policy π_F. Let $V^e(\pi_F)$ be the number of VehicleMiles traveled by empty vehicles in one day under the policy π_F. Thus, L, in years, is a function of the policy π_F with fleet size F over a feasible region:

$$L(\pi_F) = \begin{cases} \frac{mF}{365.25\left(V^e(\pi_F) + V^l\right)}, & \text{if } F \geq F^{min} \\ 0, & \text{otherwise} \end{cases}, \ \forall\, \pi_F \in \Pi$$

where:

- Π is the set of feasible policies described above, i.e., the Model-Based High-Complexity EVR Policy with different fixed fleet sizes, each of which assumes a fleet size F,
- F^{min} is some absolute minimum number of aTaxis for the state to justify even thinking about an autonomous taxi system for New Jersey, and
- V^l is the number of VehicleMiles traveled by loaded vehicles, which is independent of the fleet

size corresponding to $\pi_F \in \Pi$ as a result of my first critical assumption presented above.

In order to obtain the lifespan, we simply divide the fixed number of miles each vehicle can travel in its lifetime by the the number of VehicleMiles traveled by an average vehicle in a day under the policy π_F and adjust units accordingly, assuming that there are—on average—365.25 days in one year.

Suppose that for a given value of $F \geq F^{min}$, the total number of empty VehicleMiles is equal to 20% of the total number of loaded VehicleMiles. Then, $V^e(\pi_F) = 0.20V^l$. If we assume that V^l = 272,865,701, which is the number of VehicleMiles traveled by loaded vehicles in one day according to the parameter set $\{CD = 3, DD = 300\text{ seconds}, \text{maxCircuity} = 20\%\}$, assuming an infinite fleet size on the ORF 467 data set, then:

$$V^e(\pi_F) = 0.20\, V^l = 0.20\,(272{,}865{,}701) \approx 54{,}573{,}140$$

We do not yet know exactly the fleet sizes $F \in \left\{F : V^e(\pi_F) = 0.20V^l\right\}$ without computation via an implementation using convex optimization software and varying the given fleet size V. Instead, I attempt to visualize possible values for the lifespan, L in Figure 6.1 below. I assume that $F \in \left\{F : V^e(\pi_F) = 0.20V^l\right\}$ and V^l = 272,865,701, which is the number of VehicleMiles traveled by loaded vehicles in one day according to the parameter set $\{CD = 3, DD = 300\text{ seconds}, \text{maxCircuity} = 20\%\}$, on the ORF 467 data set in the infinite fleet case.

Finally, I choose the optimal policy π_F^*, which assumes fleet size F, that minimizes an expression for a rough expected annual cost. The components of this expected annual cost are exactly the components that I list above: realized costs from empty vehicle repositioning and fixed cost per vehicle, which breaks down to expected purchase cost and expected maintenance cost. Thus, I choose:

$$\begin{aligned}
\pi_F^* &= \operatorname*{arg\,min}_{\pi_F \in \Pi} \left(365.25\, R_\pi + \frac{p\,F}{L(\pi_F)} + a\frac{p\,F}{L(\pi_F)}\right) \\
&= \operatorname*{arg\,min}_{\pi_F \in \Pi} \left(365.25\, R_\pi + (1+a)\,\frac{p\,F}{L(\pi_F)}\right) \\
&= \operatorname*{arg\,min}_{\pi_F \in \Pi} \left(R_\pi + \frac{p}{m}\,(1+a)\left(V^e(\pi_F) + V^l\right)\right)
\end{aligned}$$

where p is the expected purchase cost per vehicle, or expected manufacturing cost per vehicle if

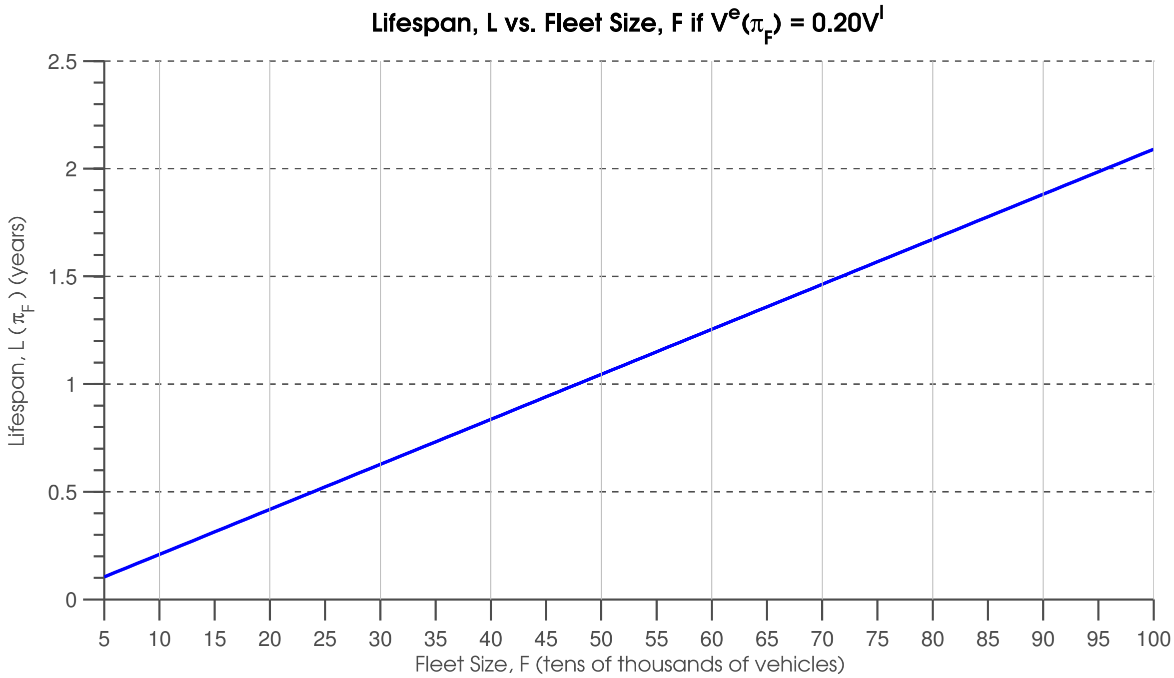


Figure 6.1: Lifespan, $L(\pi_F)$ vs. Fleet Size, F, assuming $F \in \left\{F : V^e(\pi_F) = 0.20V^l\right\}$ and $V^l = 272{,}865{,}701$.

vehicles are not bought from a third party. Sensitivity analysis on the parameter a implies that tuning a in a reasonable region does not have a drastic effect on the quantity to be minimized above. This is easy to see simply because reasonable values of $a << 1$.

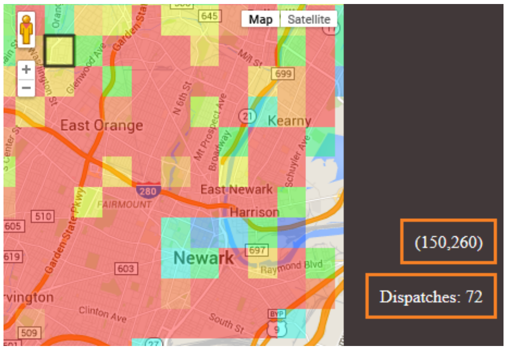
Chapter 7

Empty Vehicle Repositioning and Fleet-Sizing: Results and Analysis

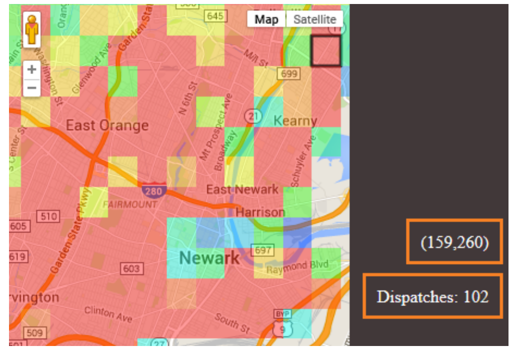
However beautiful the strategy, you should occasionally look at the results.

WINSTON CHURCHILL

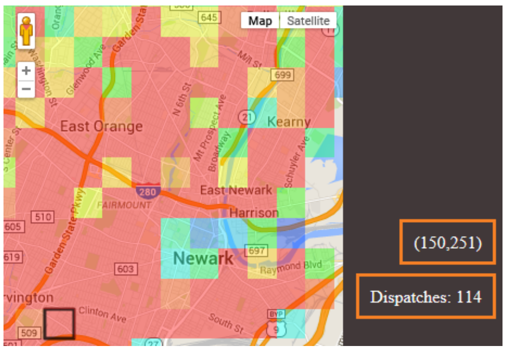
I implemented the Simplified EVR Policy with the Norm-Based Penalty that was discussed in Section 6.1.2 for empty vehicle repositioning. The results were far from expected; I suspect that I implemented the policy incorrectly. Below are some plots that describe the details of the 10 x 10 subset Pixel grid that I prepared for further analysis in the weeks to come. I use the area outline by the Pixels (150, 251), (159, 251), (159, 260), and (150, 260). I would have liked to include the Secaucus Junction Pixel, but it is in an awkward peninsula without a 10 x 10 grid that can surround it that wouldn't include a significant number of Pixels with no trips (due to water). I also think this region has more action than the area surrounding the Princeton Pixel, though I can use that if you'd like as a case study instead. Figure 7.1 below visualizes this 10 x 10 subset Pixel grid on top of a map largely over parts of Newark, Kearny, and East Orange. These corner Pixels are highlighted in black in the Figure 7.1 below. Overlaid beneath each of the highlighted corner Pixels is a color map with Pixels plotted with with various colors corresponding to their respective "departure unit" demand request rates from "yesterday's" ORF 467 data set between 9:00 AM and 9:09 AM, inclusive. These maps are taken from the Color Map page of my Interactive "Demand Request" Rates Visualization web application for 9:0x AM; I explain this web application further in Chapter 8.



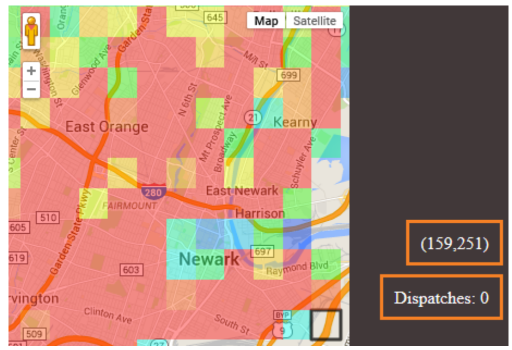
(a) $(150, 260)$



(b) $(159, 260)$



(c) $(150, 251)$



(d) $(159, 251)$

Figure 7.1: Corner Pixels, highlighted in black, overlaid on my Color Map for 9:0x AM.

Compare this to Figure 7.3 below, which contains a plot of the number of aTaxis in transit for "today's" data set constructed from Hill's Module 6 NN files within this 10 x 10 Pixel grid in the case of an infinite fleet. Plotted in red is the total number of aTaxis in transit versus time in the infinite fleet case *for "today"* within the same 10 x 10 subset Pixel grid. Like before, both the origin Pixel and destination Pixel of each "departure unit" accounted for in this plot are within the set of 100 Pixels in this grid. The blue dashed line marks the number of aTaxis in transit at midnight, 71; the green dashed line marks the maximum number of aTaxis in transit during the day, $197 + 71 = 268$, which occurs from 12:01:15 PM to 12:01:18 PM. All "departure units" in this plot, 'today's" data set constructed from Hill's Module 6 NN files, are also generated from ride share analysis on the parameter set $\{CD = 3, DD = 300 \text{ seconds}, \text{maxCircuity} = 20\%\}$.

Finally, Figure 7.4 below overlays the two plots.

As for the individual plots, both the origin Pixel and destination Pixel of each "departure unit" accounted for in this overlay are within the set of 100 Pixels in this 10x10 subset Pixel grid. Furthermore, all "departure units" in the overlay were generated by ride share analysis on the same

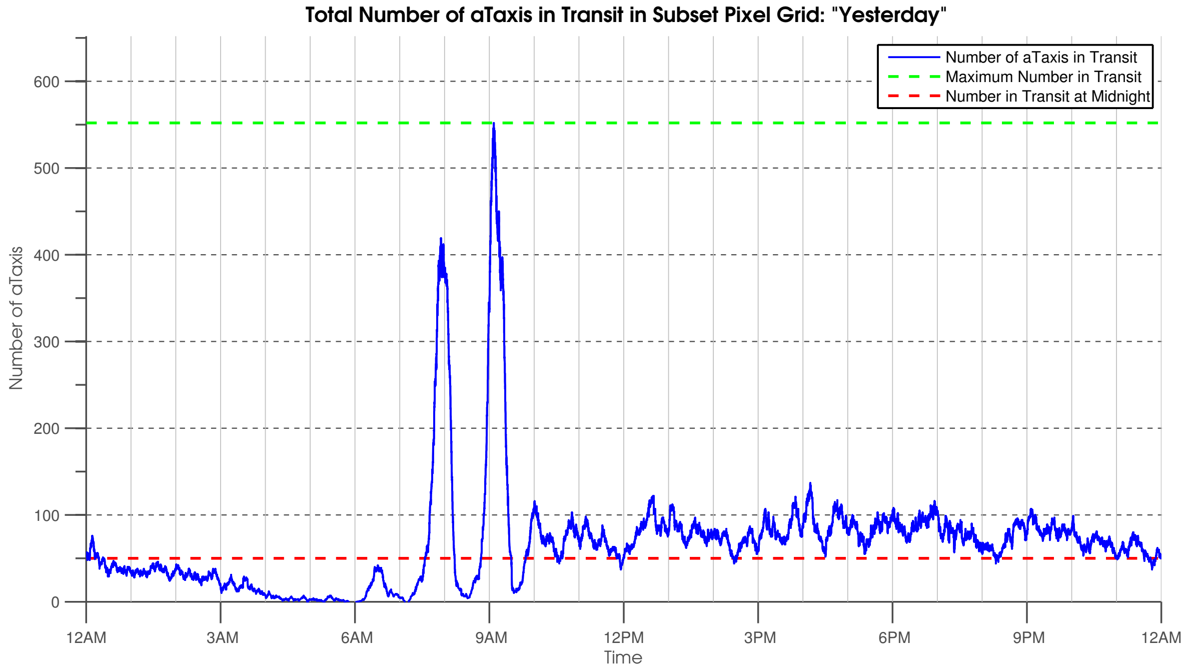


Figure 7.2: *"Yesterday's"* aTaxis in transit in the 10 x 10 subset Pixel grid in the infinite fleet case.

parameter set $\{CD = 3, DD = 300 \text{ seconds}, \text{maxCircuity} = 20\%\}$.

It is not difficult to notice the relatively extreme difference between the two plots. "Today" will require roughly 15% more total aTaxi dispatches in this subset grid than "yesterday," but the dispatches "today" are significantly more spread out, especially after noon. We see that the number of aTaxis in transit "today" in the infinite fleet case would gradually ease down from a peak of 268 at about noon, roughly linearly over longer periods of time. While the spike in traffic from approximately 10 AM until noon "today" is relatively drastic, it is not nearly as steep and magnified as the two spikes in traffic that we saw "yesterday," climbing first by about 400 aTaxis in one hour from 7 AM until 8 AM, and then climbing by approximately 550 vehicles in roughly 30 minutes from 8:30 AM until 9 AM. Though likely magnified by increased variance due to the very small subset of 100 Pixels with which we are working, these stark differences highlight the relative importance of reacting to realized "demand requests" over forecasting future "demand requests" according to "yesterday's" data or average data in general. Even if "demand requests" made by individuals were relatively similar, the effective "demand requests" for aTaxis may be less similar due to ride sharing constraints. Indeed, a parameter that dramatically scales down the impact on

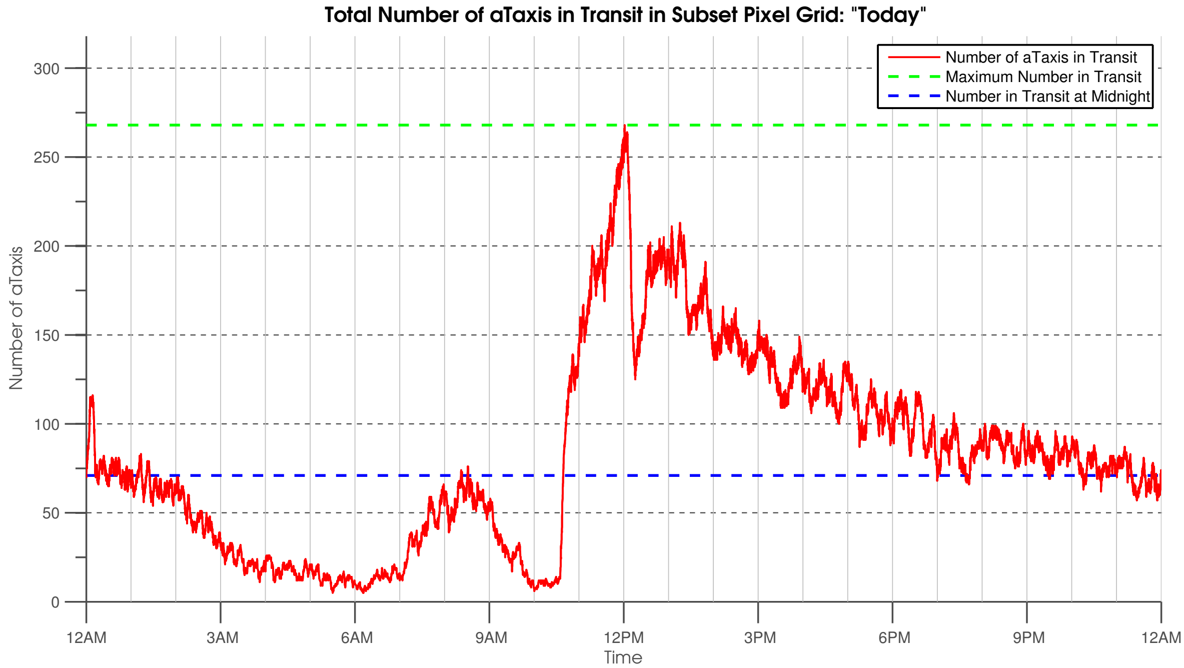


Figure 7.3: *"Today's"* aTaxis in transit in the 10 x 10 subset Pixel grid in the infinite fleet case.

the ideal vector of forecasted "demand requests" may dramatically improve results.

I used 100%, 150%, 250%, 500%, and 1000% of the absolute minimum fleet size for the 10 x 10 Pixel grid, 552, and tested each of these fleet sizes using both the 1-norm and the 2-norm. Visualizations of my preliminary results are located at `www.princeton.edu/~kwdougla`, as I describe further in Chapter 8.

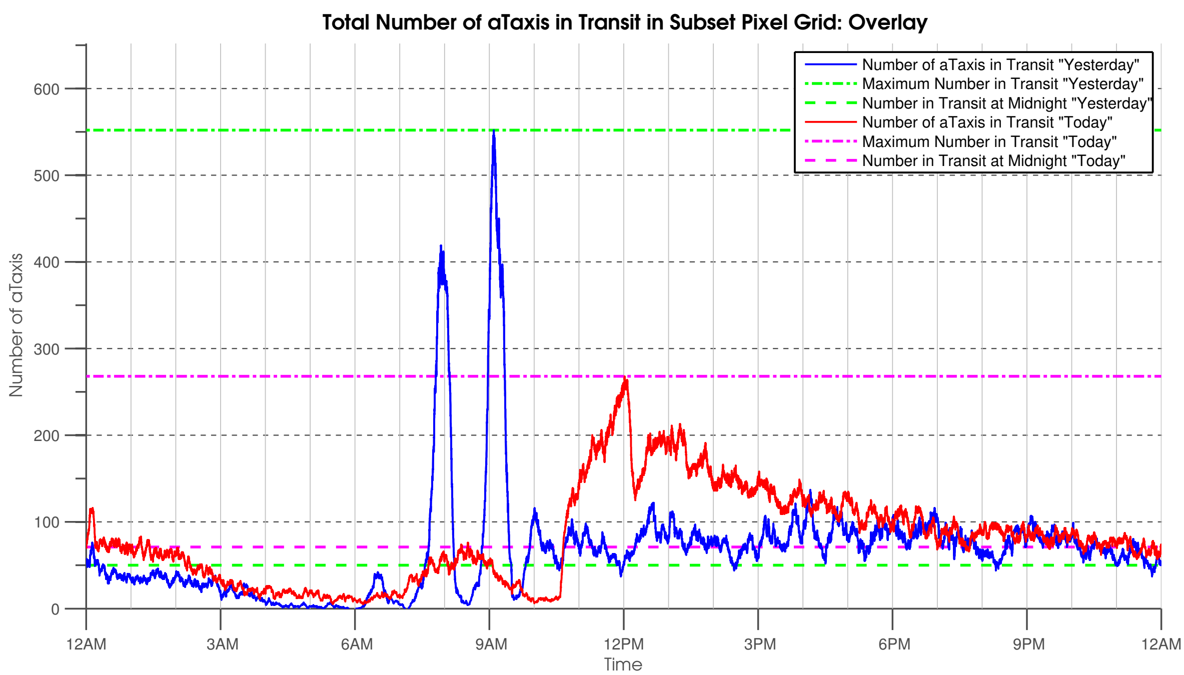


Figure 7.4: Overlay of the two plots in FIgures 7.2 and 7.3, giving the total number of aTaxis in transit versus time *for both "yesterday" and for "today"* within the 10 x 10 subset Pixel grid in the infinite fleet case.

Chapter 8

"Demand Request" and Trip Visualization

Believe only half of what you see and nothing that you hear.

Edgar Allan Poe

This chapter summarizes the functionality of two web applications that I developed for important data visualization: one for the visualization of "demand requests" over time on "yesterday's" ORF 467 data set, and one for the visualization of empty and loaded trips on a 10 x 10 subset Pixel grid in New Jersey.

Both web apps are located at `www.princeton.edu/~kwdougla`. At this URL, there are also two video-based tutorials on how to operate both web applications. I provide basic overviews of each page of both web applications below. Each web app is subject to change and continual improvement; all updates and relevant information will be reflected at this URL.

Together, the web applications serve to give the user a better understanding of the demands during a "typical day" in New Jersey that are relevant for the empty vehicle repositioning and fleet-sizing problems.

8.1 Interactive Visualization of "Demand Requests"

My first web application serves to visualize incoming realized "demand requests" made by individuals and "departure units" over time across the entire state. All "demand request" data is generated from "yesterday's" ORF 467 data set in the infinite fleet case; "demand requests" from "departure units" were created by performing ride share analysis on the parameter set $\{CD = 3, DD = 300\,\text{seconds}, \text{maxCircuity} = 20\%\}$, and accumulating "demand requests" within ten minute intervals. In the infinite fleet case, "demand requests" occur *exactly* DD before the departure times of all "departure units," and "demand requests" occur *within* DD before the departure times of all individuals. The differences in time in the infinite fleet case are negligible. Yet there are two reasons that I chose to visualize "demand requests" instead of dispatches. First, "demand requests," not dispatches, of "departure units" are relevant in forecasting demand for any empty vehicle repositioning policy in the finite fleet case. Second, the primary goal of this thesis is to make steps towards optimal empty vehicle repositioning and fleet-sizing policies. The differences in loaded vehicle dispatch times on the same sample path between the infinite fleet case and a reasonably small finite fleet case may be significant, while the "demand requests" made by individuals are identical and the "demand requests" made by "departure units" are extremely close to identical.

This web application contains four pages; a link to each page is located at the top of the web application. As of the publication of this thesis, each of these pages loads and parses a large JSON file and large CSV files that allow the user interface to be quick and agile at the expense of a longer page loading time.

8.1.1 1 Interactive Heat Map

The `1 Interactive Heat Map` page contains a slider bar with two end points near the top of the page. Beneath the slider bar is a zoomable map of New Jersey beneath a Google Maps API heat map layer in which different "heat levels" at each Pixel correspond to different numbers of "demand requests." By sliding the left end point corresponding to different ten minute intervals, the heat map of "demand requests" changes accordingly. Furthermore, there is an "Animate" button that automatically advances the heat map over ten minute intervals from the left end point to the right end point on the slider bar, subject to a user-inputted delay. There are options for toggling between individual and "departure unit" demand requests, as well as for setting the minimum and maximum "demand request" levels to display in any given ten minute interval. Mousing over each Pixel displays its coordinates and number of "demand requests" during the currently plotted ten minute interval in the bottom right of the screen. Figure 8.1 displays a screenshot of the `1 Interactive Heat Map` page.

8.1.2 2 Interactive Heat Maps

The `2 Interactive Heat Maps` page has the same functionality as the `1 Interactive Heat Map` page, except that the web browser is split into two halves so the user can view two maps, at different intervals or with different options, at the same time.

8.1.3 Interactive Plots

The `Interactive Plots` page displays a plot of "demand requests" versus time for any of the 21,641 Pixels in NJ from which at least one trip originated in "yesterday's" ORF 467 data. The user can zoom in on the map placed next to the plot and mouse over and click on any Pixel to view the plot of "demand requests" versus time for that Pixel. The user can choose to display individual "demand requests," "departure unit" "demand requests," or overlaid "demand requests." Each click on the map produces a bright green marker on the plotted Pixel; all Pixels selected in the past are identified by red markers. Figure 8.2 displays three screenshots of the `Interactive Plots` page.

8.1.4 Interactive Pixel Color Map

The `Interactive Pixel Color Map` page functions similarly to the `1 Interactive Heat Map` page. However, each plot maps each Pixel to an overlay color corresponding to the number of "demand requests" in the selected ten minute interval. The color mapping uses my own array of color code Strings, and a legend next to the map gives sample "demand request" values and their corresponding color on a band of colors. Figure 8.3 displays a screenshot of the `Interactive Pixel Color Map` page, while Figure 8.4 gives a sample band of colors from my array of color code Strings.

8.2 Trips Visualization by Policy

My second web application serves to visualize both loaded and empty vehicle trips taken "today" on the same 10 x 10 subset Pixel grid described and displayed in Chapter 7, as a result of different simplified empty vehicle repositioning policies given different fixed fleet sizes. Options for selecting the class of empty vehicle repositioning polices and fixed fleet sizes are located at the top of the page. I run Dijkstra's shortest path algorithm on an edge-weighted digraph of nodes in order to display roughly "actual" trips taken by loaded and empty aTaxis; as of now, the costs of repositioning empty vehicles in the underlying polices are still given by straight-line distance and an average speed limit. The edge-weighted digraph corresponds to major intersections in New Jersey with links containing weights corresponding to the distance between two connected nodes divided by the speed limit on the corresponding stretch of road, representing an estimation of the required travel time between pairs of connected intersections.

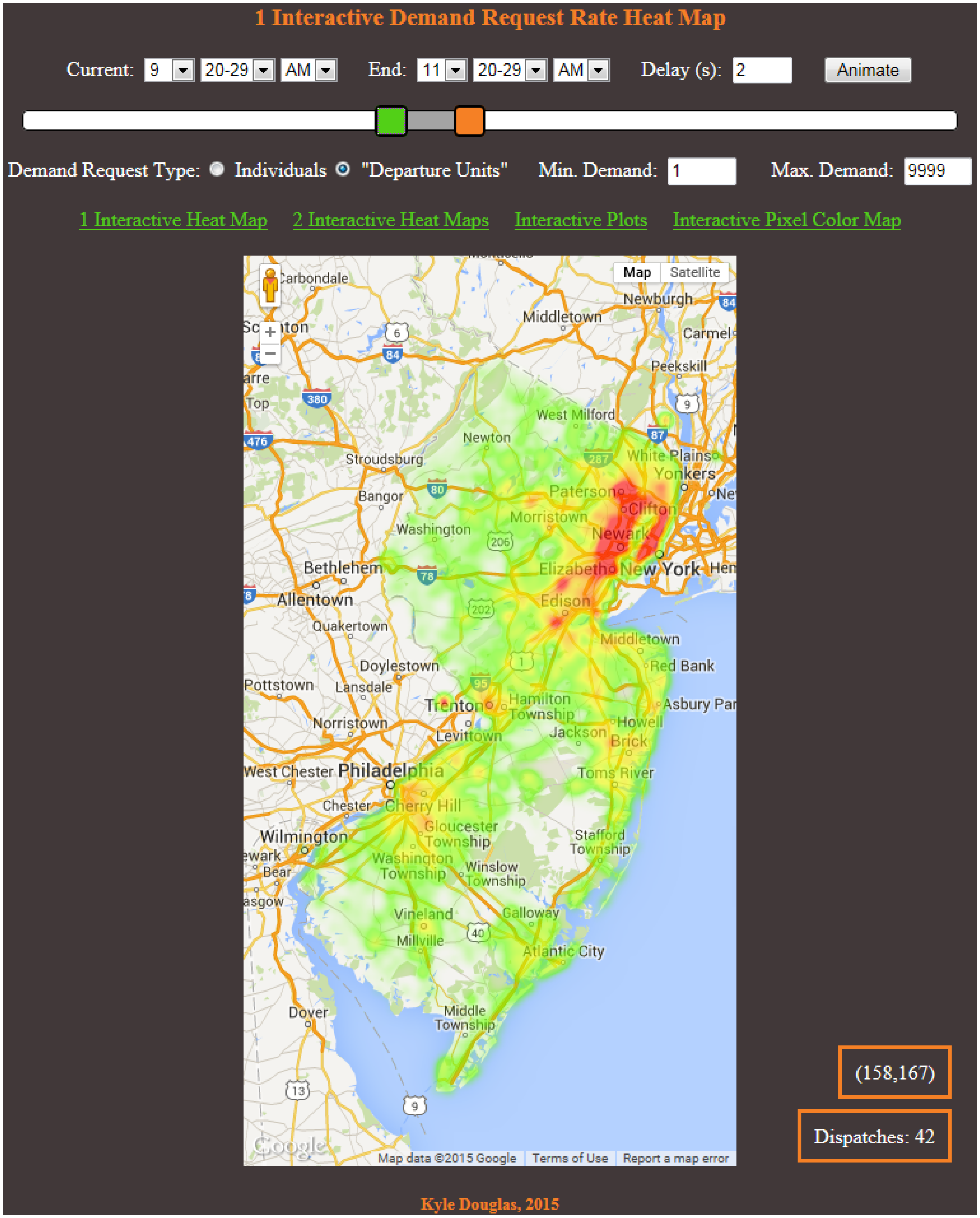
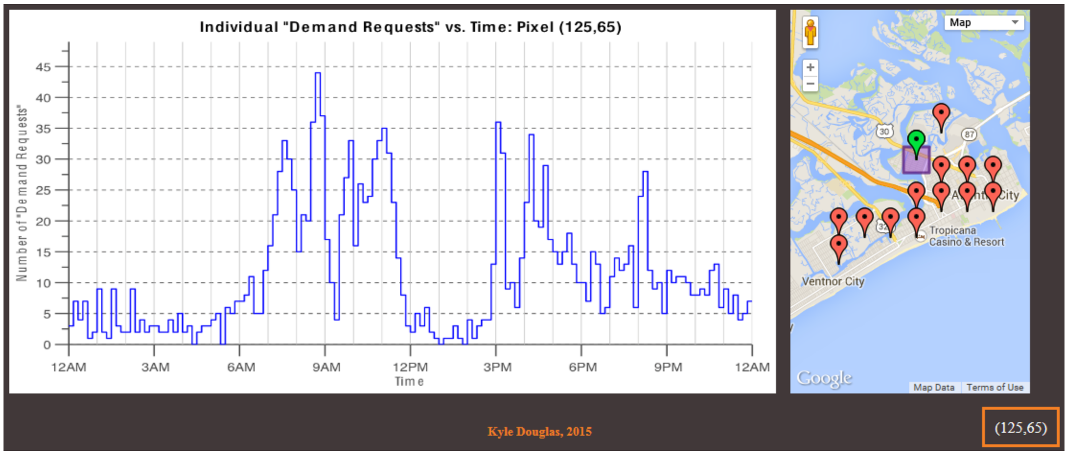
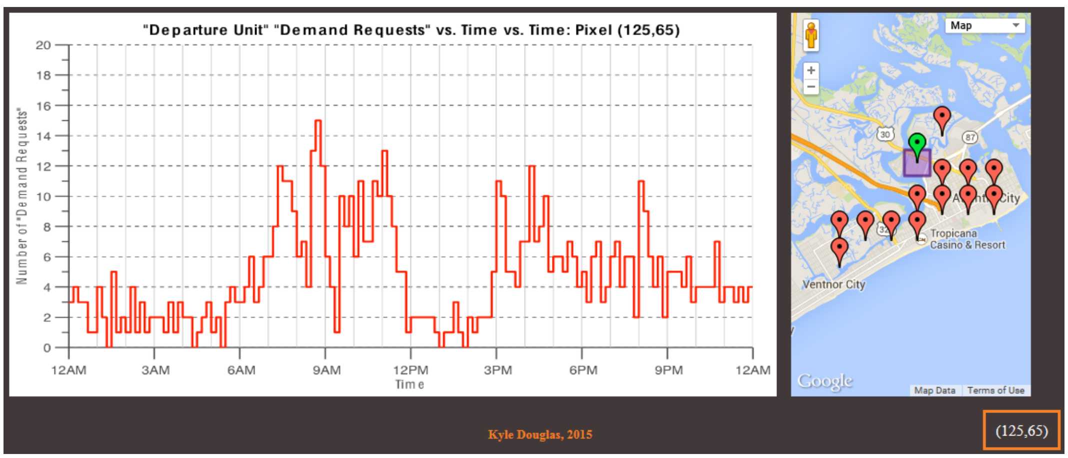


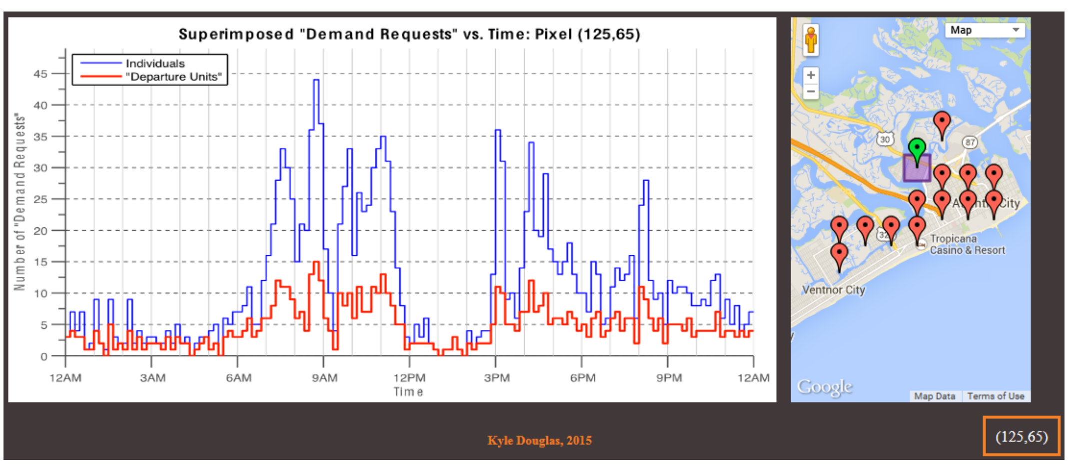
Figure 8.1: Interactive Visualization of "Demand Requests": 1 Interactive Heat Map - Screenshot.



(a) Individual "Demand Requests" vs. Time for Pixel $(125, 65)$ in Atlantic City.



(b) "Departure Unit" "Demand Requests" vs. Time for Pixel $(125, 65)$ in Atlantic City.



(c) Superimposed "Demand Requests" vs. Time for Pixel $(125, 65)$ in Atlantic City.

Figure 8.2: Interactive Visualization of "Demand Requests": Interactive Plots - Screenshots.

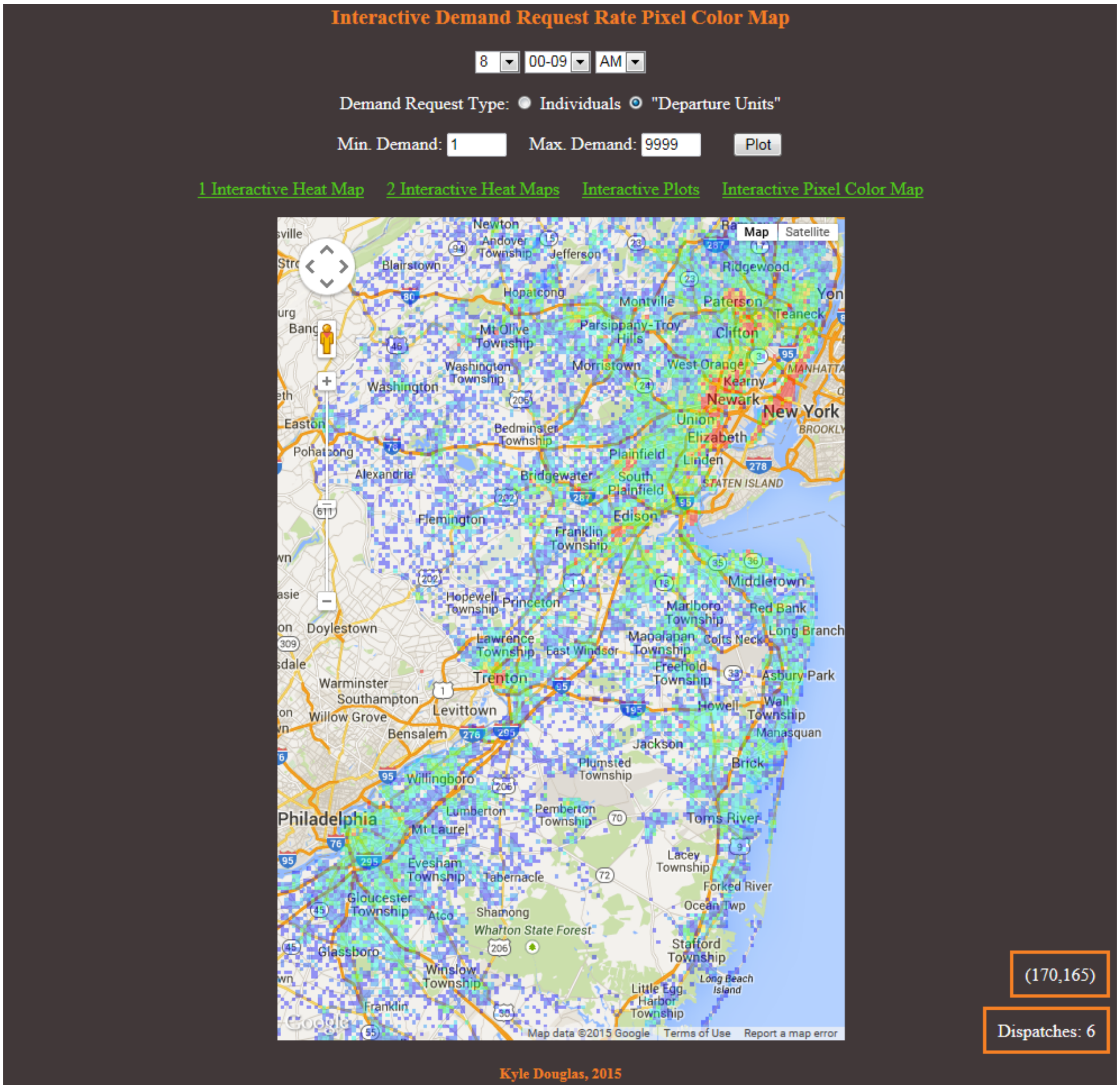


Figure 8.3: Interactive Visualization of "Demand Requests": Interactive Pixel Color Map - Screenshot.

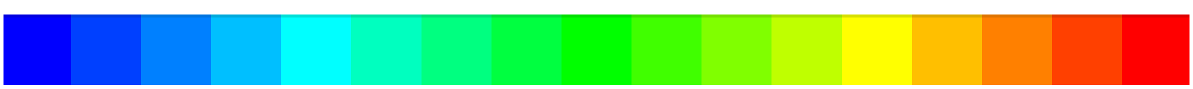


Figure 8.4: Sample band of colors from the `Interactive Pixel Color Map` page..

Chapter 9

Conclusion

If all the economists were laid end to end, they'd never reach a conclusion.

George Bernard Shaw

My conclusion is limited because I will continue to work on implementing my empty vehicle repositioning and fleet-sizing policies in the coming weeks. Below, I expand on some next steps that can be taken in further research beyond the next few weeks.

9.1 Next Steps

- *Intermediate Pickups*: When $CD > 1$, aTaxis that visit more than one TripNode may be able to pick up new passengersup to full capacity of the vehicle following drop-offat TripNodes prior to their final destination. My current ride share analysis does not account for this possibility; as of now, I finalize an entire taxi trip reaching all TripNodes at the time of departure. It is reasonable to claim that being able to perform intermediate pickups could result in closer-to-optimal handling of demand. However, the impact may perhaps be relatively small or negligible due to (1) the sparseness of the demand in many areas of the state, and (2) the fact that an intermediate pickup might eliminate an otherwise shared ride initially departing at the intermediate TripNode. When is it worth picking up one or more passengers at an intermediate stop? Is there a provably optimal condition?

- *Different Vehicle Capacities*: Currently, my initial investigation assumes vehicle capacities of two, six, fifteen, and fifty passengers. These choices were made based on an initial investigation of ride-sharing for the P2P parameter set of CD = 3, DD = 300, MaxCircuity = 20% using the updated ride-share constraints. However, these choices were not made based on any objective function or any significant analysis of different practical capacity combinations. It is worth exploring different sets of vehicle capacities to investigate whether a different practical combination is more optimal than the current set taken as given in the preliminary analysis.

- *Refueling or Recharging*: What energy propulsion system is the best for the autonomous taxi system in question? This thesis focuses on the analysis and operations research of a network of stochastic travel demands applied to an autonomous taxi system for New Jersey, and does not focus on the technology of the autonomous vehicles themselves. However, the technology has an inevitable impact on the realistic empty vehicle repositioning policy and more general control policy for any autonomous taxi system because vehicles must refuel or recharge. Do current oil prices and futures and their projections play a role in determining vehicle types, or a distribution of vehicle types? If some or all of the vehicles require gasoline, where and when is it optimal to refuel, and how is it best to accommodate the existing infrastructure and gas stations? Will there be efficient charging stations at each aTaxi station? How would recharging affect vehicle availability and repositioning?

Bibliography

Ahmadi, A. A. (2014). Notes on inner products and norms, eigenvalues and eigenvectors, elements of differential calculus, and taylor expansion. Lecture notes for ORF 363-COS 323: Computing and Optimzation. Princeton University.

Beaujon, G. J. and Turnquist, M. A. (1991). A model for fleet sizing and vehicle allocation. *Transportation Science*, 25:19–45.

Billington, D. P. (1996). *The Innovators: The Engineering Pioneers Who Made America Modern.* John Wiley & Sons, Inc.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50.

Çinlar, E. (2012). Notes on poisson processes. Lecture notes for ORF 309/MAT 309. Princeton University.

Hart, M., Roberts, M. J., and Stevens, J. D. (2003). Zipcar: Refining the business model. *Harvard Business School Press*, 9(803-096).

Hendricks, D. L., Fell, J., and Freedman, M. (2001). The relative frequency of unsafe driving acts in serious traffic crashes. *U.S. Department of Transportation.*

Howard, B. (2015). Delphi self-driving car goes coast-to-coast, autonomously.

Huet, E. (2014). Uber says it's doing 1 million rides per day, 140 million in last year.

Kornhauser, A. L. and in ORF 467, S. (2011). Personal rapid transit: A transportation plan for new jersey. ORF 467 Fall 2010 Final Project.

Mandelbaum, A. and Zeltyn, S. (2004). The impact of customers patience on delay and abandonment: Some empirically-driven experiments with the m/m/n+ g queue. *OR Spectrum*, 26(3):377–411.

Massey, W. A. (2014). Notes on traffic models and poisson processes. Lecture notes for ORF 407. Princeton University.

Mufti, T. (2013). Synthesis of spatially & temporally disaggregate person trip demand: Application for a typical new jersey weekday. Master's thesis, Princeton University, Princeton, NJ.

Powell, W. B. (1986). An operational planning model for the dynamic vehicle allocation problem with uncertain demands. *Transportation Research*, 21B(3):217–232.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, 2nd edition.

Powell, W. B. and Cheung, R. K.-M. (1994). A network recourse decomposition method for dynamic networks with random arc capacities. *Networks*, 24(7):369–384.

Powell, W. B. and Simão, H. P. (2014). Modeling. Lecture notes for ORF 411: Operations and Information Engineering. Princeton University.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Song, D.-P. and Earl, C. F. (2004). Optimal empty vehicle repositioning and fleet-sizing for two-depot service systems. *European Journal of Operational Research*, 185:760–777.

U.S. Department of Transportation (2009). Summary of Travel Trends: 2009 National Household Travel Survey. Technical report, Federal Highway Administration, Washington, D.C.

White, W. W. (1972). Dynamic transshipment networks: An algorithm and its application to the distribution of empty containers. *Networks*, 2:211–236.

Wyrough, Jr., A. P. H. (2014). A national disaggregate transportation demand model for the analysis of autonomous taxi systems. Bachelor's thesis, Princeton University, Princeton, NJ.