12-2018

Modeling movement decisions in networks: A discrete choice model approach

Larry LIN JUNJIE
Singapore Management University, larry.lin.2013@phdis.smu.edu.sg

MODELING MOVEMENT DECISIONS IN NETWORKS: A DISCRETE CHOICE MODEL APPROACH

LARRY LIN JUN JIE

SINGAPORE MANAGEMENT UNIVERSITY

2018

Modeling Movement Decisions in Networks:

A Discrete Choice Model Approach

Larry Lin Jun Jie

Submitted to School of Information Systems
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy in Information Systems

Dissertation Committee:

Shih-Fen Cheng (Supervisor/Chair)
Associate Professor of Information Systems
Singapore Management University

Hoong Chuin Lau (Co-Supervisor)
Professor of Information Systems
Singapore Management University

Pradeep Varakantham
Associate Professor of Information Systems
Singapore Management University

Kathleen M. Carley
Professor
Carnegie Mellon University

Singapore Management University
2018

I hereby declare that this PhD dissertation is my original work and it has been written by me in its entirety.
I have duly acknowledged all the sources of information which have been used in this dissertation.

This PhD dissertation has also not been submitted for any degree in any university previously.

Larry

Larry Lin Jun Jie
21 December 2018

Modeling Movement Decisions in Networks:
A Discrete Choice Model Approach

Larry Lin Jun Jie

Abstract

In this dissertation, we address the subject of modeling and simulation of agents and their movement decision in a network environment. We emphasize the development of high quality agent-based simulation models as a *prerequisite* before utilization of the model as an evaluation tool for various recommender systems and policies. To achieve this, we propose a methodological framework for development of agent-based models, combining approaches such as discrete choice models and data-driven modeling.

The discrete choice model is widely used in the field of transportation, with a distinct utility function (e.g., demand or revenue-driven). Through discrete choice models, the movement decision of agents are dependent on a utility function, where every agent chooses a travel option (e.g., travel to a link) out of a *finite* set. In our work, not only do we demonstrate the effectiveness of this model in the field of transportation with a multi-agent simulation model and a tiered decision model, we demonstrate our approach in other domains (i.e., leisure and migration). where the utility function might not be as clear, or involve various qualitative variables.

The contribution of this dissertation is therefore two-fold. We first propose a methodological framework for development of agent-based models under the conditions of varying data observability and network model scale. Thereafter, we demonstrate the applicability of the proposed framework through the use of three case studies, each representing a different problem domain.

Contents

List of Figures

List of Tables

Acknowledgement

First and foremost, I would like to thank my supervisor Associate Professor Cheng Shih-Fen for his guidance and advice from my undergraduate days, to my stint as a research engineer, to my PhD candidature days. Little had I known that what I worked on as an undergraduate project would eventually become my bread and butter. Additionally, I would like to thank my co-supervisor, Professor Lau Hoong Chuin, as well as Associate Professor Pradeep Varakantham for their guidance and advice, which helped to shape this dissertation, as well as chapter 4.

I am thankful to Professor Lim Ee-Peng for the scholarship from Living Analytics Research Centre, as well as the opportunity to participate in the overseas training residency at Carnegie Mellon University. I would also like to express my gratitude to Professor Kathleen M. Carley (Carnegie Mellon University) for her guidance and kind encouragement, as well as for chapter 6 of the dissertation.

Last but not least, thank you to my wife (Liping) and daughter (Emma) for their unconditional support throughout this journey.

Chapter 1

Introduction

1.1 Motivation

Movement is an essential part of everyday life. People move for multiple reasons, such as getting from an origin to destination (e.g., home to workplace), or for livelihood purposes (e.g., taxi drivers in search of passengers). Additionally, movement can occur at different frequencies, whether its continuous movement in matter of seconds (when driving), to ad-hoc movement which requires a contemplation period of years (e.g., migration).

Agent-based models have been developed to model the movement of agents in a network environment. There could be various reasons to model the movement of agents in networks, such as utilizing agent-based models as an economical approach to evaluate various recommender systems and policies. Therefore, it allows policy makers to evaluate the effectiveness of such systems before an actual system-wide implementation.

The development of such agent-based models involves a bottom up approach, which traditionally suffers from issues such as data observability. This is mitigated with the proliferation of mobile devices and applications, which increased the ease of obtaining spatio-temporal data for analysis and modeling. Such spatio-temporal data generally suffer from another

problem: it could be hard to figure out users intentions and decisions from the raw data which usually involves a $< time, location >$ tuple. Other than the challenge of inferring intent, most of such agent-based models often involve rule-based movement decisions, or representation of movement through transitional *origin-destination* matrices, which suffers from a lack of robustness in dealing with changes in the network environment.

Discrete choice models have been proposed to model the movement of agents in the transportation domain, where the utility form is distinct (e.g., driven by revenue or demand). The discrete choice model is a good candidate for modeling of movement decisions as it represents the movement decision of an agent well (i.e., agent selects from a set of alternative links, roads, or zones to travel to). Additionally, the movement decision satisfies the requirements of the discrete choice model: (1) alternatives are *collectively exhaustive* (e.g., links or nodes), (2) alternatives are *mutually exclusive* (choosing to travel to link B means that the agent does not travel along any other links), and (3) set contains a *finite* number of alternatives (set of links or nodes to travel).

On the other hand, the discrete choice model is less utilized in other domains where the utility form is less distinct (e.g., consisting of qualitative factors). In this dissertation, we make the contributions as follows:

- **Representation of movement decision via discrete choice model.** Previous works have adopted the discrete choice model for domains (e.g., transportation) where the utility takes on a distinct form (e.g., revenue or demand driven). We show that the discrete choice model can also be used in domains (leisure, migration) where the utility form might not be as distinct (i.e., qualitative).

- **Propose methodological framework for development of agent-based models.** We propose a framework on how a user can build *high quality* agent-based models involving movement in a network

environment. We first define problem domains under different conditions of network model scale and data observability. Thereafter, we provide guidelines on dealing with potential domain-specific issues and steps for developing agent-based models in these problem domains.

- **Illustration of applicability.** We illustrate how the proposed framework could be utilized by applying it in three concrete case studies, each representing a problem domain.

- **Adoption of data-driven approach** for development of agent-based simulation models, even in the scenario of partial data observability (e.g., leisure). In the event of partial data observability, instead of creating a small scale model of a sampled sub-population (previous approaches), we make use of (1) data collected from a sample population to train the decision model, and (2) use of simulation model as a ground truth generator for generating the mobility traces of an entire population.

Summary of the use cases introduced in chapter 4, 5 and 6 is as shown in table 1.1. In particular, the table shows the network type, movement type and utility function of the agent-based and network models. This will be further discussed in the individual chapters.

Table 1.1: Problem Domains: networks/movement types and utility model.

Area	**Network Type**	**Movement Type**	**Utility Function**
Leisure	Network of attractions in theme park	Non-recurring, ad hoc movement behavior. Visitors at a theme park are less likely to visit attractions that they have visited before.	Consists of thrill, dark, wet level of attractions.
Transport	Road network: consists of zones, links and junctions	Continuous; drivers seek passengers continuously when in roaming (i.e., free) mode.	Historical demand at zones, along with probability of a passenger pickup at road links.
Migration	Country networks: migration, economic disparity, etc.	Movement from one node (country) to another. Movements here are typically a *one-off* type of movement, which requires more contemplation and consideration as opposed to the previous two forms of movement.	Country indicators such as economic disparity, alliance/hostility, migrants network, just to name a few.

1.2 Organization of Dissertation

The rest of this dissertation is organized as follows. In chapter 2, we describe the methodological framework that is proposed as a guideline for users in development of agent-based models involving movement in a network environment. In chapter 3, we first discuss some research that are relevant to the overarching topics of behavioral and agent-based modeling. In chapters 4, 5, and 6, we present our research works in the problem domains of leisure, transportation, and human migration respectively. Lastly, we summarize the contributions and highlight some future directions in chapter 7.

Chapter 2

Methodological Framework: Foundation to Building Agent-Based Models

In this chapter, we describe our proposed methodological framework that provides a guideline to users on the development of *high quality* agent-based models involving movement in a network environment. We describe how users can develop their models under different data observability and network model scales. To provide illustrations on how this framework could be applied to different application domains, we provide three case studies in chapter 4, 5 and 6, each representing a unique setting.

2.1 Agent-Based Model

2.1.1 Model Characteristics

For clarity, within the scope of this dissertation, the term agent-based models refers to models with the following characteristics. Agents are initialized to be associated with nodes within a given network environment, and move around from node to node at every discrete time interval. The

movement decisions are based on application-specific utility functions, where agents are probabilistically more likely to travel to nodes which gives a higher utility.

This makes the discrete choice model a viable option for modeling the decision model of agents, i.e., selecting a travel alternative (e.g., link) out of a set of possible alternatives based on a utility function. From an agent's perspective, the utility obtained is essentially the reward (quantitative or qualitative), minus the *cost of action* (e.g., distance to travel). Additionally, there is the concept of *congestion-awareness* in the networks. In the case of positive network externalities, an agent's utility would increase with more agents making the same decision. In the case of negative network externalities, it would be a congestion function, where an agent's utility would decrease with increasing number of agents making the same movement decision. What this means is that agent's movement decisions are not independent of one another.

Last but not least, time is discretized; at every discrete time interval (could vary from minutes to years), agents make movement decisions by considering either historical or current network-level observations.

2.1.2 Model Design

When developing an agent-based model (with characteristics specified), it is important to consider the following series of questions, which will help in the initial model design:

1. What is the scale of network that we are trying to model? (i.e., macro or micro-level)

2. What is the data observability? (i.e., partial or fully observable)

3. What are the agents in the model? Are the agents individuals, entities, corporations, just to name a few.

4. What are the factors or variables that might affect an agent's movement decision, i.e., the utility function?

Answering these questions are essential in order to come up with an initial agent-based model design for development.

2.2 Notations

We highlight some common notations in the class of agent-based models described in section 2.1.1. We define a common set of travel options by N, which is commonly used for representing the set of nodes in a network, although it could also represent travel options at other levels, such as links. The origin node (agent's current node at a discrete time interval t) is represented as node i, while the destination node (node that agent is considering to travel to) is represented by node j. Movement in the network is represented by $P^t_{i,j}$, which is the probability of moving from a node i to j at discrete time interval t. Note that i and j could also take in other forms, such as moving from zone i to j, edge or road i to j, just to name a few. $U^t_{i,j}$ is utilized for modelling the utility that an agent gets by travelling from node i to j at discrete time interval t, which corresponds to $P^t_{i,j}$. The utility function could consist of various factors such as historical or current node-level observations, along with the distance to travel from node i to j.

Another form of movement representation would be $P^t_{d,n}$, which is the probability of an agent d travelling to node n at discrete time interval t. This utility here is represented by $U^t_{d,n}$, which considers 1) node-level information (similar to $U^t_{i,j}$), as well as 2) agent-level information (e.g., agent's preferences). For clarity, the agents movement model in chapter 4 takes on the $P^t_{d,n}$ form (node and agent-specific), while the movement model in chapters 5 and 6 takes on the $P^t_{i,j}$ form. This will be further

elaborated in the individual chapters.

2.3 Proposed Methodological Framework

In this section, we describe the proposed methodological framework for developing high quality agent-based models in various problem domains under the conditions of network model scale (micro or macro) and data observability (partial or full observation). For each problem domain, we describe the potential issues surrounding them, and provide ways to overcome these domains-specific problems. Then, in chapters 4, 5 and 6, we provide concrete use cases for each problem domain, and illustrate our proposed methodological framework in these use cases. The mapping of use cases to problem domains are as illustrated in figure 2.1. However, these guidelines are definitely *non-exhaustive*; certain components such as the utility function (factors contributing to the agents movement) would differ on a case-by-case scenario, even for the same problem domain as specified. Additionally, it is important to note in figure 2.1 that the 'Macro-level, Fully Observable' problem domain is left blank (indicated in grey); to the best of our knowledge, there is no use case that falls in this category.

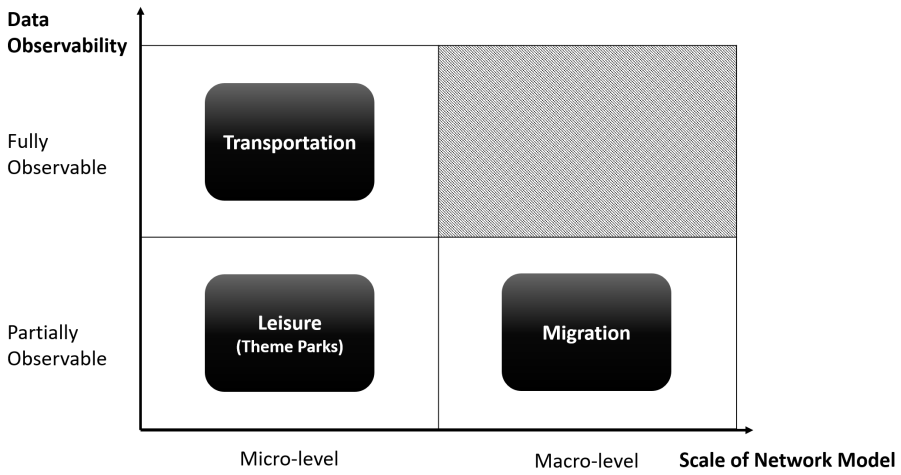

Figure 2.1: Problem domains: network model scale and data observability.

2.3.1 Micro-Level, Partially Observable (MiPO)

For the first problem domain (MiPO), the main challenge faced would be data observability, where data is observable for only a small subset of population that we are trying to model. In this case, the question here would be whether the agent-based models built from the small set of micro-level data would be representative of the entire population.

The process flow for users developing agent-based models in this problem domain is as follows:

1. Calibrate utility-driven decision model based on a small set of (observable) micro-level data.
2. The utility-driven decision model will be input to an agent-based simulation model.
3. Make use of the agent-based simulation platform as a ground truth generator to generate the movement data for an entire simulated population.
4. Compare or validate the aggregate-level observations from the simulation model against that of actual ground truth data. Some common approaches would be to validate via sum of errors or weighed sum of errors.

In addition to the following points, a user would also need to consider the factors that might contribute to the utility function of an agent. A concrete example of the application of this approach is in chapter 4.

2.3.2 Micro-Level, Fully Observable (MiFO)

The micro-level, fully observable (MiFO) problem domain is defined by two conditions: (1) micro-level network model, and (2) data is observable for the *entire population* that we are modeling. In such instances, the data

collected, while being of a larger quantity vis-à-vis that of MiPO, might not be as qualitatively similar.

The process flow for users developing agent-based models in this problem domain is as follows:

1. To deal with the noisier set of data, we will introduce some pre-processing steps, either for 1) cleaning of data or 2) intent inference. A concrete example of this will be elaborated in chapter 5.

2. Calibrate utility-driven decision model based on the processed data.

3. The utility-driven decision model will be input to an agent-based simulation model.

4. Compare or validate the aggregate-level observations from the simulation model against that of actual ground truth data. Some common approaches would be to validate via sum of errors or weighed sum of errors.

In addition to the following points, a user would also need to consider the factors that might contribute to the utility function of an agent. A concrete example of the application of this approach is in chapter 5.

2.3.3 Macro-Level, Partially Observable (MPO)

The macro-level, partially observable (MPO) problem domain is a special scenario where it is impossible to model the movement decisions of individual agents in the model. This can be explained from a computational perspective, where there are too many agents and their decisions to compute, as well as from a data perspective, where it is impossible to get the individual-level data.

The process flow for users developing agent-based models in this problem domain is as follows:

1. Model simulation agents as entities, which could be corporations, countries, just to name a few. In this case, the flow between the agents (i.e., entities) would be people from one entity to another.

2. Calibrate utility-driven decision model based on the macro-level data.

3. The utility-driven decision model will be input to an agent-based simulation model. A thing to note about the agent-based model here is that while it is a macro-level model, it is by no means a simplistic representation of the movement and underlying dynamics. This is ensured by having sub-populations at agents-level. In this case, the user would need to consider what kind of sub-populations to model, such as departments, sections, race or language speakers, just to name a few. A concrete example of this will be applied in chapter 6.

4. Compare or validate the aggregate-level observations from the simulation model against that of actual ground truth data. Some common approaches would be to validate via the sum of errors or the weighed sum of errors.

In addition to the following points, a user would also need to consider the factors that might contribute to the utility function of an agent. A concrete example of the application of this approach is in chapter 6.

Chapter 3

Related Work

3.1 Agent-Based Models

One of the classical use of agent-based model (ABM) would be the segregation model [67] by Thomas C Schelling. It involves simple rule-based mechanisms that is based on the Cellular Automata (CA) model [31], where agents would move out of their current dwelling zone if the proportion of racially similar neighbors exceed a certain threshold. It was a simple but representative example of how agents with simple rule-based decisions can produce complex emergent phenomenon at aggregate levels. Since then, ABMs have evolved a lot in terms of complexity, and is utilized in many fields, ranging from political science, sociology, epidemiology, finance, environmental science, and economics [77]. In this dissertation, we are interested in agent-based models involving movement. This is defined as movement of agents moving from a node i to j.

In the previous chapter, we propose a methodological framework for modeling movement decisions of agents in a network environment. We further discussed the use of the discrete choice model approach under various problem domains, by exploring the interaction of network model scale and data observability (illustrated in figure 2.1). In this chapter, we

review several modeling works according to the relevant categories (MiPO, MiFO, MPO) in our framework. We then highlight the main differences between the existing literature and our approach for modeling under the three different categories as specified in figure 2.1.

For the relevant domain-specific (i.e., leisure, transportation, and migration) references, they will not be discussed in this chapter. Instead, they will be as discussed subsequently in chapters 4, 5, and 6.

3.2 Micro-Level, Partially Observable Domain

Example of models in the Micro-Level, Partially Observable (MiPO) category would be epidemic models (e.g., smallpox containment models [28]), civil violence models [28], and revolution models [29], e.g., the 2011 Arab Spring. Another example would be crowd simulation models, such as for modeling of egress in emergency events [62], which adopts a multi-tiered agent behavior, ranging from locomotion, to steering and social (in increasing complexity).

For the aforementioned models, common approaches would be to (1) use the existing data to model a small scale network that is not representative of the entire population, or (2) skip the data completely, and create purely "illustrative" models to highlight various phenomenons under certain network conditions. This differs from our approach whereby we (1) collect data from a sampled population, (2) use an agent-based model as a ground truth generator, and (3) validate aggregate-level observations from simulation runs against that of actual ground truth. This results in high agent-based simulation models that serve as *realistic* ground truth generators that is based on data from a small set of the entire population.

A relevant recent project worth pointing out would be the Ground Truth program [24] by the Defense Advanced Research Projects Agency (DARPA). As mentioned in [23], the social science field has long been faced

with technical and logistical limitations when it comes to studies of large, representative populations. What the Ground Truth program aims to do is to validate the effectiveness of various social science modeling methods through the use of artificial computational social system simulations with built-in "ground truth" causal rules [24].

Essentially, this means that simulation models are used as ground truth generators, in the hope that it can effectively 'reverse engineer' artificial social systems through accurate identification of the causal rules in the simulation itself [24]. This is similar to our approach for modeling under the MiPO category, where we utilize a small set of sample data from a subpopulation, and utilize the utility-driven decision model calibrated from the sample data as input to a simulation model that models the movements for an entire representative population. With this, we hope create models that serve as effective ground truth generators in the event that micro-level data is not observable for an entire representative population.

3.3 Micro-Level, Fully Observable Domain

Models in the Micro-Level, Fully Observable (MiPO) category mainly revolve around domains where (1) data is readily observable and/or (2) sensors are present and can easily capture data of an entire population. It is made possible through the ubiquitousness of mobile and sensing devices, which allows researchers to develop smartphone applications (e.g., for Android and iOS) to capture user Global Positioning System (GPS) locations without major human interventions. This allows models to be built revolving around traffic simulation [8] and train transportation networks [66].

A common concept revolving around such models would be the Origin-Destination (OD) matrix and flows, which is essentially a representation of the flow of traffic or vehicles travelling from an origin to destination

location. A normalized representation would be through transition probabilities matrix, which is the probability of travelling from an origin to destination. In our work, we adopt a two-tiered decision model for modeling the movement of taxis in a road network. While this increases the realism of the system, having a more complex representation (compared to a single-tier OD representation) also leads to the challenge of figuring the intent of drivers, especially at the zonal level decision. This will be further discussed in chapter 5.

3.4 Macro-Level, Partially Observable Domain

A common macro-modeling approach would be through the use of Dynamic Programming (DP) models [7]. It is a mathematical optimization and computer programming method that simplifies complicated problems by breaking them into simpler sub-problems in a recursive manner. In such an approach, systems within the environment are modeled as a whole, without any individual elements.

Another way of modeling macro-level models is through Principal-Agent Models [42], which generally is applied in the area of political science and micro-economics. Actors in the system are divided into agents and principals, where agents appoint the principals to make decisions on their behalf. This is generally used in modeling of high level macro phenomenon, such as the interactions between politicians (agent) and voters (principal), corporate management (agent) and shareholders (principal).

A major difference between the aforementioned approaches and our work is that the models mainly model group and actors as homogeneous entities, with no individual elements. In our work, even though we model agents as entities, we incorporate some form of individual element via several distributions, for example, age and origin-destination distributions.

This means that we model heterogeneity through the modeling of sub-populations, instead of having homogeneous population pools. Not only does this help in creation of realistic models, it also serves as useful information for policy makers, where they can observe shifts in the distribution of entities (which could be corporations or countries) through certain policies or implementations.

Chapter 4

Micro-level, Partially Observable: Leisure Domain

Overview

In this chapter, we look into the development of high quality agent-based simulation models in the micro-level, partially observable domain. A representative use case here would be the leisure domain, in particular the modeling of agents at theme parks. Theme parks are a major driver of the global leisure and tourism sector. However, theme parks are known to suffer from congestion problems, where visitors could spend up to two hours waiting for popular attractions. This can be attributed to lack of (1) global queue time information, and (2) central congestion management mechanism, which can be costly to implement and test on a theme park wide scale.

An alternative would be a simulation-based approach, which serves as an economical approach to evaluate congestion management strategies, but these models often face the challenge of low data availability, where micro-level movement data are obtained via sample-based mobility studies - raising questions on the representativeness of such data. To this end, we

collected movement of patrons at a theme park to build a baseline visitor movement model. Then, utilizing an agent-based simulation model as a ground truth generator, we tested out a sample-based approach, similar to that *on-the-ground*. By comparing with a baseline approach (diffusion dynamics model), our results have shown that even with a small sampling rate of 1%, the sample-based approach is able to capture the movement of patrons at a theme park fairly well, which supports the use of sample-based mobility studies. Through two additional case studies, we demonstrated the effectiveness of an agent-based simulation model as an evaluation tool for (1) a dynamic route guidance application and (2) effect of spatial layout on crowd distribution.

4.1 Introduction

Theme parks play an important role in the global tourism and entertainment industry. However, with the rise of theme parks as an important driver of tourism, this has also led to congestion and an increase in queue times at theme park attractions, which could be as long as two hours for the popular attractions. Consequently, this has led to a decrease in experience for theme park visitors.

Congestion at theme parks can be attributed to: (1) a small set of popular (or star) attractions being preferred by all visitors, (2) the spatial layout of theme parks that leads to bottlenecks at several path segments, and (3) lack of global information on attraction wait times.

Different strategies have been adopted by theme park operators to deal with the long waiting lines. An example would be the FASTPASS (implemented at Disney parks), which incentifies visitors to visit popular attractions at a later time of the day - through the use of express passes as a reward. Another example would be the Express Pass implemented

at Universal Studios parks, which allows visitors to utilize priority queues by paying for it, thereby reducing their wait time at attractions. Other ways includes having street shows and character greetings, so as to prevent visitors travelling from different directions to converge at potential bottleneck (congested) regions.

While there has been experience management strategies in place to deal with congestion and crowd management, it is often costly to test out and evaluate such measures at a large scale prior to implementation, due to logistic and financial reasons. An alternative to this would be simulation-based approaches, where we could make use of agent-based simulation models as an evaluation tool for testing out such experience management strategies prior to implementation. However, a challenge here would be data availability, as it is often costly and logistically impossible to obtain comprehensive micro-level data at a large scale (entire theme park). To this end, we introduce SimLeisure, a agent-based simulation model for two purposes: (1) to serve as a ground truth generator in the scenario of a lack of individual agent-level trajectory based data, and (2) to serve as an evaluative tool to test out a experience management strategies, as well as the effect of spatial layout on crowding.

Our approach is as follows. Firstly, we conducted mobility surveys to collect micro-level visitor data, which includes raw Global Positioning System (GPS) trajectories and QR scan records. Then, with the combination of micro and macro level (i.e., queue time at attractions) obtained from our theme park operator, we developed a baseline behavioral model for movement around the theme park. As micro-level data is not available for the entire theme park visitor population, we input the behavioral model into an agent-based model, and made use of the agent-based model as a ground truth generator. We then proceed to evaluate the sample-based approach by comparing with a baseline approach, the diffusion dynamics

model.

Lastly, we included two additional case studies to demonstrate the effectiveness of an agent-based simulation model as an evaluation tool for a dynamic route guidance application, and the effects of spatial layout on crowding.

4.2 Literature Review

In this section, we review some of the existing literature in the area of agent-based models and congestion management for theme parks, and how our work differs from them.

4.2.1 Agent-Based Model as Ground Truth Generator

In our work, in order to deal with the issue of low data availability, we made use of an agent-based simulation model as a ground truth generator, where we investigate the effectiveness of a sampled-based approach (i.e., sampling of population in the selection of participants) which exists in typical ground surveys. A similar effort would be the the Ground Truth (GT) program ([21] and [22]) by the Defense Advanced Research Projects Agency (DARPA), where the aim is to "develop models that accurately 'reverse engineer' an artificial social system by correctly identifying the causal rules designed into the simulation". Another notable project from DARPA is the Next Generation Social Science (NGS2) program ([20] and [19]).

4.2.2 Modeling and Optimization: Theme Park Domain

[58] gave a overview of agent-based models and their advantages, as well as how they can be applied in the tourism industry. They also included a survey of applications of Agent-Based Modeling in Tourism. [82] adopted

a mixed methods approach combining statistical and spatial analyses to examine theme park visitors movement - done through utilizing theories on intuitive and rational choice, and looked into both attraction attributes and spatial layout attributes. [72] implemented an ACS-ILS algorithm to find best locations of RFID readers. [12] looks into crowd control mechanism, through the use of mobile devices. [65] looks at the impact of recession to national parks visitation. [83] focuses on the theme park carrying capacity.

4.2.3 Behavioral Modeling: Theme Parks and Beyond

[4] focuses on the aspect of revisiting intention, which could be an interesting aspect to explore in future. [69] looks into amusement park attributes by patrons of theme parks, which is similar to the thrill, dark and wet aspects of attractions that we implemented in our model. [9] and [26] are also similar works on the utility model that we are using for visitors attraction attendance. [56] looks into behavioral modeling for Scottish theme park visitors. [1] evaluates of the attractiveness of a new theme park. [53] looks into factors affecting attendance. [39] predicts how long theme park visitors would spend at attractions, through the use of a ordered logit model.

4.3 Methodology

In this section, we describe the data that is utilized in building the behavioral model for theme parks visitors, followed by the processing, and evaluation of the sample-based approach as proposed.

4.3.1 Data Collection

For the purposes of this study, we partnered with a theme park operator at a major Asian city. The attraction-specific attributes (thrill, dark and wet) were also decided in consultation with the theme park operators. Data

utilized in this study is divided into two categories, micro and macro-level data.

Macro-level data is directly obtained from our partner, which consists of attraction wait time collected at 30 to 90 minutes interval (time period of September 2011 to August 2012). This data is utilized in the dynamic route guidance application, as well as utilized by the diffusion dynamics model (discussed subsequently) for estimating visitor transition probabilities.

For micro-level data, we do not have comprehensive data on visitors and their preferences, along with their decisions (e.g., queuing up at attractions, movement, etc.) - all of which is required for us to construct the visitors behavioral model. To do this, we conducted a mobility survey (December 2012) for a total of 50 participants, where they would download a mobile application (developed in-house) prior to a visit to the theme park. The interface of the mobile application is as illustrated in figure 4.1.

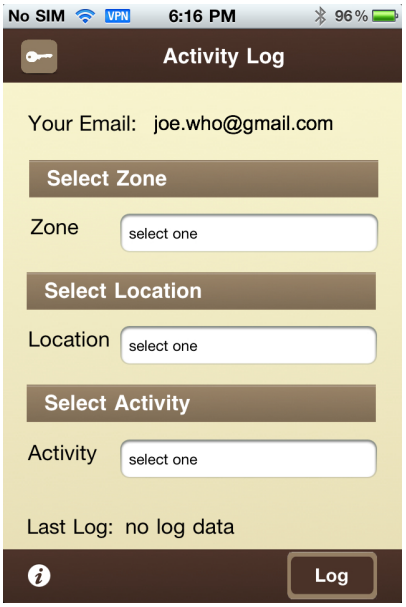

Figure 4.1: Interface of Mobile Application for Logging of Visitor Activities.

With the mobile application, we collect the GPS (Global Positioning System) traces of the participants throughout the theme park, along with attraction attendance (participants would scan a QR code prior to entering an attraction). We would also collect other details such as the group type

(individual or family) and preference (towards thrill, dark, wet rides - on a scale of 0 to 3).

4.3.2 Data Processing

With the data collected, we would still require some processing to infer other information which is not available in the existing dataset. Additionally, we processed filtering to retain only 17,534 data points (for analysis) out of the 86,376 data points collected, with data points removed for reasons such as duplication, *out-of-bounds* error (due to poor GPS reception and other reasons), as well as those with timestamps that were outside of the theme park operating hours.

As mentioned previously, we captured the participants entry into an attraction via means of QR scans, where participants will have to scan a QR code before they entering an attraction. However, while we do have information on the entry to attractions, we do not have information on whether a visitor joined the queue for an attraction. More specifically, in the event that a participant joins the queue for an attraction, but subsequently decides to leave the queue, there will be no data captured on our end (as no QR scan is performed). In order to infer this information, we first divide the trajectories into thirty minutes intervals, and get the region (i.e., zone) that a participants spends the most time in (for that time interval). We subsequently perform a match between the attraction that the patron spends the most time during that time interval, and the actual QR scans that is performed by that patron. If there is a match, it means that the patron went to queue at the attraction, and entered the attraction. Otherwise, it meant that the patron did queue at the attraction, but left after waiting for some time. This is an important piece of information that is subsequently used in the behavioral model (logit choice model).

4.3.3 Behavioral Model for Theme Park Visitors

In this subsection, we describe the behavioral model for theme parks visitor agents. To model the movement decisions of visitors at theme parks, we used a form of discrete choice model, the logit model. The logit choice probability model is one of the most widely used discrete choice model, and chosen as the formula for the logit choice probability model takes a closed form and is readily interpretable [71]. Following consultations with the theme park operators, we decided on the factors thrill, dark, wet levels, along with distance (to travel) for an agent's movement decision.

Formally, the utility obtained by a visitor v going to attraction n is defined as:

$$U_{v,n} = \beta_1 T_{v,n} + \beta_2 D_{v,n} + \beta_3 W_{v,n} + \beta_4 \delta_{v,n} + \epsilon_{v,n}$$

where $\epsilon_{v,n}$ denotes unobserved factors that contribute to visitor's decision. Also, $\delta_{v,n}$ represents the distance a visitor has to travel from his/her current location to the attraction n, while $T_{v,n}$, $D_{v,n}$, and $W_{v,n}$ represents the absolute difference in thrill, dark and wet preferences between the visitor and that of an attraction. For example, for a thrill level T_v and T_n for a visitor v and attraction n, the difference is formally defined as:

$$T_{v,n} = |T_v - T_n|$$

With the utility level $U_{v,n}$ defined, the probability of visitor v choosing to travel to an attraction n is then formally defined as:

$$P_{v,n} = \frac{e^{U_{v,n}}}{\sum_{j \in N} e^{U_{v,j}}}, \forall v, n$$

We then proceed to generate the itineraries for the agents as follows. As the preference for an entire theme park population is unknown to us, we proceed to generate C number of visitors class, with each class representing a (randomly generated) user preference (thrill, dark, wet). Thereafter, with the logit choice model defined previously, we generate the itineraries of visitors *sequentially* (attraction-by-attraction), up till a time threshold is exceeded. In this case, an attraction is randomly selected, and probability of the attraction being added would be as defined by the probability $P_{v,n}$ defined previously. Also, we set a time budget of 320 minutes, which is the average time that a participant spends in the theme park, and the total time incurred per itinerary should not exceed the (expected) time spent visiting all attractions. We proceeded to generate N_c itineraries per visitor class, and the final number of itineraries generated would be $N_c \times C$. During the course of simulation, an agent would then select at random one of the itineraries at the point of initialization.

4.3.4 Agent-Based Model as a Ground Truth Generator

As mentioned previously, it is practically (logitically and financially) impossible to obtain comprehensive data of an entire theme park patron population - which is why we used an agent-based simulation model ([16] and [45]) as a ground truth generator to evaluate the sample-based approach. The agent-based model is developed in NetLogo 5.1, and includes additional customization (Java extensions and external files) for agents behavior and changing of physical environment (e.g., distance matrix generation) in the simulation. This makes the model highly configurable, and it can be

changed easily to fit the landscape and spatial layout of other theme parks. The interface of the model is as shown in figure 4.2, where the colored bars are used as indication of the level of crowding at attractions.

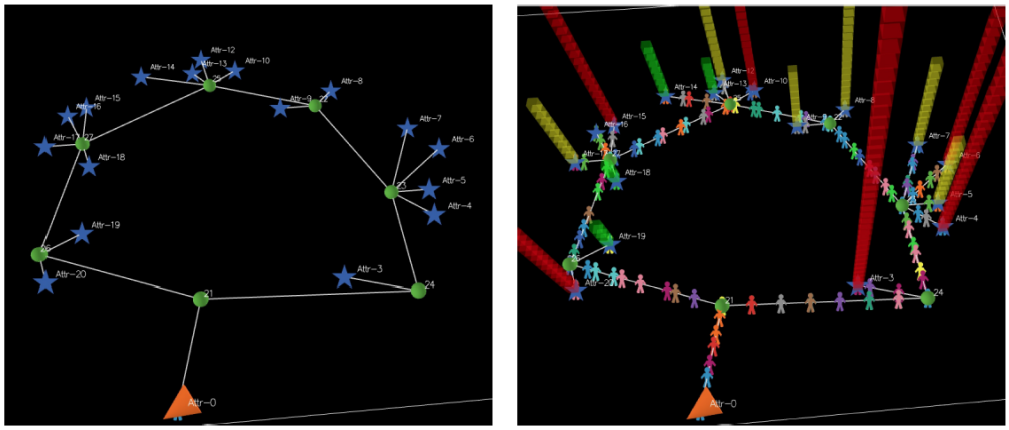

Figure 4.2: Graphical interface of simulation model at (left) initialization and (right) during course of simulation. Colored bars are indicative of crowding at attractions.

Some additional details and features of the simulation model is as follows:

- The model allows for intervention schemes such as street shows (similar to that found in theme parks) and break down of attractions. This allows policy makers to test out various *what-if* scenarios via the simulation model.

- Every agent has its agent-level attributes, such as profile (individual, family), preference (thrill, dark, wet), time spent waiting at attractions, number of attractions visited, as well as (depending on agent type) a list of attractions to follow. These allow us to easily track the performance of agents subsequently.

- Every attraction has its attraction-specific attributes, such as service rate, service capacity, as well as queues. Each attraction would manage a queue, which is essentially a list of visitor ID that are waiting in line, managed in a FIFO (first in, first out) manner. Also,

depending on the queue length and service rate, this would allow us to compute the estimated waiting time at attractions, similar to that *on-the-ground*.

- Additionally, the model allows for priority or express queues, similar to that found in theme parks. Under the express queue scheme, we could customize certain groups of visitors (who purchased express tickets) to utilize a separate queue, while other visitors would utilize the normal queue. It could also potentially be used in certain incentive schemes, where patrons who follow recommendations from a central system could be rewarded with express queue tickets.

It is also important to position the agent-based model as a mesoscopic model, between that of a macroscopic and microscopic model. This means that it captures some of the micro-level details such as agent-level decision and queues at attractions, while other details such as movement along links (i.e., streets at theme park) and visiting of restaurants/restrooms are excluded, so as to reduce the model complexity. With the agent-level decision and queue at attraction incorporated, it would then allow us to observe emergent behaviors, such as congestion at attractions, which is a classic reason for utilization of agent-based modeling and simulation ([11]) techniques. This also separates our work from previous works ([54] and [55]) in theme park operations simulation, which utilizes only macro-level data for the simulation model.

4.3.5 Comparison Baseline: Diffusion Dynamics Model

As mentioned previously, we compare the results of the sample-based approach to that of the diffusion dynamics model, which is utilized to derive flow patterns using observed congestion (i.e., visitor queues) at attraction nodes [27]. The diffusion dynamics model is an optimization model that learns

the probability of a visitor going from node i to node j by utilizing waiting times at attractions. This allows us to overcome the issue of a lack of actual on-the-ground movement data for comparison purposes (with the sample-based approach). Additionally, the output (time-dependent transition matrix $P^t_{i,j}$) from the diffusion dynamics model is identical to our approach. Last but not least, the diffusion dynamics model has been shown to provide prediction accuracy of up to 80% for the popular attractions, making it a good baseline comparison approach.

The diffusion dynamics approach ([27]) models the movement of visitors as a multinomial distribution based diffusion model, which is a form of dependent cascade model. Formally, the likelihood is defined as:

$$\mathcal{L}(p; x, n) = \prod_{d \in D} \prod_{i \in A} \prod_{t \in T} \frac{(\sum_j x_{d,t,i,j})!}{\prod_{j \in A} x_{d,t,i,j}!} \prod_{j \in A} p_{t,i,j}^{x_{d,t,i,j}}, \tag{4.1}$$

where the total outflow of visitors from attraction node i at time t is $\sum_j x_{d,t,i,j}$, which corresponds to the total number of trials in the multinomial distribution.

Variable	Description
D	observed cascades
A	set of all attractions in the theme park
T	set of time slices
S_i	service rate at attraction i
$n_{d,t,i}$	number of visitors waiting at node i, time t in cascade d
$x_{d,t,i,j}$	corresponds to the number of people moving from node i to j
$p_{t,i,j}$	probability of a visitor moving from node i to node j at time t

Through the modeling of visitors movement, the probability of a visitor moving from node i to j at time t ($P_{i,j}^t$)) is then estimated (through a maximum likelihood estimation) from aggregate (queue-time) observations, which are an output of the simulation model (ground truth generator). The final probability matrix $P_{i,j}^t$ for all approaches (sample-based and diffusion dynamics) is then evaluated for performance comparison. This is as discussed subsequently.

4.3.6 Performance Measure

To compare the performance of that of the sample-based approach versus that of the diffusion dynamics model, we would first define two performance measures, a *sum of errors*, as well as a *weighted sum of errors*. Formally, the sum of errors performance metric is defined as:

$$SE_a = \sum_{i,j,t} |P^t_{i,j} - P^{a,t}_{i,j}|$$

where SE_a represents the sum of error for an approach a, which could be (1) the diffusion dynamics model, (2) 5% sampling approach, or (3) 1% sampling approach. Correspondingly, $P^{a,t}_{i,j}$ represents the time-dependent transition probability of going from an attraction node i to j at discrete time interval t, for one of the three approaches a. $P^t_{i,j}$ represents the time-dependent transition probability for the entire population (generated ground truth).

The next measurement is weighted sum of errors. It is a weighted variant of the sum of errors, where we multiply the error for a $< i, j, t >$ tuple with the flow $F^t_{i,j}$ of the entire population (generated ground truth), which penalizes errors that are made at links with high transition flows. Therefore, this places higher emphasis on prediction accuracy for links with high visitor flow between them. Formally, the weighted sum of error for an approach a is defined as:

$$WSE_a = \sum_{i,j,t} |P^t_{i,j} - P^{a,t}_{i,j}| \times F^t_{i,j}$$

Similar to the sum of errors, approach a could take any one of the three approaches (mentioned previously). Results for both measurements will be presented subsequently in section 4.4.

Input Variables	**Description / Values**
Thrill, Dark, Wet	[0,3]
Number of Visitor Classes	10
Number of Itineraries per Visitor Class	10
Number of Visitor Agents	15,000
Detour Threshold	320 minutes
Dependent Variables	**Description / Values**
Transition Matrix	$P^t_{i,j}$.
Number of Replications	30

4.4 Results

4.4.1 Sum of Errors and Weighted Sum of Errors

The results of both the sum of errors and weighted sum of errors at varying time intervals (30 minutes, 1 hour, 2 hours) is as shown in table 4.1. From table 4.1, we can observe that the sample-based approach (both 5% and 1% sampling rate) outperforms that of the diffusion dynamics model, although the 5% sampling rate approach gives the best results.

Additionally, not only does the sample-based approach outperform that of the diffusion dynamics model, it does so especially well at the links with high visitor flows, as observed by the lower weighted sum of errors in table 4.1. This is even at a low sampling rate of 1%, where the sample-based approach outperforms that of the diffusion dynamics model.

An additional point to note here would be the performance of the sample-based approach dropping with decreasing time interval (from 2 hours to 1 hour to 30 minutes), which can be explained by a decrease in data points available per time interval, thereby resulting in poorer prediction accuracy.

To further illustrate the performance of the sample-based approach against that of the diffusion dynamics model, we adopt a heat map visualization for comparing the results obtained for all approaches. This is done by representing the transition probabilities obtained for an approach a, for

Table 4.1: Sum of Errors (SE) and Weighted Sum of Errors (WSE) result at different time intervals (30 minutes, 1 hour, 2 hours). Best performing approach for each setup indicated in bold, and standard deviation in brackets.

		5% Sampling	1% Sampling	Diffusion Dynamics
SE	30 min	**49.77 (0.04)**	101.08 (0.06)	362.82 (0.11)
	1 hr	**16.02 (0.02)**	35.25 (0.04)	176.85 (0.1)
	2 hrs	**5.6 (0.01)**	12.04 (0.02)	85.54 (0.11)
WSE	30 min	**18,260.35 (0.04)**	38,106.3 (0.06)	170,084.54 (0.11)
	1 hr	**12,889.96 (0.02)**	26,096.17 (0.04)	169,722.64 (0.1)
	2 hrs	**9,068.67 (0.01)**	18,476.5 (0.02)	173,172.33 (0.11)

$< i, j >$ tuple as $P^a_{i,j}$. This is essentially an aggregation of the time-dependent transition matrix $P^{a,t}_{i,j}$, and is formally defined as $P^a_{i,j} = \sum_t P^{a,t}_{i,j}$.

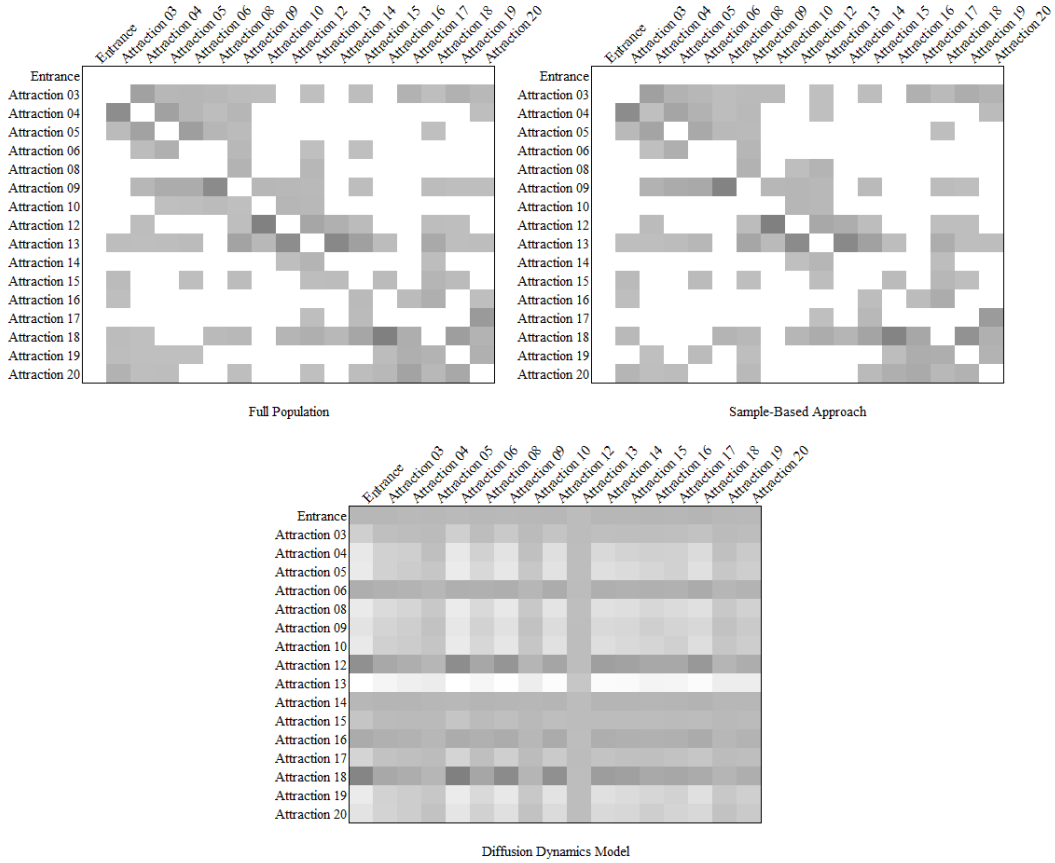

Figure 4.3: Heat map of transition probabilities for full, sampled and diffusion dynamics model. Darker shades indicates higher transition probabilities.

Figure 4.3 shows the visualization for the (1) full population, (2) sample-

based approach, as well as the (3) diffusion dynamics model. Attractions with high transition probabilities are indicated in darker shades of gray, and vice versa. From figure 4.3, we can observe that the sample-based approach performs well in representing the transition movements of the entire population, as observed by the closer color representation. This further confirms the usefulness of sample-based mobility surveys in capturing the movement decisions of theme park visitors.

4.4.2 Sensitivity Analysis: Sampling Rate

As noted in table 4.1, having a sampling rate of 5% would yield a better prediction accuracy as opposed to just having a sampling rate of 1%. This is an expected outcome, as a larger sample population would definitely represent an (entire) population better. However, given that sample-based mobility studies often involve huge logistical and financial costs, it would be beneficial for researchers to find a "good sampling rate", one which gives them the best value in terms of (1) data representativeness and (2) financial and/or logistical costs involved.

To this end, we conducted sensitivity analysis tests to further look into the effect of sampling rate on the performance of the sampling-based approach. Virtual experiments were conducted with sampling rates of 1% to 9% (increments of 1%), followed by 10% to 90% (increments of 10%). Results of the virtual experiments are as summarized in figure 4.4.

From figure 4.4, we can observe that when increasing the sampling rate from 1% to 10%, we get the greatest gain in terms of performance (sharpest decrease in sum of errors and weighted sum of errors). From 10% onwards to 90%, we continue to obtain gains in performance, but the rate of performance increase is not as pronounced as that before 10%. This is an important finding for researchers conducting sample-based mobility studies, and are looking to make informed decisions on the sampling rate.

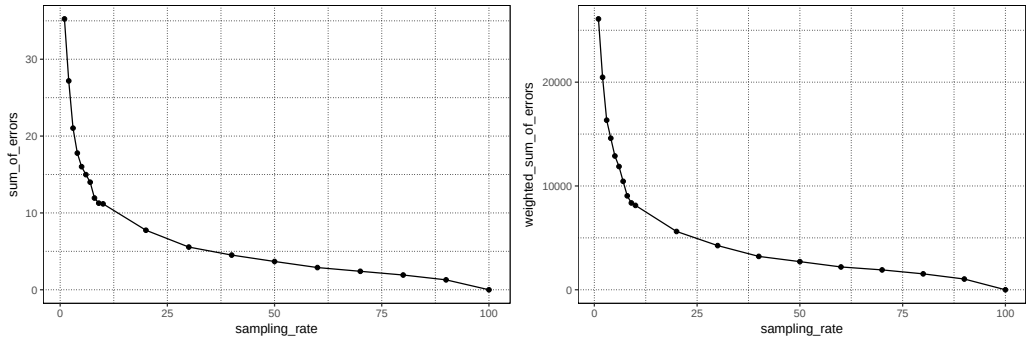

Figure 4.4: Sensitivity analysis results for (left) sum of errors and (right) weighted sum of errors under varying sampling rates.

4.5 Case Studies: Agent-Based Model as an Evaluation Tool

In this case study section, we present two case studies on the use of an agent-based model to evaluate (1) a dynamic route guidance application and (2) effect of spatial layout on crowds in theme parks. These will be further elaborated in the subsequent subsections.

4.5.1 Case Study 1: Evaluation of Dynamic Route Guidance App

In the first case study, we provide a use case for an agent-based simulation model in theme park visitor experience management, where it serves as an evaluation tool ([16]) for a dynamic route guidance application that is developed in-house. This case study serves as an illustrative example for future works on the usage of agent-based models as an economical approach to evaluate various experience and congestion management strategies, in both theme park and other crowd simulation scenarios.

4.5.1.1 Motivation: Attraction Wait Times

We first present some findings from the attraction wait times obtained (period of September 2011 to August 2012) from our theme park operator

partner, which subsequently serves as a motivation for the development of the dynamic route guidance application. The first finding is as summarized in table 4.2, which contains the average wait times at attractions during peak days, sorted (descending) by wait times. From figure 4.2, we can observe the disproportionate wait times at attractions, where the most popular attraction (attraction T) experiences 67% higher wait times than the second most (attraction C) popular attraction.

Table 4.2: Average wait time (in minutes) at attractions on peak days.

Attraction	**Avg wait time**
T	82.0
C	49.1
H	40.0
J	37.0
BH	31.4
BC	26.2
S	26.1
M	25.8
D	24.9
L	22.7
R	20.7
A	12.2
E	11.7
P	10.2
K	7.8

To offer another perspective, figure 4.5 shows the wait times for the four most popular attractions throughout the day. Combining the results from table 4.2 and figure 4.5, we can observe that while attraction T experiences 67% higher (on average) wait times that the second-most popular attraction, there are times of the day where the wait time drops, not just lower than the second-most popular attractions, but even lower than the fourth-most popular attraction (J). This is an important finding, as not only does this confirm on the strategies adopted by theme parks (e.g., FASTPASS by Disney parks to encourage visitors to visit popular attractions at later times of the day), it also supports the usefulness of a dynamic route guidance

application in this case, so as to exploit the peak and non-peak periods of attractions at various times of the day.

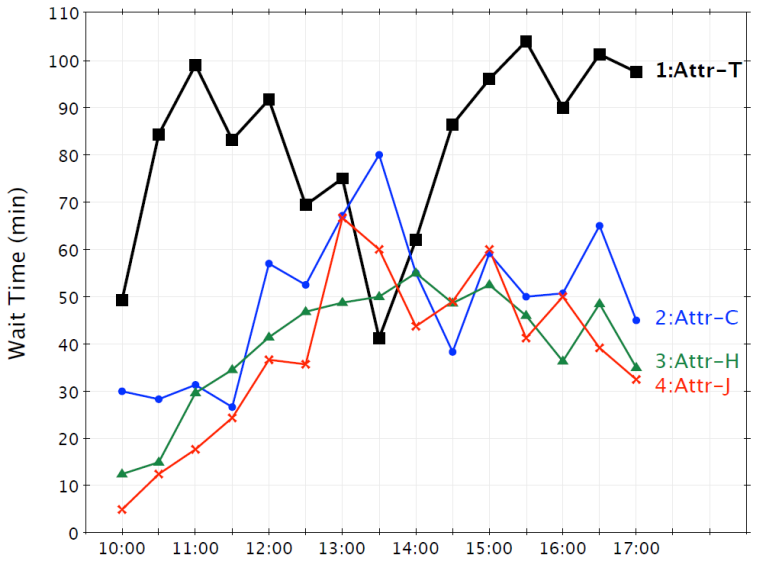

Figure 4.5: Wait time for four most popular attractions throughout the day.

4.5.1.2 Visitor Experience Management: Dynamic Stochastic OP

In order to design an application for visitor experience management in theme parks, we first look at the aim of providing personalized route guidance to theme park visitors as a Dynamic Stochastic Orienteering Problem (DSOP) ([43]). Applying a standard Orienteering Problem (OP) ([73]) to the case of theme park visitors, the aim here would be to find a series of nodes (i.e., attractions) in a graph to traverse that gives the maximum reward (to the visitor), while ensuring that the total distance or time taken to traverse those nodes does not exceed a threshold level. In this case, the user reward gained is dependent on the user-specific preferences (profile), as well as the total time spent (1) queuing at attractions and (2) traveling between attractions.

The DSOP is a (single-agent) variant of the OP, with the difference being that the attraction queue times and travel times from attraction

to attraction are non-deterministic and dynamic (varies with time). Also, each attraction has a attraction-specific operating time window and status, which is dynamic and may change according to weather, technical failures, just to name a few. The aim here is similar to that of the OP scenario, while ensuring that the time period of visiting an attraction node falls within the attraction-specific operating time period.

The dynamic route guidance application comes in two modes, a generation mode (for generating an initial list of attractions to visit), as well as an intervention mode, where, as a feedback to a change in the environment (e.g., ride breakdown and other anomalies) - will result in a new list of attractions being regenerated. Additionally, it is important to point out that this model adopts a slightly different variant of the visitors movement model (as opposed to that in section 4.3), which involves 17 zones, each of which represents a major attraction at the theme park. The visitor movement model is then simply a Markovian representation $P^t_{i,j}$ of visitors' movement from a current zone i to the next zone j - and is dependent only on the current zone of a visitor. This is as opposed to the logit choice model itinerary generator that was discussed in section 4.3.

4.5.1.3 Virtual Experimental Design

The design of the virtual experiment is as follows. First, we define two key performance indices that will ultimately affect a visitor's experience in a theme park: (1) total time spent waiting at attractions, and (2) number of attractions visited. Also, we define two types of agents, a guided one (i.e., visitors following a list of recommended attractions from the dynamic route guidance app), as well as an unguided one (following the visitor movement model - for comparison with the guided agents). Simulation instances were run with 15,000 agents, and we conducted sensitivity analysis to evaluate the performance of the dynamic route guidance application under various

percentages of guided visitors.

4.5.1.4 Virtual Experimental Results

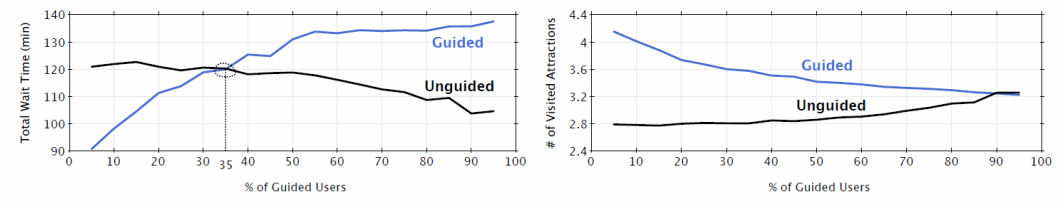

Figure 4.6: Total wait time (left) and number of visited attractions (right) per visitor type (on average).

Results of the virtual experiment is as summarized in figure 4.6. From figure 4.6, we can observe that at lower percentages of guided visitors (e.g., up to 35%), we do observe a better performance (lower wait times and higher number of attractions visited) for the guided visitors as opposed to the unguided ones. However, at higher percentage of guided users, we start to observe a decrease in performance of the guided visitors, especially from 35% onwards for the total wait time.

This observation of deteriorating performance (of guided agents) with increasing percentage of guided agents is in line with existing dynamic route guidance research in the transportation literature. A plausible explanation for this, which is also a limitation of the current dynamic route guidance model, is that the wait times at attractions are not a function of the routes generated. At lower guidance percentages (percent of guided visitors), this will not be an issue, as attraction wait times will not be significantly affected by the guided visitors movement. However, at higher guidance percentages, the attraction wait times (throughout the course of day) would most probably be significantly affected by the routes of the (larger) group of guided visitors, and not factoring this in the planning of visitors routes will most definitely result in a deterioration of guided visitors performance. This could potentially be an interesting direction for future works to consider,

by factoring generated visitor routes into the attraction wait times throughout the day.

4.5.2 Case Study 2: Impact of Spatial Layout

The second case study was inspired by a recent work on effect of spatial layout in theme park. In the work by ([82]), the theme park attributes that the authors looked into are distance between attractions, path network, entrance location, and attraction distribution ([82]). For the purpose of our investigation, we do not consider attributes such as distance between attractions. Instead, we focus on one of the key findings, where the authors mentioned was that the "beginning and midpoints of the main visit route (either direction) in a theme park with a circular path" would be more "likely to face greater capacity pressure than attractions in other locations" ([82]).

The reason why we chose to focus on this particular finding was due to the fact that the theme park that we were working with happens to have a circular path, and therefore it would be easy (as well as interesting) for us to test out such a finding via a simulation-based approach. Figure 4.7 shows the spatial layout of the theme park, where the attractions are indicated in yellow (popular attractions) and blue (normal attractions) stars. Intersection nodes are indicated by the green spheres, while the entrance is represented by the orange cone. From figure 4.7, we can observe that the popular attractions are spatially located near the entrance, with 3 out of 4 popular attractions (connected via intersection node 23) located near the entrance.

To evaluate the effect of proximity to entrance on crowding at theme parks, we introduce three different scenarios for comparison purposes. The first scenario is the baseline scenario, which is just as indicated in figure 4.7, where attractions are located in the same manner as that in the actual

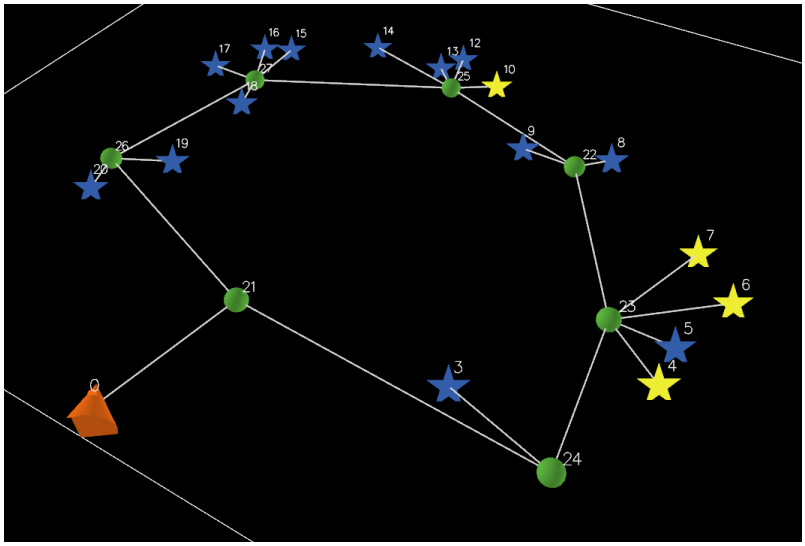

Figure 4.7: Spatial layout of attractions in theme park. Popular attractions are indicated in yellow.

theme park. The second and third scenario are added to evaluate the findings from [82], where scenario 2 represents popular attractions located further from entrance, and scenario 3 represents popular attractions located nearer to the entrance. Scenarios 2 and 3 are as illustrated in figure 4.8.

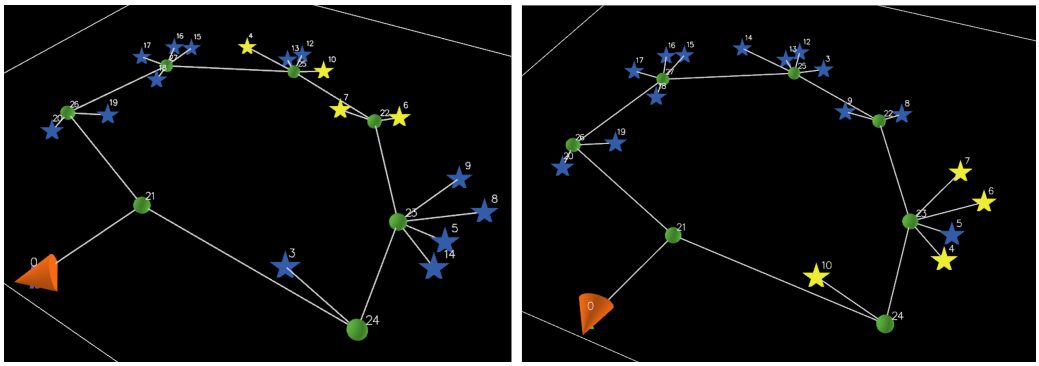

Figure 4.8: Spatial layout of attractions for scenario 2 (left) and scenario 3 (right). Popular attractions are indicated in yellow.

4.5.2.1 Performance Comparison and Results

To compare the results of various scenarios of spatial layout of attractions, we measure the performance of each scenario via an entropy measure for distribution of crowds. Formally, the entropy for distribution of crowds at attractions is defined as:

$$H(N) = -\sum_{n \in N} p(n) \log_2 p(n)$$

where $p(n)$ represents the normalized crowd distribution at an attraction n. As we are measuring the entropy of distribution of crowds in the theme park, a *higher entropy value* would equate to a relatively even distribution of crowd, which indicates better performance.

The results of the virtual experiments are as summarized in table 4.3, which represents the entropy of average crowd distribution at attractions through the entire day. From table 4.3, we can observe that having popular attractions further from entrance would yield better results (than the baseline), while having popular attractions nearer to the entrance would yield poorer results, as observed by the entropy values. This concurs with the findings as mentioned by ([82]).

Table 4.3: Entropy results for scenario (1) baseline scenario, (2) popular attractions further from entrance, and (3) popular attractions near entrance. Best performance is indicated in bold.

	scenario-1	scenario-2	scenario-3
entropy	3.209763	**3.255398**	3.131249

An additional result is as illustrated in figure 4.9, which shows the entropy values by hour of day for a single replication of simulation. From the figure, we can observe that other than a single hour of the day, the second scenario generally outperforms that of the first scenario. This further confirms the finding by ([82]). Another interesting observation is that scenario 3 outperforms the baseline towards the end of day (last 3 hours). A possible explanation could be that the initial crowding (i.e., popular attractions near entrance) has cleared by that time of day, and visitors are spread more evenly to the remaining attractions (current model does not allow revisiting of attractions).

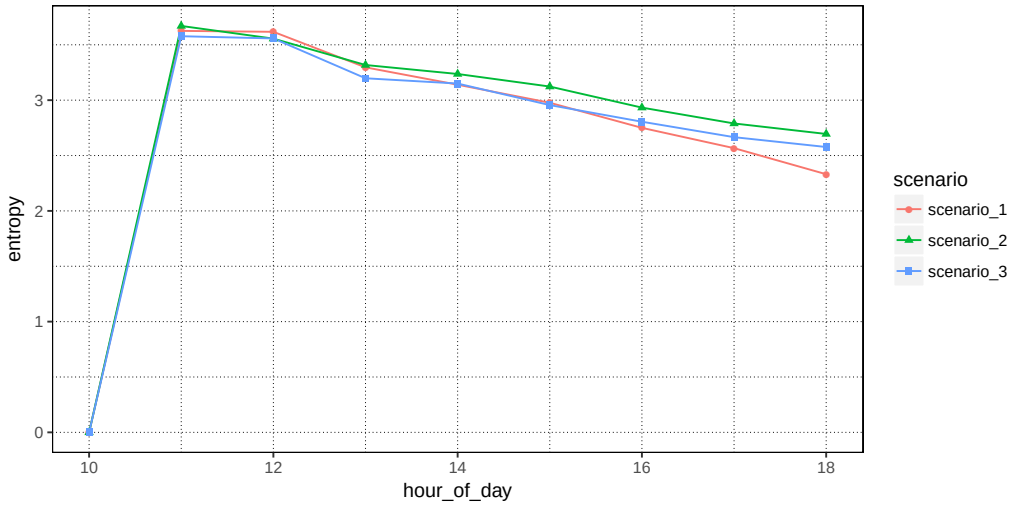

Figure 4.9: Entropy for different scenarios by hour.

It is important to point out here that this is but just one of the findings that is mentioned by ([82]) - and is used as a simple case study on how an agent-based simulation model can be used as a evaluation tool for spatial layout strategies. Future research could involve other findings (or a combination thereof) to further investigate the effect of spatial layouts on congestion at attractions.

4.6 Discussion

In this section, we discuss the practical implications of the result obtained from our work, along with some future research directions.

4.6.1 Practical Implications

The usage of sample-based mobility studies for crowd simulation related research has long been established, especially in the transportation domain. This work aims to answer the question on the representativeness of such sample-based mobility studies, particularly in the theme park domain. While it is practically impossible to collect large scale mobility traces of theme park patrons and their attraction attendance decision, we overcome

this by utilizing an agent-based simulation model as a ground truth generator, and compared the sample-based approach with a baseline method (diffusion dynamics). The findings support the usage of sample-based mobility studies, which confirms its use in various crowd simulation domains as a means of data collection. Additionally, it also demonstrates the potential of agent-based models as a ground generator in the event of low data availability.

Additional takeaway for researchers conducting sample-based mobility studies is as discussed in the sensitivity analysis results, where an increase in sampling rate up till 10% improves the rate of increase in performance, following which increasing the sampling rate further leads to smaller increases in performance. This helps researchers to make informed decisions in determining the sampling ratio prior to conduct of their mobility studies.

4.6.2 Future Research

This work is a combination of our exploration on the utilization of agent-based models in a particular scenario of crowd simulation (theme park visitors), where we have a group of heterogeneous agents with ad hoc objectives, differing preferences, and the flow between attraction are uneven. We hope that our work will inspire future works in this domain, with several examples as listed below:

- Use of agent-based models in evaluation of impact of spatial layout on congestion in theme parks ([82]). ([82]) concluded on the effects that path network and entrance location on waiting times at attractions. Therefore, it would be an interesting direction to look into spatial layout of popular attractions to reduce congestion, such as avoiding the start and midpoint of a popular visit trajectory for a theme park with circular path (one of the findings from [82]).

- Use of agent-based models to evaluate queue management strategies.

For example, theme park operator could set various queue time threshold, and redirect visitors to other less crowded attractions when the popular attractions are experiencing long wait times.

4.7 Conclusion

This study attempts to establish the usefulness of sample-based mobility studies in a special class of crowd simulation, where agents are heterogeneous with differing preferences and ad hoc goals. Also, our work attempts to establish the use of agent-based simulation models as a ground truth generator, in the event of low data availability. For the domain of crowd simulation (in particular simulation of theme park visitors), it is often impractical (logistically and financially) to carry out comprehensive studies on visitor decisions on an entire population level, and researchers often have to resort to sample-based mobility studies, where a sampled population are recruited as participants for a mobility survey, leading to the question of whether the decisions of the sampled population is representative of that of the entire population. The issue is further exacerbated by the fact that ground truth data (i.e., micro-level decisions made by entire population) is not available, rendering any form of ground truth-based validation and comparison impossible.

We answered this by first conducting sample-based mobility surveys to produce a baseline visitor movement model. Next, we made use of an agent-based simulation model as a ground truth generator, so as to evaluate the effectiveness of a sampled-based approach (similar to that *on-the-ground*, vis-a-vis that of sample-based mobility studies). This was done by comparing with a diffusion dynamics model, which is shown to provide a prediction accuracy of approximately 80% for the popular attractions ([27]). Our approach outperforms the diffusion dynamics model, even

when the sampling rate is as low as 1%. This establishes the usefulness of sample-based mobility studies in scenarios such as that in theme parks, where the flow between (attraction) nodes are shown to be uneven. It also demonstrates (1) the usefulness of mobility studies in behavioral modeling in the leisure domain, as well as (2) the capabilities and usefulness of agent-based simulation models as a ground truth generator in the event of low data availability.

We further looked into the sampling rate as a means of controlling the performance, by varying it through the conduct of a sensitivity analysis test. Our results show that the biggest gains are made when increasing the sampling rate from 1% to 10%, following which the rate of increase in performance starts to deteriorate (from 10% to 90% sampling rate).

Last but not least, we included two additional case studies to further illustrate the effectiveness of agent-based models in investigation of a visitor experience management application and effect of spatial layout on crowding. Our work paves the way for future investigation of (1) building highly accurate crowd simulation models, (2) use of agent-based models as a ground truth generator in the event of low data availability, and (3) use of agent-based simulation models in investigation of theme park crowd management strategies and policies.

Chapter 5

Micro-level, Fully Observable: Transportation Domain

Overview

In this chapter, we look in development of high quality agent-based simulation models in the micro-level, fully observable domain. A representative use case here would be the transportation domain, in particular taxis. Taxis play a fundamental role in many cities around the world, providing a faster alternative to public transportation modes, such as buses and trains. However, taxis have been shown to suffer from efficiency problems, i.e., supply-demand imbalances at various times of the day. This can be mainly attributed to a lack of a central coordination guidance mechanism, where drivers rely only on local knowledge to get to places at various times of the day. While various approaches have been proposed to deal with this issue, such as through guidance systems, such approaches are often expensive to implement and test, especially on an island wide scale.

An alternative would be a simulation-based approach, where it is imperative to develop highly realistic simulation models that are capable of emulating the actual movement decisions of taxi drivers on the ground. This will

serve as economical approach to effectively test the performance of such guidance systems. To this end, we propose a two-tiered decision model for modeling the roaming movement decisions of taxi drivers. The proposed model consists of link and zonal level decisions, where agents would switch between the two levels based on roaming dwell time. Additionally, through the use of case studies, we demonstrate how we can utilize such highly realistic agent-based simulations to investigate the effectiveness of a driver guidance system, that is currently under the field trial phase. The results conclude that by adhering to recommendations from the driver guidance system, drivers will experience an increase in performance, across various times of day and adoption ratios. The proposed approach provides a promising way to produce highly accurate multi-agent simulations in the transportation domain.

5.1 Introduction

Taxis are an important mode of transportation in many urban cities, where it serves as a faster and more reliable alternative to traditional public transportation modes (e.g., buses and trains). It also accounts for a significant number of daily transportation services in cities such as Singapore, where it represents 13.3% of total public transportation trips [49]. However, despite the benefits that it brings, taxi services are traditionally known to suffer from efficiency problems, as observed by the *supply-demand imbalances* during the peak and non-peak periods. This can be mainly attributed to the lack of a global coordination mechanism. As a result, the movement of taxi drivers are purely based on local information or past experience.

Recently, approaches have been proposed to coordinate the movement of taxi drivers in an Uber-like manner [38], where they leverage on the availability of real-time streaming data to perform demand prediction and

driver guidance with consideration of future supply and demand movements. While such approaches could potentially increase the efficiency of drivers and solve the issue of *supply-demand imbalances*, they are often logistically and economically costly to implement and test on a city-wide scale - involving field trials with thousands or potentially tens of thousands driver participants.

To enable the rapid evaluation of such approaches, one way would be to make use of simulation models as an economical approach to first test out the effectiveness of various approaches and recommendation models prior to actual field trials. In this case, it is imperative to develop highly realistic simulation models that gives a close representation of the taxi drivers movements, similar to that observed *on-the-ground*.

To this end, the focus of this work is on the modeling of movement behavior of taxi drivers. Due to the job nature of taxi drivers, it is subjected to two types of movements: voluntary (when searching for customers), and involuntary (when driving passengers to their destination). Taxis, in the hired state (i.e., after a street pickup or booking), are subject to involuntary movements, where their movements are constrained mainly by the passengers' objectives (i.e., getting to the passenger's choice of destination), with little or no variation in the route traveled, often by the path of shortest distance or the least travel cost incurred.

On the other hand, when in the roaming state (i.e., "free" status), taxi drivers make their own travel decisions, which involves traveling to places with potentially greater revenue or passenger demand. For this reason, the *voluntary movement* is of greater interest to us. In particular, we propose the development of a *two-tiered decision model* to emulate the actual movement decisions made by taxi drivers in the roaming state. The two-tiered decision model proposed consists of zone and link level decisions, where drivers would switch between the movement levels based on zonal dwell time.

Our approach is as follows:

- We perform statistical analysis on a dataset that is obtained for taxi fleets across Singapore. This allows us to test how factors such as distance and historical demand can have an effect on the taxi drivers zonal movement model.

- We utilize TaxiSim 2.0 (works on previous [17]) as a simulation platform to evaluate the performance of a *two-tiered decision model* against that of a traditional *single-tiered decision model*. TaxiSim 2.0 is a multi-agent simulation platform that combines (1) data-driven modeling, (2) tiered decision modeling and (3) agent-based modeling techniques to emulate actual *on-the-ground* movement of taxi fleets.

- We compare aggregated output (e.g., trip accumulation and zone visitation) of the two-tiered decision model against that of a traditional single-tiered model to demonstrate the performance of the two-tiered decision model. Results show that the two-tiered decision model outperforms the single-tiered decision model, supporting the usage of a two-tiered decision model in modeling the movement decision of taxi drivers.

- We further introduce two case studies to demonstrate the effectiveness of TaxiSim 2.0 as an evaluation tool for a Driver Guidance System (DGS), which was developed to improve the operations of taxi drivers.

We believe that the *two-tiered decision model* would make for a better decision model that captures the movement of taxi drivers - resulting in highly realistic simulation models.

5.2 Related Work

Our work is motivated by existing literature in the domains of data-driven agent-based modeling, tiered decision modeling and applications of multi-agent simulation. These concepts and models may or may not be applied in the application area of transportation, in particular taxis. However, to the best of our knowledge, there is no existing work that combines all three elements in the development of a data-driven, multi-agent simulation model with a two-tiered decision model. On top of that, we showed the usefulness and applicability of the model developed by introducing case studies involving the utilization of the model as an evaluation tool for a driver guidance system that is currently under the field trial phase. This is the main contribution and novelty of our work.

5.2.1 Data-Driven Agent-Based Models

In the area of data-driven agent-based models, [81] adopted a data-driven approach in the development of an agent-based model for rooftop solar adoption, where the authors make use of actual data from the Center for Sustainable Energy to define an adoption model based on peer effects and Net Present Value (NPV). [44] developed a data-driven agent-based model for the purpose of crowd simulation, where the movement of agents are dependent on their environment and other nearby agents, which is learned from actual movement dynamics (i.e., trajectories) derived via video processing.

Both work by [81] and [44] differs from traditional agent-based simulation models, where the action or movement of agents is often based on simple rule-based mechanism, without consideration of actual ground truth data.

In our work, we adopt a data-driven approach in the development of TaxiSim 2.0, where we make use of existing dataset available, along

with some inferred data (discussed in section 5.9.1) for both the decision and simulation model. This includes multiple aspects, which ranges from initialization of supply and demand, to link and zonal level movements, to the decision of making a zonal jump.

5.2.2 Tiered Decision Modelling

The tiered decision model is not a new concept. In [62], a 3-tiered decision model was proposed for crowd simulation, with the 3 tiers being (from lowest to highest) locomotion, steering and social. Unlike the hierarchical decision model proposed by [62], focus on the locomotion aspect is not within the scope of this thesis.

In the work by [79], the authors introduced a two-stage approach for modeling the movements of vacant taxis, which we define as taxi drivers behavior in *roaming mode*. The first stage model proposed by the authors involves sequential, independent zone-based decisions, modeled with a multinomial logit model (MNL) whereby a taxi decides whether to (1) stay in a current zone or (2) move to an adjacent zone in search of passengers. This is similar to our strategic (zone) level movement, except that we allow for zonal transitions to non-adjacent zones as well. This is a major differentiating factor, as only 46.1% of zonal movements are to adjacent zones (based on our pre-processing by zone-specific dwell time), which is why we include non-adjacent zones into the zonal transition model. Another distinction would be that the second stage is modelled cell-based, whereas our second (operational) level movement is modeled as a link-based movement. Additionally, the two-stage model proposed by [79] works with data from a fleet of 460 taxis, while *TaxiSim 2.0* (our model) works with data from more than 26,000 taxis.

Another work by [78] introduces a bi-level decision for taxi drivers in roaming mode as well. In this work, the levels of decision are (1) whether

to travel to a nearest taxi stand after dropping off a passenger and (2) whether to join the queue at the nearest taxi stand once they have reached there. For the purpose of this work, we do not consider movements to and fro taxi stands, but rather focus on link and zone-based movements.

5.2.3 Multi-Agent Simulation in Transportation

Multi-agent simulations have been shown to be useful in the domain of transportation. [5] covers the usage of multi-agent simulations done by various works in the area of traffic management and traffic control. It is also used in general vehicular simulations such as [60], which is used for modeling the dynamics of vehicles and their surrounding vehicles.

Agent-based modeling and simulation (ABMS) techniques are also used extensively in modeling the movement of taxis [36] [35] [10] [50]. A major benefit of this is that agent-based models (ABM) serves as an economical approach for evaluating intervention or coordination approaches, without having to implement them on the ground. In our case, we utilize the simulation model for evaluation of a Dynamic Guidance System (DGS). This is done in preparation of a upcoming field trial, where we use the simulation model as a test-bed of the effectiveness of such as system in improving the supply/demand imbalances at zones at various times of the day.

An existing vehicular simulation model would be SUMO (Simulation of Urban MObility) by [6] [41]. While SUMO does include additional features such as traffic lights and even pedestrians [6], this do not fall within our scope of investigation of the effectiveness of a two-tiered decision model. Additionally, as mentioned by the authors, SUMO requires a explicit definition of a vehicle's route, i.e., the "complete list of connected edges between a vehicle's start and destination. This means that the movement is essentially of a deterministic, one-tiered (link-level) nature,

which differs from ours, which is a stochastic, two-tiered level decision model, allowing us to test the effectiveness of a two-tiered versus single tier decision model.

Another example of vehicular simulation model would be [52] [51]. In this work by [52] [51], the focus is more on modeling of movement of drivers to-and-fro taxi ranks (also known as taxi stand or designated spots where taxis wait for passengers). This differs from our work, where we focus on movement along links and zones, guided via the two-tiered decision model.

5.3 TaxiSim 2.0: A Data-Driven Multi-Agent Simulation Platform

In this section, we first discuss various features in TaxiSim 2.0 that are relevant to the virtual experiments that we are conducting. TaxiSim 2.0 (builds on existing [17]) is utiliyed as an evaluation testbed on the effectiveness of a two-tiered decision model for representing the movement decision of taxi drivers (sections 5.4 and 5.5). For additional details on the TaxiSim and its usage, refer to [17].

TaxiSim 2.0 is developed in Java 1.7, along with customized Java extensions. The road network data is based on OpenStreetMaps [61], with the movement of taxis specified by the *two-tiered decision model* (discussed further in section 5.4). Figure 5.1 shows the interface of TaxiSim 2.0 during the course of simulation. Several aspects of the virtual experimental setup and TaxiSim 2.0 components are as discussed in subsequent subsections.

Models in transportation can generally be classified as microscopic, macroscopic and mesoscopic models. Microscopic models look into agent or individual-level details, such as dynamics of queues and vehicles overtaking. Macroscopic models, on the other hand, looks into traffic and vehicular movements at an aggregated level, example of which would be equilibrium

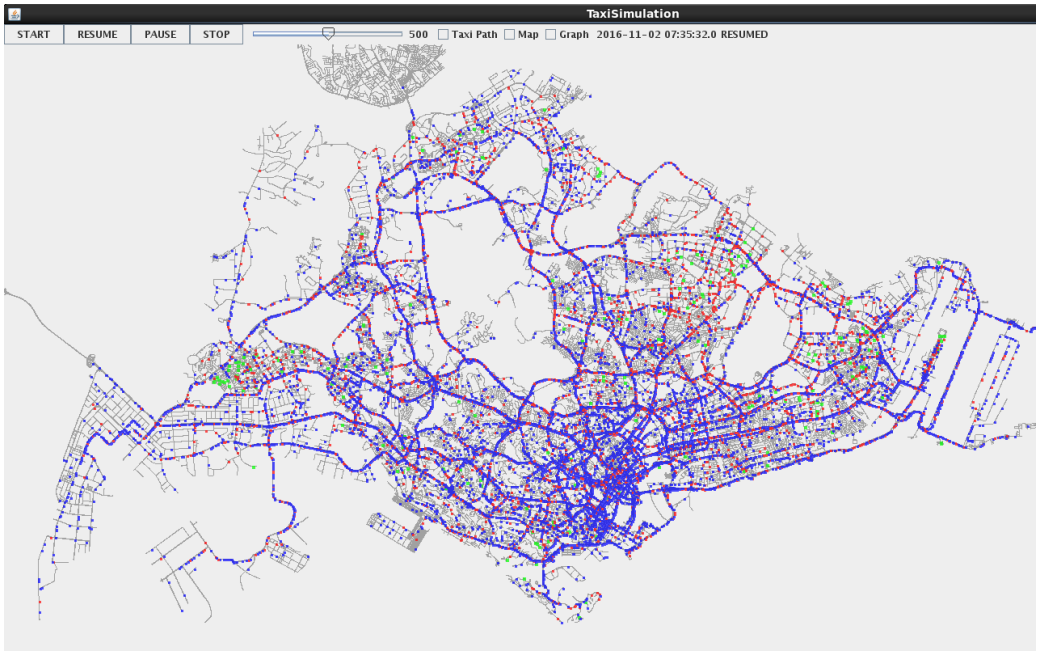

Figure 5.1: Interface of TaxiSim 2.0.

network flows [18]. TaxiSim 2.0 is designed as a mesoscopic simulation model, in that it differs from other microscopic and macroscopic models. For example, TaxiSim 2.0 includes queues (e.g., at taxi stands), which are characteristics of a microscopic model, but the queues are not at traffic junctions or for purposes of traffic flow (e.g., vehicles at traffic lanes and overtaking). This substitutes some level of granularity for performance.

5.3.1 Simulation Agents: Taxis

In this subsection, we discuss various attributes and behaviors pertaining to taxi agents in the TaxiSim 2.0 simulation model, ranging from initialization to operation mode to roaming behavior.

5.3.1.1 Supply Generation

For initialization of taxis (supply generation), we could either (1) initialize all taxis at the start of simulation or (2) initialize the taxis in batches (i.e., various times of the day), reflecting the work shifts of drivers. In our case, we do not have explicit information on shifts of drivers, as only taxi ID

(and not driver ID) is available in the data provided by the operators. A proxy for this would be a change in status to "Change Shift", although this might not be adopted by all drivers, causing the data to be incomplete.

Considering the set of data available, we decided to take up the first option, where we adopted a data-driven approach for initializing all taxis at the start of simulation. The initialization of taxis are based on actual distribution across different zones throughout the island at the start of day. While this is slightly different from actual *on-the-ground* operations of drivers, we found that the number of taxis that were initialized (24,000) matches the total number of "active taxis" (i.e., non-stationary taxis) throughout the day.

This is illustrated in figure 5.2, where we generated the number of "active taxis" for the month of November 2016 as follows. For each hour, we retrieved the total number of unique taxi IDs with at least one data point with an "active status" (i.e., not offline). From the results generated in figure 5.2, we can observe that the number of active taxis (i.e., active supply) per hour is approximately around 24,000, which matches the number of simulated taxis in our virtual experiments.

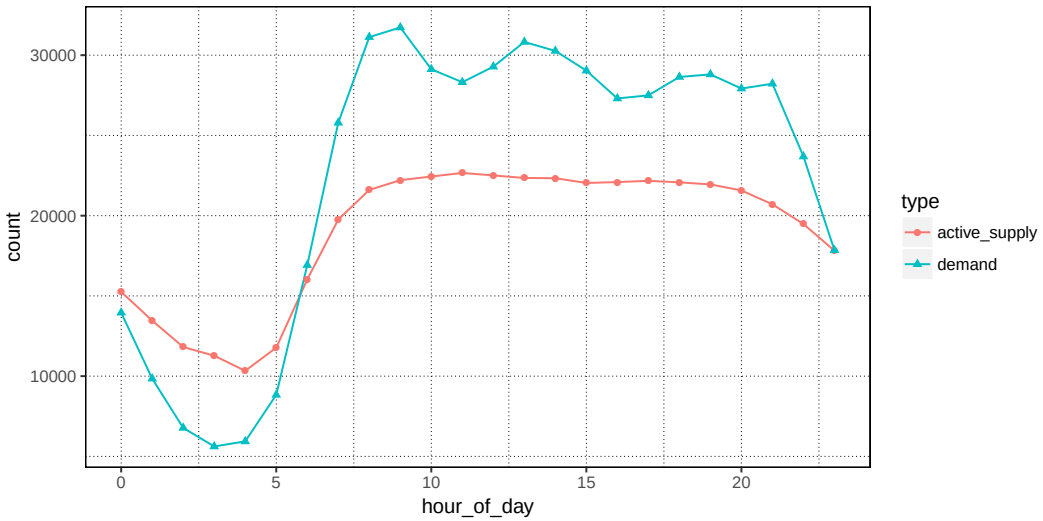

Figure 5.2: Number of active supply and demand by hour of day (Month: November 2016).

5.3.1.2 Mode of Operation

We currently assume two modes of operation ("hired" and "free") for the taxis in TaxiSim 2.0. The behavior of drivers under these two modes are as discussed subsequently. The overview of the movement model is as shown in figure 5.3. In the hired mode, the taxi will make its way from the origin to destination of a trip based on shortest path [30]. In actual situations, passengers would usually make route choices based on shortest path or lowest fare. However, given that we do not have fare information associated with the logs, taking shortest path from origin to destination would be a reasonable assumption. The movement of taxis in the roaming state ("free" status) is then defined according to the *two-tiered decision model* that is defined earlier. In the case of the *single-tiered decision model*, taxis will move according to only link-level decision model (discussed further in section 5.4).

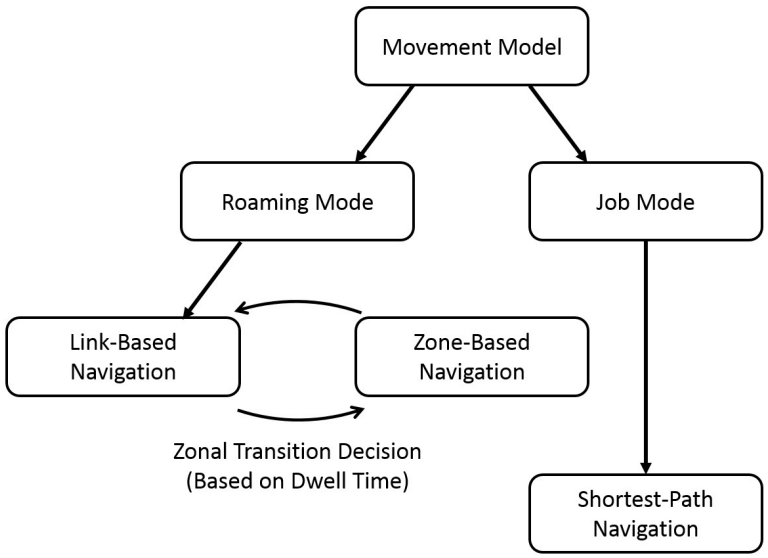

Figure 5.3: Sequence of Decision in Roaming Mode.

5.3.1.3 Roaming Behavior

TaxiSim 2.0 currently allows for three modes of roaming behaviors: (1) 1-tier or link-level (2) 2-tier or zone+link level and (3) replay of historical

movements. The single and two-tiered roaming behavior will be further elaborated in section 5.4. For the purpose of this work, we focus on the use of roaming behaviors (1) and (2) only.

5.3.2 Simulation Agents: Passengers

In this subsection, we discuss various attributes and behaviors pertaining to passenger agents in the TaxiSim 2.0 simulation model.

5.3.2.1 Demand Generation

For demand generation, we adopt a data-driven approach, where we utilize a Poisson model for initialization of passengers. First, we divide the trips data by their origin zone z (i.e., zone that pickup/booking occured). We then derive for each zone z and discrete time interval t an average number of pickups that occured, defined as $\lambda_{z,t}$. The $\lambda_{z,t}$ obtained is then input to the poisson demand generation model. Formally, the probability of k demand (i.e., passenger agents) being generated in a zone z at a discrete time interval t is defined as:

$$P_{k,z}^{t} = e^{-\lambda_{z,t}} \frac{(\lambda_{z,t})^k}{k!} \qquad ,\forall z \in Z$$

A caveat here will be that using the trips information only gives us the *true positives* (passengers that "appeared" at zone z and boarded a taxi), but not the *false positives* (passengers that "appeared" at zone z, but did not manage to board a taxi), as this information is not available in our data.

For destination of every demand or passenger, we also adopt a data-driven approach, where the destination is generated in a stochastic mannner following that of actual trip Origin-Destination distributions. Note that

initialization of a passenger may or may not equate to an actual trip (i.e., fulfilled trip), depending on whether there's any available taxis nearby.

5.3.2.2 Demand Type

Taxi jobs in the simulator are divided into two types: booking and street pickup. For street pickups, taxis will pickup passengers that are on the same road link, similar to that *on-the-ground*. For booking jobs, a passenger will make a booking, and the job information will be disseminated to taxis in "free" mode that are within a search radius (discussed subsequently) from the passenger's location. We adopt a data-driven approach here as well, where the poisson demand generator (discussed previously) is distinct for both types of demand (i.e., each demand type has their own $\lambda_{z,t}$ value).

5.3.2.3 Waiting Behavior

TaxiSim 2.0 incorporates a waiting behavior into the demand, which serves as a job expiry period. Therefore, after passengers are initialized, they will wait at the point of initialization for a certain amount of time, after which they will be removed from the simulation. This will constitute towards a *miss scenario*, or an *unfulfilled service request*. This behavior is similar to that on-the-ground, where passengers do not wait indefinitely for a taxi to arrive. As mentioned previously, as we do not have information on unfulfilled demand, we assign a value of 15 minutes for the waiting time threshold, which is a reasonable assumption.

5.3.3 Matching of Supply to Demand

In this subsection, we discuss the matching mechanism of supply to demand within the TaxiSim 2.0 simulation model. This will be discussed subsequently.

5.3.3.1 Demand Type: Booking

For booking jobs, a passenger will make a booking, and the job information will be disseminated to taxis in "free" mode that are within a search radius from the passengers location. To improve the performance of the simulation model, we currently employ a *grid-based* distance measure (where each grid is 500 by 500 meters), instead of a *radius-based* distance measure. The booking process is as illustrated in figure 5.4. For every booking that is initialized, the passenger will reach out to cabs in search zones 1, 2, and followed by 3, depending on (1) whether there's taxis in "free" mode in the search zone and (2) whether the booking job was accepted (*probabilistic* acceptance by drivers).

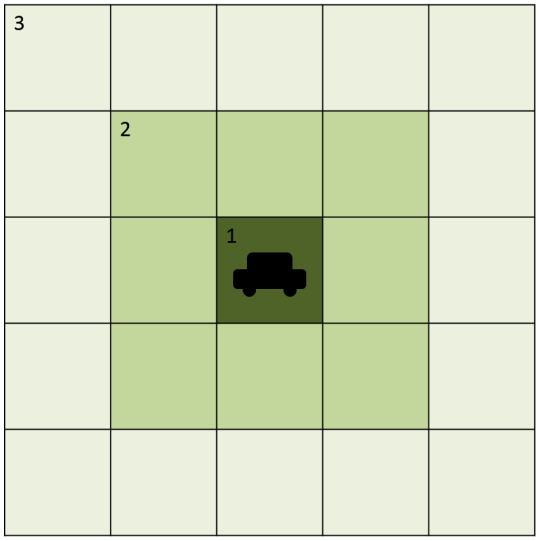

Figure 5.4: Illustration of Radius Concept for Booking. Each grid is 500 meters by 500 meters.

5.3.3.2 Demand Type: Street Pickup

For street pickup demand, it follows a similar *grid-based* matching mechanism, but with only a single grid (unlike the three search zones for bookings illustrated in figure 5.4), and a smaller granularity of 200 meters by 200 meters. This is to simulate a similar mechanism *on-the-ground*, where drivers can only observe potential passengers that are within their line of

vision (e.g., exists on the same link or adjacent link). Similar to the booking demand, the taxi has to be in "free" status before a passenger pickup.

5.3.3.3 Probabilistic Acceptance

On top of the matching via the search radius (for bookings) and link-level matching (for street pickups), TaxiSim 2.0 also incorporates probabilistic acceptance of (1) taxis accepting booking calls and (2) passengers cancelling booking jobs after driver accepts the job. As actual numbers are not available in our existing dataset, we assign a value of 5% for the supply (taxis) and demand (passengers) cancellation probability/rate. This percentage is used in both the virtual experiment and case studies, with results presented in section 5.5 and 5.6.

5.4 Methodology: Designing the Two-Tiered Decision Model

The two-tiered decision model proposed for taxi drivers is as illustrated in Figure 5.5. A driver will begin roaming by traveling at link-to-link (operational) level. With increasing time without a success passenger pickup, the driver will then make a decision on whether to make a "zonal jump" (i.e., zonal transition that could be to non-adjacent zones).

Subsequently, the driver will then need to make a strategic (zone) level decision, where it decides on the zone to travel to based on factors such as distance and historical demand (at destination zone vis-a-vis the driver's current zone). After the driver has moved to the destination zone, it will then revert back to the operational level decision. Further details of the decision model will be discussed in the subsequent sub-sections.

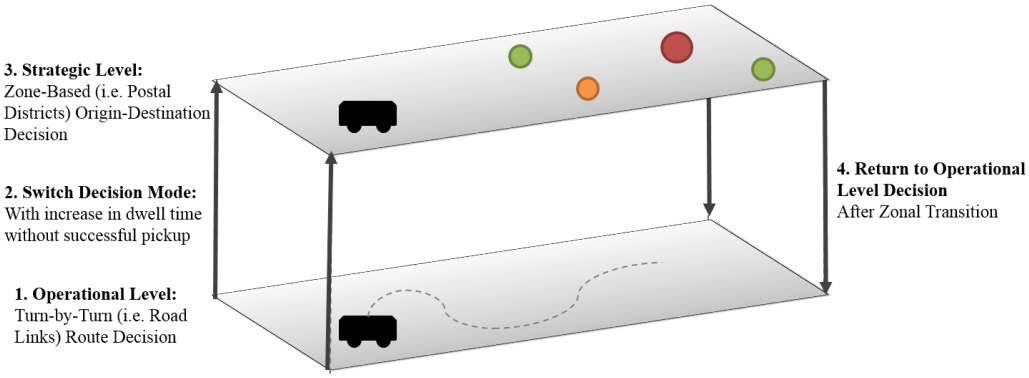

Figure 5.5: Two-Tiered Decision Model for Taxi Drivers

5.4.1 Strategic: Zone Level Movement

The first level that we'll look at is the strategic level (zonal) movement. That is, an individual taxi driver will make a decision to go from an origin to a destination zone based on several factors, such as distance and historical demand at the destination zone. Figure 5.6 supports the motivation for the construction of this level of strategy. In figure 5.6, we illustrated the (zonal) distribution of demand for taxis at two time periods on a typical weekday in September 2015. The demand is normalized against total demand for that hour, and highlighted on a *red-yellow-green* scale, indicating *high-medium-low* demand. From figure 5.6, we can observe that the demand distribution of taxis at different zones varies at times of the day (7am and 7pm). Therefore, it is intuitive that taxi drivers will choose to go to different zones at different times of the day, depending on the demand. Note that demand here is defined as number of trips originating from a zone, which is a subset of *actual demand* at a zone, which also includes the *unfulfilled demand*, which cannot be observed in our data.

After a driver has decided to make a zonal transition (discussed in the next subsection), he or she would then need to make a decision regarding which zone to move to, with the set of zone choices being discrete and finite - this decision is best modeled using a discrete choice model. In this case, the driver makes a "zonal jump", with the zonal selection decision

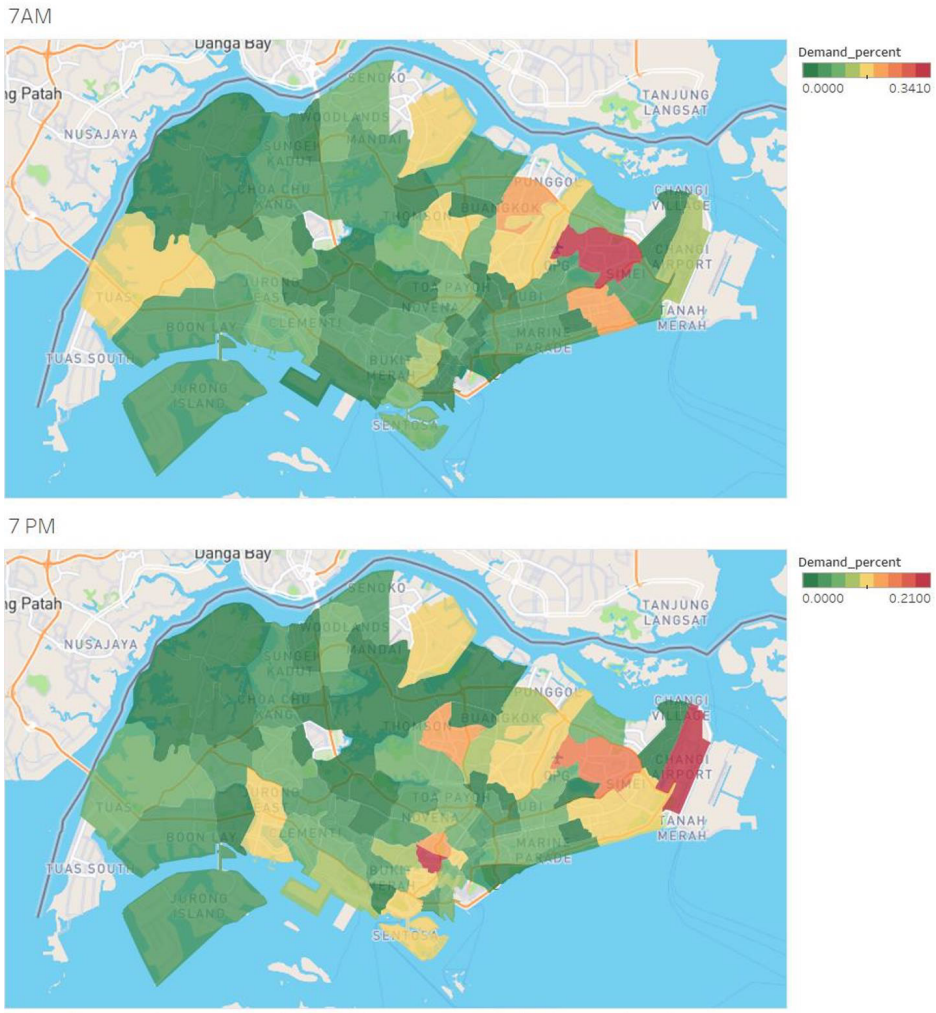

Figure 5.6: Distribution of demand (trips starting) at times of day: (top) 7am and (bottom) 7pm. Demand is normalized against total demand for the hour.

being dependent on the utility that driver gets by going from a zone u to a zone v is a combination of the distance (defined as $d_{u,v}$) and the demand at destination zone vis-a-vis the origin zone.

There are a number of potential functions that we can use in converting utility values to the probability distribution over choices. In our study, we adopt the logit function in modeling discrete choices made by taxi drivers in selection of zones. The logit choice probability model is one of the most widely used discrete choice model, and chosen due to the fact that the formula for the logit choice probability model takes a closed form and is readily interpretable [71]. As mentioned earlier, the second component of

the utility definition is a comparison of the demand at destination and origin zone.

We first define the demand at zone z (part of zone set Z) at time interval t to be D_z^t. In order to better cater for differences in total demand at various times of the day (peak and non-peak), we incorporated a normalized demand, and define it as:

$$\delta_z^t = \frac{D_z^t}{\sum_{w \in Z} D_w^t}, \forall z \in Z, t \in T.$$

Thereafter, we define the normalized demand attraction level, $\alpha_{u,v}^t$. That is, how *attractive* zone v is to a driver who is at zone u at time t. Formally, this is defined as:

$$\alpha_{u,v}^t = \frac{(\delta_v^t - \delta_u^t)}{(\delta_v^t + \delta_u^t)}, \forall u, v \in Z, t \in T.$$

The utility obtained from travelling from zone u to v at time interval t is then defined as:

$$U_{u,v}^t = \beta_0 + \beta_1 d_{u,v} + \beta_2 \alpha_{u,v}^t, \forall u, v \in Z, t \in T.$$

Given the set of zones Z, the probability of migrating from zone u to v at time interval t is then defined as:

$$P_{u,v}^t = \frac{e^{U_{u,v}^t}}{\sum_{z \in Z} e^{U_{u,z}^t}}, \forall u, v \in Z, t \in T.$$

To provide support for the logit model that is proposed for the strategic-level decision, we conducted a simple logit regression for the zonal movement

model. For the regression, we selected the zonal transition records for a typical weekday. Results of the logistic regression is as shown in table 5.1. From table 5.1, we can see that the regression results provide support for our proposed model, as shown by the significance level and coefficients of the variables.

Table 5.1: Logistic regression result for zonal decision model, with the probability of moving from zone u to v at discrete time interval t as independent variable, and the distance between zone u and v and demand attraction as dependent variable.

	Estimate	Std. Error	z value
(Intercept)	-2.862 ***	0.004319	-662.6
distance	-0.16 ***	5.234E-07	-312.5
demandAttraction	0.5847 ***	0.004909	119.1
Observations	185,984		

*Significance Level: * p <0.05; ** p <0.01; *** p <0.001*

As mentioned previously, one of the components of the two-tiered decision model is the concept of a zonal jump, where taxi agents (with increasing dwell time at a zone) would choose to make a zonal transition. To the best of our knowledge, existing works which models taxi driver movements (e.g., [79]) only includes adjacent zones into the zone transition model.

While the "non-adjacent" zonal transition is one of the novelty of our approach, a valid question here would be whether this makes an impact to the decision model for taxi drivers, particularly to the zonal level decision. The support for this zone movement is as illustrated in figure 5.7, which is the CDF distribution of zone transition distance (obtained via pre-processing described in section 5.9.1). From figure 5.7, we can observe that there's a significant amount of zone transitions that spans across a long distance (from origin to destination zone). To further put this into perspective, only 46.1% of all zone transitions were to adjacent zones. This supports the importance of incorporating a zonal transition that spans

across non-adjacent zones.

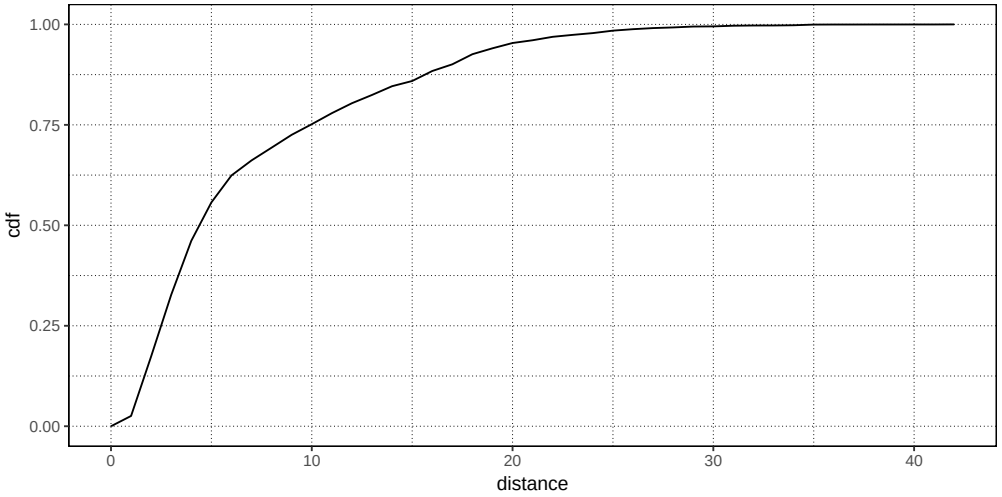

Figure 5.7: CDF Distribution of Zone Transition Distance (in km).

5.4.2 Operational-to-Strategic: Dwell Time Driven

In designing the zone-level movement, we would expect the decision to make a zone transition to be correlated with dwell time. That is, the longer a taxi spends searching for passenger in a zone, the more likely we would expect the taxi driver to make a "zonal jump" (i.e., zonal transition). Figure 5.8 shows the cumulative distribution function for dwell time before taxi drivers make a zonal transition.

From figure 5.8, we can observe that the zonal transition decision function takes on a logarithmic shape, whereby even at a dwell time of 20 minutes, we have approximately 80% of taxi drivers who have made a zonal transition. Therefore, we model the decision as a logarithmic function, which we define formally as:

$$P(transitZone) = \beta_0 + \beta_1 \log(dwellTime)$$

Table 5.2 shows the OLS regression result for the zonal transition decision. From the results, we can observe that the model represents the decision to

Table 5.2: Individual OLS regression result for the zonal transit decision model, with the probability of making a zonal jump transition as independent variable, and the log of dwell time spent at zone as dependent variable.

	Estimate	Std. Error	t value
(Intercept)	-0.21992 ***	0.03884	-5.663
log(dwellTime)	0.32016 ***	0.01169	27.393
Residual Std. Error	0.06554	Observations	185,984
Multiple R^2	0.9306	Adjusted R^2	0.9293
F-statistic	750.4	p-value	<2.2e-16

Significance Level: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

make a zonal jump very well, as observed from the significance values and R^2 values. An additional illustration of the fitted curve is as shown in figure 5.8, where the dotted line represents the cumulative distribution function (cdf), and the fitted curve is represented by the blue line. From figure 5.8, we can observe that the fitted log function (in blue) represents the actual zonal transition decision (in black) very well.

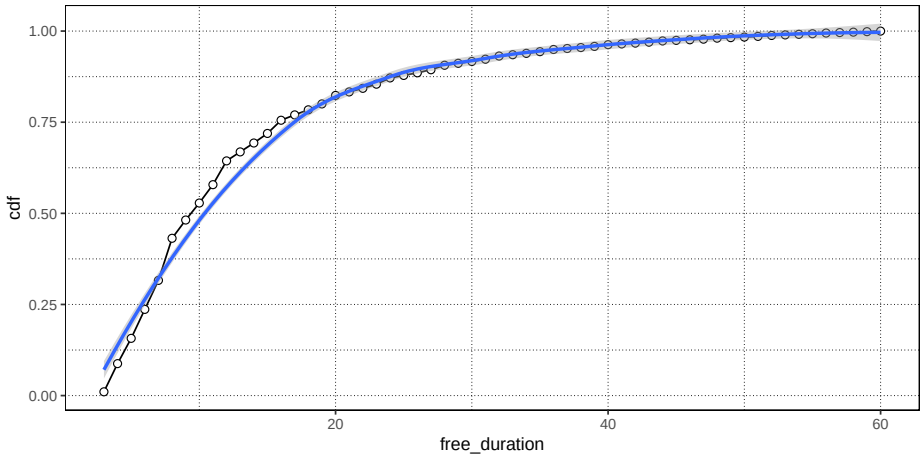

Figure 5.8: Cumulative Distribution Function: Dwell Time.

5.4.3 Operational: Link Level Movement

The operational decision is defined as link-to-link movement decisions. This link-to-link decision is represented by $P^t_{i,j}$, which is the probability

of moving from a link i to j at time interval t. As mentioned in the previous section, we derive link-level information by making use of the Map Matching algorithm [57] that is based on GraphHopper [34]. The algorithm essentially uses a Hidden Markov Model (HMM) representation for the road segments and their transitions across time. The Viterbi-based algorithm then computes the best path through the HMM latice.

From the link-level information derived (described in section 5.9.1), we then obtain the flow $F^t_{i,j}$, which is the number of drivers going from link i to link j at a discrete time interval t. With the flow information, link transition probability (denoted $P^t_{i,j}$) is derived according to the formula shown below. Since $P^t_{i,j}$ refers to transition probability, $\sum_{j \in L} P^t_{i,j} = 1$.

$$P^t_{i,j} = \frac{F^t_{i,j}}{\sum_{k \in L} F^t_{i,k}}, \forall i, j \in L, t \in T.$$

5.4.4 Analysis: Two-Tiered Decision Model

In the previous subsections, we give a layer-by-layer perspective of the two-tiered decision model that is proposed, from the zonal level movement, to the link-to-zonal switch, and finally to the link level movement. This is done via several regression analysis, where we demonstrated the soundness of individual levels of movement decision.

To provide support for the two-tiered decision model though, we would need to demonstrate the soundness of the two-tiered decision model in its entirety. This is in spite of the fact that we do not have explicit data regarding the zonal level decision of drivers (i.e., indication of when a driver makes a conscious zonal movement decision) - which makes it impossible to perform any form of statistical analysis on the ground truth data available.

To achieve this, we make use of TaxiSim 2.0 as an evaluation testbed for comparing the performance of the two-tiered decision model as opposed

to a traditional single-tiered decision model. The approach is as follows. We first keep track of of zonal transitions $P^t_{u,v}$ and time spent in each zone (i.e., dwell time) by performing the pre-processing as specified in 5.9.1. Thereafter, we make use of the microscopic movement model (single or two-tiered decision model) to generate the transition between zones (defined by $\widehat{P^t_{u,v}}$ and $\widetilde{P^t_{u,v}}$ for the single and two-tiered model respectively). Finally, we compute the average of errors for the single tier ($\widehat{ASE}$) and two-tiered ($\widetilde{ASE}$) decision model.

Formally, the average of errors for the single-tier ($\widehat{ASE}$) and two-tier ($\widetilde{ASE}$) decision model are defined as:

$$\widehat{ASE} = \frac{\sum_{u,v \in Z, t \in T} |P^t_{u,v} - \widehat{P^t_{u,v}}|}{|Z| \times |Z| \times |T|}$$

$$\widetilde{ASE} = \frac{\sum_{u,v \in Z, t \in T} |P^t_{u,v} - \widetilde{P^t_{u,v}}|}{|Z| \times |Z| \times |T|}$$

Additionally, we introduce a weighted average of errors, which penalizes the errors for high probability transitions. Formally, the weighted average of errors for the single-tier ($\widehat{WASE}$) and two-tier ($\widetilde{WASE}$) decision model are defined as:

$$\widehat{WASE} = \frac{\sum_{u,v \in Z, t \in T} |P^t_{u,v} - \widehat{P^t_{u,v}}| \times P^t_{u,v}}{|Z| \times |Z| \times |T|}$$

$$\widetilde{WASE} = \frac{\sum_{u,v \in Z, t \in T} |P^t_{u,v} - \widetilde{P^t_{u,v}}| \times P^t_{u,v}}{|Z| \times |Z| \times |T|}$$

The results of the average of errors and weighted average of errors are

as shown in table 5.3 and 5.4 respectively. From the table, we can observe that the two-tiered model gives a better representation of the inferred *on-the-ground* zonal transition, as represented by the lower average of errors, and especially so for the weighted average of errors, which means that the two-tier model is performing better for the zone transitions with higher probabilities. This gives an initial confirmation of our intuition on the use of a two-tiered model in modeling the behaviors of taxis in roaming mode. To further demonstrate the performance of the *two-tiered decision model* proposed, we introduce additional performance measures, which will be subsequently discussed in section 5.5. We then introduce three case studies in section 5.6, two on utilization of the model for evaluating a Driver Guidance System (DGS), and an additional case study on investigating the zonal jump that is proposed in the *two-tiered decision model*, which differs from some of the existing literature. This will be elaborated in the subsequent sections.

Table 5.3: Zone transition comparison using average of errors.

	Single-Tier Model	Two-Tier Model
Average of Errors	2.68%	2.58%
Standard Deviation	6.41%	5.59%

Table 5.4: Zone transition comparison using weighted average of errors.

	Single-Tier Model	Two-Tier Model
Weighted Average of Errors	28.37%	24.69%
Standard Deviation	145.73%	113.08%

5.5 Virtual Experiments: Results

As mentioned previously, in order to evaluate the performance of the *two-tiered decision model*, we compare to a *single-tiered decision model*. To do this, we run TaxiSim 2.0 under both decision models. In the *single-tier*

Input Variables	**Description / Values**
Number of Driver Agents	24,000
Detour Threshold	320 minutes
Dependent Variables	**Description / Values**
Cumulative Trip	R^t
Zone Visitation	V_z^t
Number of Replications	30

mode, movements of taxis in roaming mode will be guided by just the link-level decision model. In the *two-tier* mode, movements of taxis in roaming mode will be guided by link-level decisions first. When the zone-specific dwell time threshold is reached (without a successful passenger pickup), a "zonal jump" will be triggered, and the taxi will pick another zone based on historical zone transitions. The aggregate level outputs (under both modes) from TaxiSim 2.0 are then compared with actual ground truth data, which is discussed subsequently.

5.5.1 Virtual Experiment Setup

TaxiSim 2.0 is run under two different movement models, a *two-tiered decision model* and a *single-tiered decision model*. The single tiered model is where the agents move about via the link-level transition model. We setup the virtual experiments with 24,000 agents per replication, with a total of 30 replications per setup (single or two-tiered agents). Other aspects of the model are as discussed subsequently.

5.5.2 Evaluation Method 1: Cumulative Trip

The first method of comparison is the cumulative trip curve of drivers. From the predicted cumulative trip curve of drivers, we then compare against that of the actual cumulative trip curve of drivers throughout a day. To make the comparison, we make use of two metrics, namely the

area between curves, as well as the *maximum difference*. We first define the trip accumulation variable R^t, which is the accumulated trip for drivers at a time interval t. Note that the trip accumulation is normalized against total number of trips for the day. The ground truth trip accumulation variable is then compared against that for the one-tier decision model (defined as $\widetilde{R^t}$), and the two-tier decision model (defined as $\widehat{R^t}$).

The cumulative distribution function (CDF) curve for both one and two-tiered decision model is as shown in figure 5.9. The CDF of the ground truth, single-tier and two-tier model are represented by red, blue, and green lines respectively. From the figure, we can observe how closely the two-tiered decision model (green) to that of the *ground truth* (red), as opposed to that of the single-tiered decision model (blue). Additionally, the maximum error for the two-tiered model is 2.48%, as opposed to a maximum error of 6.22% for the single-tiered decision model. This shows that the two-tiered decision model performs better in representing the trip accumulation throughout the day.

5.5.3 Evaluation Method 2: Zone Visitation

The second output we are comparing is the zone visitation. Note that zone visitation here refers to zone visitation by taxis in "free" mode only. We define zone visitation by notation V_z^t, which is the number of vehicles entering zone z at time interval t, and compare the difference between one-tier decision model (defined as $\widetilde{V_z^t}$) and two-tier decision model (defined as $\widehat{V_z^t}$) against that of the ground truth data. In this case, we assess the performance of both decision models based on two criteria, the *average of errors*, and the *weighted average of errors*.

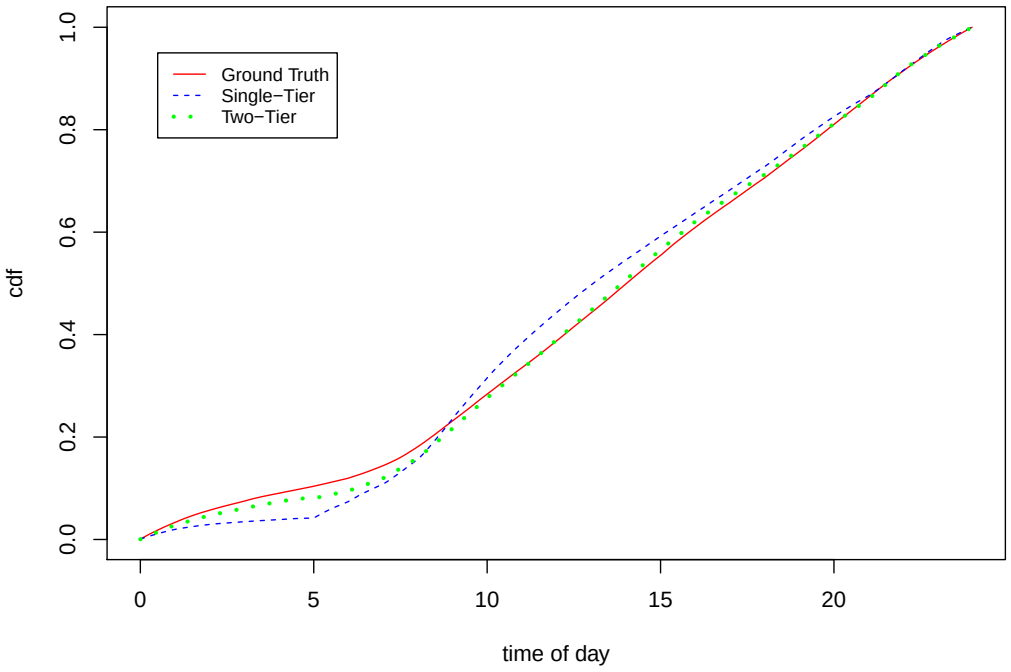

Figure 5.9: CDF: Trip accumulation for ground truth (red), single-tiered (blue) and two-tiered decision model (green).

5.5.3.1 Metric 1: Average of Errors

The average of errors measure the difference in zone visitation for the single and two-tiered decision model against that of the ground truth. We use the notation $\widetilde{AE}$ and $\widehat{AE}$ to denote the *average of errors* for the single-tier and two-tiered decision model respectively. Formally, the *average of errors* are defined as:

$$\widetilde{AE} = \frac{\sum_{z,t} |V_z^t - \widetilde{V_z^t}|}{|Z| \times |T|}$$
$$\widehat{AE} = \frac{\sum_{z,t} |V_z^t - \widehat{V_z^t}|}{|Z| \times |T|}$$

Table 5.5 shows the results of the *average of errors*. From the table, we can observe that the two-tiered decision model performs better in representing the zone visitation on the ground, as observed by the lower average of error

and standard deviation.

Table 5.5: Zone visitation comparison using average of errors.

	Single-Tier Model	Two-Tier Model
Average of Errors	1.00%	0.65%
Standard Deviation	1.50%	0.64%

The results for the zonal visitation is also as illustrated in figures 5.10 and 5.11. Figures 5.10 and 5.11 represent the zone visitation for a particular non-peak and peak hour respectively. Zones with higher visitation are indicated with darker shades of gray, and vice versa. From the figures, we can observe that the zone visitation from the two-tiered decision model represents that of the ground truth better, at both peak and non-peak hours of the day.

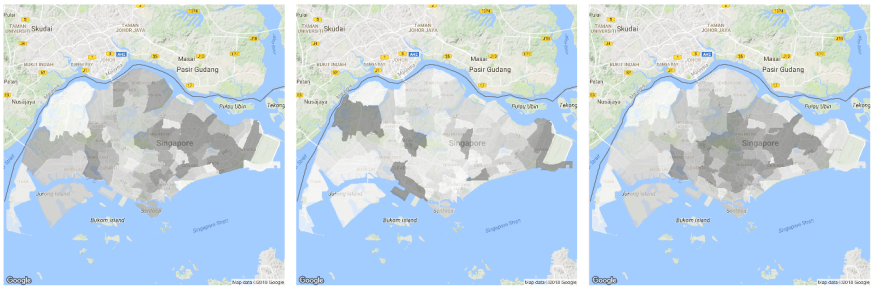

Figure 5.10: Zone visitation at non-peak hour: ground truth (left), single-tier (middle) and two-tier model (right).

Figure 5.11: Zone visitation at peak hour: ground truth (left), single-tier (middle) and two-tier model (right).

5.5.3.2 Metric 2: Weighted Average of Errors

The *weighted average of errors* is a weighted variant of the *average of errors*, where we multiply the absolute error with the actual (ground truth) zone visitation, V_z^t. We use the notation $\widetilde{WAE}$ and $\widehat{WAE}$ to denote the *weighted average of errors* for the single-tier and two-tiered decision model respectively.

Formally, the *weighted average of errors* are defined as:

$$\widetilde{WAE} = \frac{\sum_{z,t} |V_z^t - \widetilde{V_z^t}| \times V_z^t}{|Z| \times |T|}$$
$$\widehat{WAE} = \frac{\sum_{z,t} |V_z^t - \widehat{V_z^t}| \times V_z^t}{|Z| \times |T|}$$

Table 5.6 shows the results of the *weighted average of errors*. From the table, we can observe that the two-tiered decision model performs better in representing the zone visitation on the ground. This further confirms the performance of the two-tiered model in not only representing the zone visitation well, but especially for the zones with high visitation rates.

Table 5.6: Zone visitation comparison using weighted average of errors.

	Single-Tier Model	Two-Tier Model
Weighted Average of Errors	1.41%	0.95%
Standard Deviation	2.87%	1.71%

5.6 Case Studies: Evaluation of DGS System

In the previous section, we have presented our work on a two-tiered decision model for improving the modeling of taxi drivers and their movement decisions. In the case study section, we present three case studies. The first two case studies will focus on utilization of TaxiSim 2.0 as an evaluation testbed on the impact of a Driver Guidance System (DGS) for improving

the performance of taxi drivers. TaxiSim 2.0 was used recently [38] to evaluate the effectiveness of the DGS system developed, but it was for a single day demand scenario (i.e., November 2, 2016) under various penetration ratios. In our case, we perform a comprehensive set of virtual experiments to further investigate the effectiveness of DGS under different scenarios, over an extended period of time (i.e., single month used for training and testing each). In particular, we make use of TaxiSim 2.0 to (1) evaluate the effectiveness of the DGS system at various times of the day (e.g., peak and non-peak), and (2) conduct sensitivity analysis on the effectiveness of the DGS system at different penetration ratio. More details of the DGS system and virtual experimental setup will be as discussed subsequently.

5.6.1 Driver Guidance System

The driver guidance system is a centralized decision support system. The model provides recommendation to taxi drivers to move to a potential location in order to fetch a demand. The recommendation model utilizes the information of the current supply of taxis in each region (the size of a region is 1km x 1 km based on the cluster of demand) gathered using the real-time stream of taxi-logs along with the estimated passenger demand for a fixed time horizon. The recommendation model uses a multi step stochastic optimization formulation based on the models proposed in [47] [48]. It maximizes the total expected revenue obtained by taxi drivers (after deducting the movement cost) over a set of demand samples subject to following constraints:

- At any timestep, for any region, the number of taxis going out of a region is equal to number of taxis available in the region.
- At any timestep, for any demand sample, between any pair of regions i and j, the number of requests served is less than the number of

requests available between regions i and j.

Formally, the objective function for the optimization model is defined as:

$$\max(-\sum_{i \in Z} \sum_{j' \in Z} Cost^{1}_{ij'} \cdot)$$

In the current system, we only consider street bookings where for any demand sample a trip is assigned to taxi if and only if it is present in the same region as the pickup location. But the optimization formulation can also handle booking requests, where taxis can be assigned to customer requests from nearby regions.

The optimization formulation groups the taxi drivers in a region and provides recommendation in the form of number of taxi drivers which should move from region i to region j for the current timestep. The multistep optimization model is run at every timestep in a rolling horizon fashion in order to provide the real-time recommendations.

If only a subset of taxi drivers are guided using the recommendation model: The recommendation model formulation optimizes the total expected revenue of the guided taxis. If all the taxis in the system are not guided, the recommendation model first simulates the behaviour of unguided taxis by using the historical probabilities for transitioning from one region to another. Based on the movement of unguided taxis, they are assigned trips from the demand samples with a fixed probability (say 0.5). The demand samples are then modified to exclude the trips which are assigned to unguided taxis. The optimization formulation is executed for the guided taxis using the modified demand samples.

5.6.1.1 Simulation and Virtual Experimental Setup

To prepare for the evaluation, we introduce a new mode of taxi navigation: the dynamic route guidance, which is linked to the simulator to compute itineraries for routed visitor agents. We vary the fraction of agents having this technology (defined as penetration ratio), and assume that all other agents would follow the default two-tiered decision model described earlier. Additional, we have various time periods where DGS movement takes over from the normal roaming mode (i.e., two-tiered decision model), which allows us to test the effectiveness of DGS application at various times of the day. This will be further discussed in subsequent subsections.

5.6.2 Use Case 1: Performance at various times of day

For the first use case, we look into the performance of the driver guidance system at various times of the day/week, which includes (1) weekday peak, (2) weekday non-peak, and (3) weekends. For each scenario, we include various multiple time period for virtual experimental purposes.

Figure 5.6 explains the motivation for including multiple peak/non-peak period for virtual experimental purposes. From figure 5.6, we can observe distinct differences in the Origin-Destination travel patterns of passengers at different peak time periods of the day. It would therefore be important for us to not only evaluate the performance of DGS at various scenarios (peak/non-peak), but also at different time periods with varying Origin-Destination travel patterns.

We define the following time periods for evaluation of the DGS system under various demand scenarios, which is illustrated in 5.12. Figure 5.12 shows the demand on a typical weekday, with the peak and non-peak period indicated in red and orange respectively. The time periods are: M-Peak

(morning peak: 6am to 9am), A-Peak (afternoon peak: 12pm to 3pm), E-Peak (evening peak: 6pm to 9pm), M-Non-Peak (morning non-peak: 9am to 12pm), and A-Non-Peak (afternoon non-peak: 3pm to 6pm). Multiple peak/non-peak periods are added due to difference in demand "hotspots" at various times of the day. Additionally, we compared the effectiveness of using the application on weekdays and weekends for an entire day (defined as 6am to 9pm). To ensure fairness, we have an equal duration (3 hours per time period) for comparison purposes.

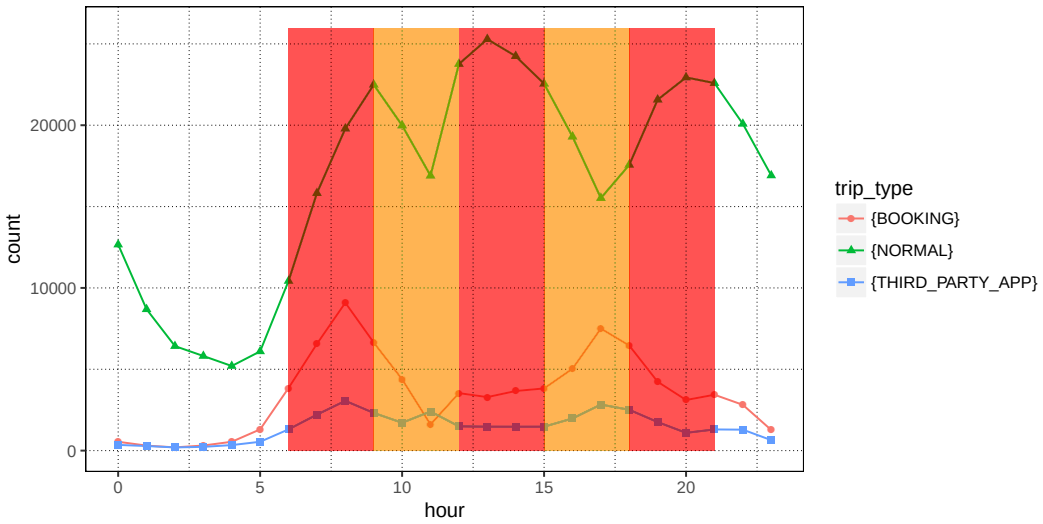

Figure 5.12: Peak (red) and non-peak (orange) periods on a typical weekday in November 2016.

5.6.2.1 Virtual Experiment Evaluation

We define a couple of KPIs (Key Performance Indices) for evaluating the performance of the DGS application at various times of the day. Firstly, we have fulfilment rate, which is essentially the proportions of trips that were fulfiled within an virtual experiment time period (e.g., 9am to 12pm). Formally, this is defined as:

$$fulfilmentRate = \frac{fulfilledTrips}{fulfilledTrips + unFulfilledTrips}$$

Thereafter, we define another KPI: zonal imbalance. The rationale for this is to test if the DGS application does indeed resolve imbalances in supply and demand during the DGS guidance period. Formally, this is defined as:

$$zonalImbalance = \sum_{t \in T} \sum_{z \in Z} |demand_z^t - supply_z^t|$$

Additionally, we included another performance indicator for zonal starvation, which is essentially measuring zonal imbalance, but only if the demand is greater than supply. This serves as another perspective on the performance of the DGS system. Formally, it is defined as:

$$starvation_z^t = \begin{cases} (demand_z^t - supply_z^t) & \text{if } demand_z^t \geq supply_z^t \\ 0 & otherwise \end{cases}$$

$$zonalStarvation = \sum_{t \in T} \sum_{z \in Z} starvation_z^t$$

5.6.2.2 Virtual Experiment Results

Results of the virtual experiments is as shown in table 5.7.

Table 5.7: Performance of DGS application at various times of the day.

Time Interval	6am-9am	9am-12pm	12pm-3pm	3pm-6pm	6pm-9pm
Fulfilment Rate	91.8%	92.8%	90.9%	90.5%	89.5%
Zonal Imbalance	208.5	171.3	185.3	178.2	189.59
Zonal Starvation	34	51	75	44.5	58.1

5.6.3 Use Case 2: Performance under various penetration ratio

To evaluate the performance of the guided vehicles under different penetration ratios, we utilized the following KPIs for evaluation: (1) fulfillment rate, (2) inter-job time, and (3) trips per taxi. Details of the virtual experiment results will be discussed subsequently.

5.6.3.1 KPI 1: Fulfillment Rate

The first key performance index is the fulfilment rate. Similar to the first use case, it is formally defined as:

$$fulfilmentRate = \frac{fulfilledTrips}{fulfilledTrips + unFulfilledTrips}$$

Figure 5.13 shows the fulfillment rate under different penetration ratios of DGS taxis. From the figure, we can observe that the fulfillment rate increases from 90.01% to 93.7% at 80% DGS, after which there's a slight drop thereafter (93.25% at 100% DGS).

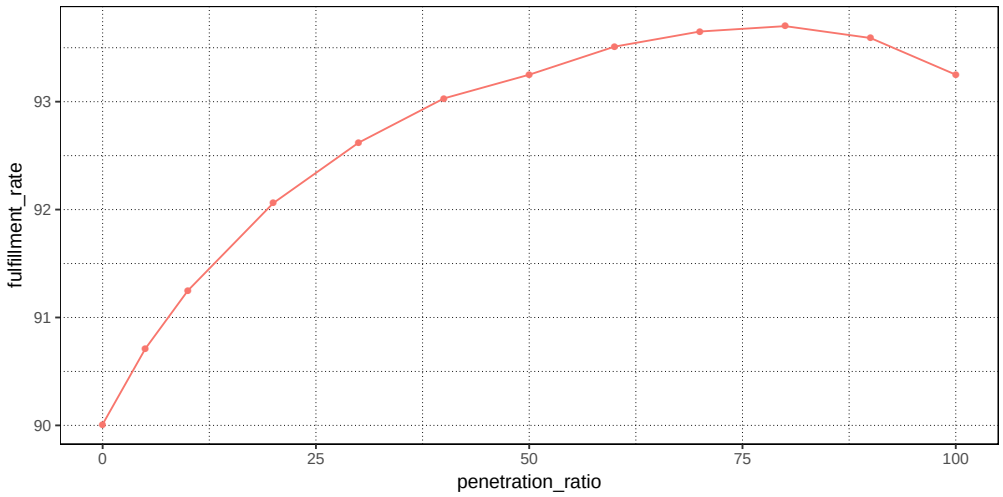

Figure 5.13: KPI: Fulfillment Rate

5.6.3.2 KPI 2: Inter-Job Time and Percentage

The next key performance index that we adopt is the inter-job time/percentage. Given that time spent roaming do not generate revenue for the driver agents, we would want to reduce this roaming time (defined as inter-job time) across all drivers. Formally, the Inter-Job Percentage is defined as:

$$interJobPercentage = \left(\frac{freeModeTime}{freeModeTime + hiredModeTime} \times 100 \right)\%$$

Figure 5.14 provides an illustration for the computation of inter-job time and percentage for a single driver, with active or working hours from t_0 to t_5. The time period where the driver spends time in *job mode* (i.e., status = "hired") is indicated by the shaded time periods. The *Inter-Job Time* (IJT) would therefore be the periods of time where driver spends searching for passengers (indicated by "Free Mode"), while the *Inter-Job Percentage* (IJP) normalizes this against the total active time period.

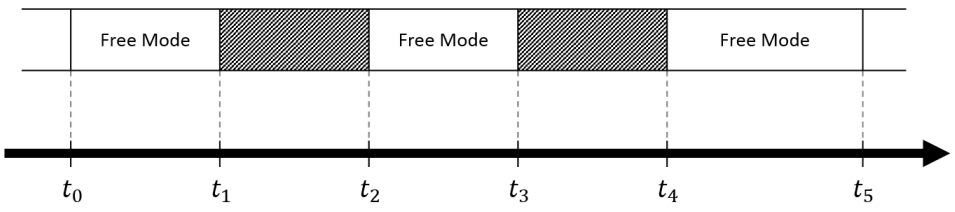

Figure 5.14: KPI: Illustration of Inter Job Time/Percentage Computation

Figure 5.15 and 5.16 shows the average inter-job time (under different penetration ratios of DGS taxis). From the figure, we can observe that the performance (defined as least inter-job time) for the DGS taxis is the best at 5% DGS (i.e., penetration ratio), following which the performance drops subsequently.

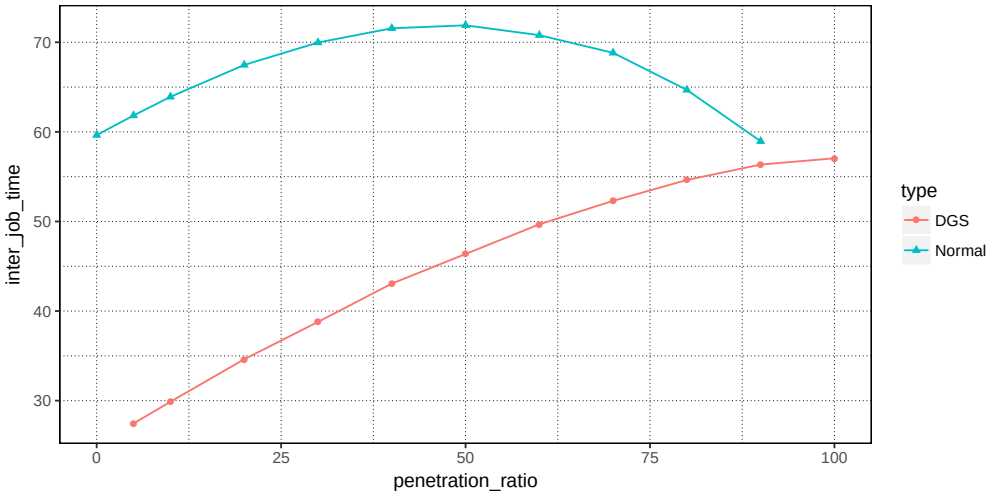

Figure 5.15: KPI: Inter-Job Time for Street Hail.

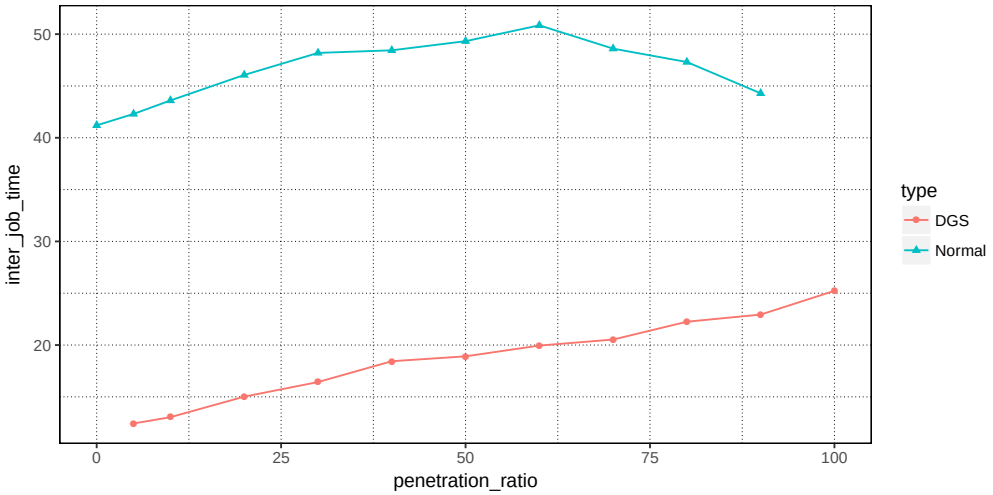

Figure 5.16: KPI: Inter-Job Time for Bookings.

5.6.3.3 KPI 3: Trips per Taxi

Last but not least, we use the average number of trips per taxi as the third key performance index, which is defined as:

$$averageTrips = \frac{numberOfTrips}{numberOfTaxis}$$

Figure 5.17 shows the trips per taxi under different penetration ratios of DGS taxis. Similar to the figures for inter-job time, we can observe that the performance (defined as a higher number of trips) is the best at 5%,

following which it decreases with an increase in penetration ratio. We can observe that the trips per taxi for DGS vehicles is higher than non-DGS vehicles, which applies under different penetration ratios (from 5% to 90% penetration ratio). Additionally, we can observe that the average trips per taxi for the entire population (i.e., DGS and non-DGS taxis) increases till about 60% (illustrated by red line in figure 5.17), after which it decreases slightly.

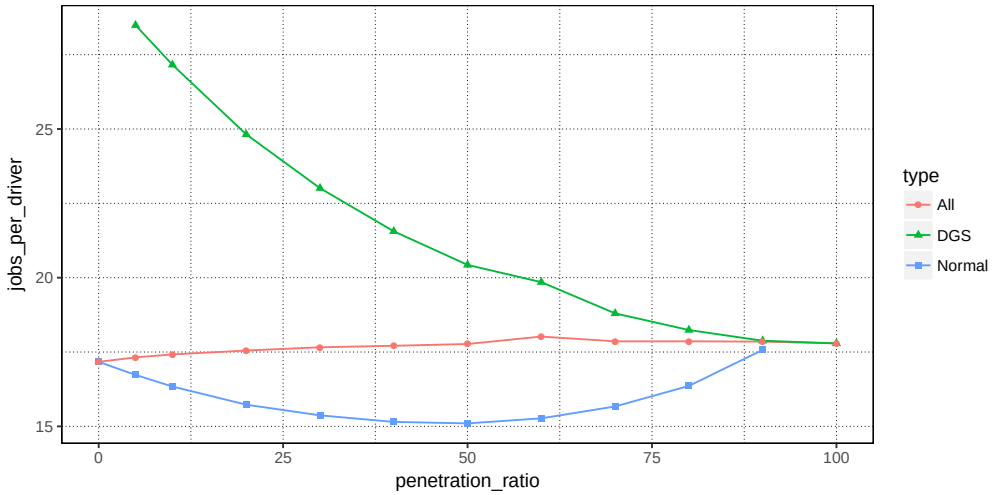

Figure 5.17: KPI: Trips per Taxi

5.7 Future Directions

This work is just the beginning of our research in the application of hierarchical decision models in building realistic multi-agent simulation systems in the transportation domain. In this section, we discuss some future directions that we could potentially look into.

Firstly, in this work, we look into hierarchical decision models in terms of road network levels, particularly in terms of link and zone level. For future work, we look into hierarchies in terms of various levels of thinking and reasoning sophistication. Example of such a model would be the Cognitive Hierarchy Model by [14]. Given that we do not expect drivers to think at the same level of sophistication, it would be interesting to

test out such a model with the existing data and model available. Such a model is intuitive and can be observed in the actual data, where there is differing revenues earned by drivers who worked for a same amount of time. Successful implementation of such a model could further increase the realism of the existing model.

Additionally, another future direction could be application of the *two-tiered decision model* in other sub-domains of transportation, where individual agents seek to maximize their gained utility.

5.8 Conclusion

A lot of effort has been directed to research in the area of transportation. Similarly, a lot of work has been done in building agent-based simulation models that seek to mimic the behavior of agents in transportation. However, many of these existing models often make over-simplifying assumptions, or are mainly used for visualization purposes, with lesser emphasis on the calibration efforts.

With increasing research efforts in AI being applied in the area of transportation, it becomes imperative to create realistic simulations that serve as an economical approach for evaluating agent strategies, in order to solve problems in transportation that includes congestion and inefficient utilization of resources. To this end, we introduce a *two-tiered decision model* that mimics the decision-making process by taxi drivers that seek to maximize individual utilities (i.e., revenue). On top of that, we utilized TaxiSim 2.0 as a simulation platform, and compared the performance of a two-tiered decision model against that of a single-tiered decision model. Our initial results show great promises, supporting the usage of tiered decision models in the transportation domain.

5.9 Appendix

5.9.1 Data and Pre-Processing

In this section, we describe the data utilized for the model, as well as some data pre-processing that was done to obtain the intermediate variables that is required for the two-tiered decision model and analysis.

5.9.1.1 Data Description

To illustrate the *two-tiered decision model*, we worked with operational data from all taxi operators in Singapore. Every taxi in Singapore is equipped with a Mobile Data Terminal (MDT) that is designed to provide basic GPS functionalities and transmit current vehicle positions every 30 to 60 seconds to the central server, along with other information such as taxi status. For the purpose of this study, we worked with time period of November 2016 (26,026 unique taxi ids) and December 2016 (26,072 unique taxi ids) for training and testing respectively.

Some fields in the taxi logs data obtained are as shown in table 5.8. As observed in table 5.8, the raw data only contains basic information such as latitude, longitude and status, which is not sufficient for our link and zonal transition model (discussed in section 5.4 subsequently). We therefore proceed to infer links/zones/trips information, as discussed in the subsequent subsections.

Table 5.8: Description of Fields in Logs Data.

Variable	**Description**
timestamp	Date and time of log (sent every 30 to 60 seconds)
taxi-id	Anonymized ID of taxi vehicle
longitude	Longitude of vehicle at the time of log
latitude	Latitude of vehicle at the time of log
speed	Speed of vehicle at the time of log
taxi-status	Status of taxi: Free, Hired, Busy, On Call or Offline

5.9.1.2 Pre-Processing: Obtaining Link and Junction Information

As mentioned previously, link/junction information is missing in the raw GPS taxi logs that is obtained. A simple approach here would be to map each lat/lon data point to the nearest link/junction. However, such an approach suffers from inaccuracies of GPS data, which occurs for various reasons e.g., when going through a tunnel. To better infer the link and junction information, we made use of the Map Matching algorithm [57] that is based on GraphHopper [34], which maps the latitude and longitude information of every taxi log to a series of links (defined as a path). The algorithm essentially uses a Hidden Markov Model (HMM) representation for the road segments and their transitions across time. The Viterbi-based algorithm then computes the best path through the HMM latice. The final output will then be a most probable path of links traversed by a driver.

5.9.1.3 Pre-Processing: Obtaining Zonal Information

With the link information derived, it then becomes easier to obtain zonal information. Zonal definition is obtained from URA's website [75], where postal sectors refer to the first one or two digits of postal code of an area (which can be five or six digits long). The zone defintion is as illustrated in figure 5.18. To derive the zonal information, we get the polygon for every zone in Singapore, and links (derived previously) that fall within the polygon are determined as such.

5.9.1.4 Pre-Processing: Obtaining Trips Information

As trip information is not available in the raw data, we would have to infer trip-level information by tracking change in status. The trips information derived would then be used for purposes such as determining zonal demand. That is, when the status of a taxi changes from "free" or "busy" to hired, this is considered as the start of a trip. Similarly, when the status changes

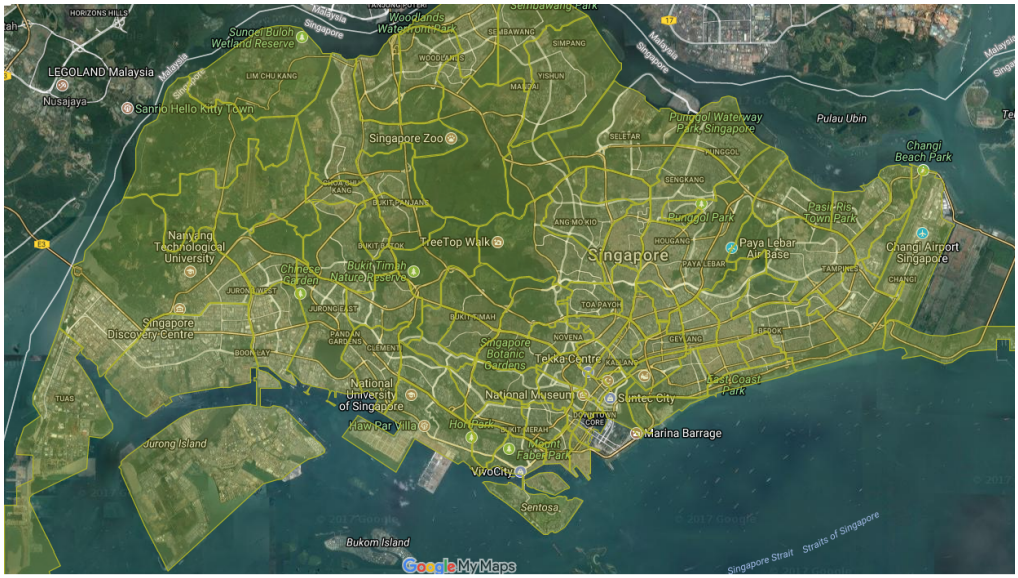

Figure 5.18: Zone Definition of Singapore. Boundaries are indicated by white lines.

from "hired" to "free", this indicates the end of a trip. After the initial trips are generated, we remove anomalies by additional filters: (1) Trips with only one data point (e.g., system error caused temporary change in status) and (2) Trips where the taxi did not move at all. Further details of the trips information inference is as summarized in table 5.9.

Table 5.9: Summary of Trips Inferred from Change in Status

Status Change	**Trip Type Inferred**
Free ->Hired ->Free	Street Hail
Free ->On Call ->Hired ->Free	Normal Booking
Free ->Busy ->Hired ->Free	Third Party Booking

From the inference of trips via change in status, we then proceed to derive other information such as the trip duration, distance travelled, etc. A thing to note here is that as the trips-specific information are inferred (instead of obtaining directly from the operators), we do not have trip information such as revenue (collected from a trip). Therefore, in the designing of the zonal level movement (discussed further in section 5.4), we only make use of number of trips (i.e., demand) as a proxy for "zonal attractiveness". As an example, some trip-level information that are inferred

is as shown in table 5.10.

Table 5.10: Description of Fields in Trips Data.

Variable	**Description**
taxi-id	Anonymized ID of taxi vehicle
start-time	Start time of the trip
end-time	End time of the trip
duration	Duration of the trip
start-latitude	Latitude of vehicle at the start of trip
start-longitude	Longitude of vehicle at the start of trip
end-latitude	Latitude of vehicle at the end of trip
end-longitude	Longitude of vehicle at the end of trip
distance	Distance travelled
start-zone	Zone that the driver started the trip
end-zone	Zone that the driver ended the trip

5.9.1.5 Pre-Processing: Obtaining Zonal Dwell Time and Transition

Subsequently discussed in section 5.4, one of the key concepts of the two-tiered decision model is the zonal jump, where a taxi would decide to travel to a zone (could be adjacent or not) based on a discrete choice zonal-level decision model. As the current set of pre-processed data only contains movement of agents moving through links/zones, we do not have actual data on agents making a conscious decision to make a "zonal jump".

More specifically, we are not interested in the zonal movements where (1) drivers move into another zone due to roaming at link-to-link level and (2) driver is "passing by" other zones (defined as *transient zones*) when getting to their destination zone in search of passengers. This raises the problem of how we can better understand the zonal movement intention in a coherent manner, i.e., correctly identify the zone *origin-destination* decisions made by taxi drivers.

One way to do this is to determine the *dwell time* spent in zones. That is, the longer the time spent in a zone, the more likely that the zone was the drivers destination zone. Therefore, we introduce a *zone-specific dwell time threshold*, which filters out zonal transitions where a taxi does not

spend a minimum period of *dwell time* at a zone. With the zone-specific dwell time threshold introduced, it then allows us to effectively filter out the zonal transitions that are *transient*, i.e., the zones that a taxi would "pass by" when getting from its origin to destination zone.

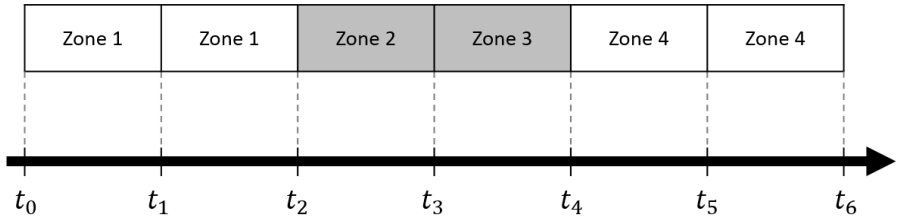

Figure 5.19: Obtaining Zonal Transition with Zone-Specific Dwell-Time Threshold.

An illustration of the *zone-specific dwell time threshold* filtering is as shown in figure 5.19. From the individual logs of taxis, we derive the zone that a taxi spends most of its time in for every 5 minutes interval, as well as the dwell time. For zone dwell time records that are less than the *zone-specific dwell time threshold* value (indicated in gray), the zone transition will be taken as an *invalid* or *transient* zone movement. An example would be the transition from period t_2 to t_4, which we assume to be less than the threshold for now. Therefore, this renders the transitions in gray as invalid. The final record will therefore be a transition from zone 1 to zone 4, with a dwell time of $t_2 - t_0$ at zone 1 before transiting.

Below are the steps followed to derive the *zone-specific dwell-time threshold*:

- We first filter by taxis in "Hired" mode.
- We then computed the time taken for taxis in hired mode to "cross" the zones, and generate a cumulative distribution for them. Example of the cumulative distribution function for a single zone for a peak hour is as illustrated in figure 5.20.
- From the cumulative distribution generated, we then took the 95th percentile value as the *zone-specific dwell-time threshold*. What this

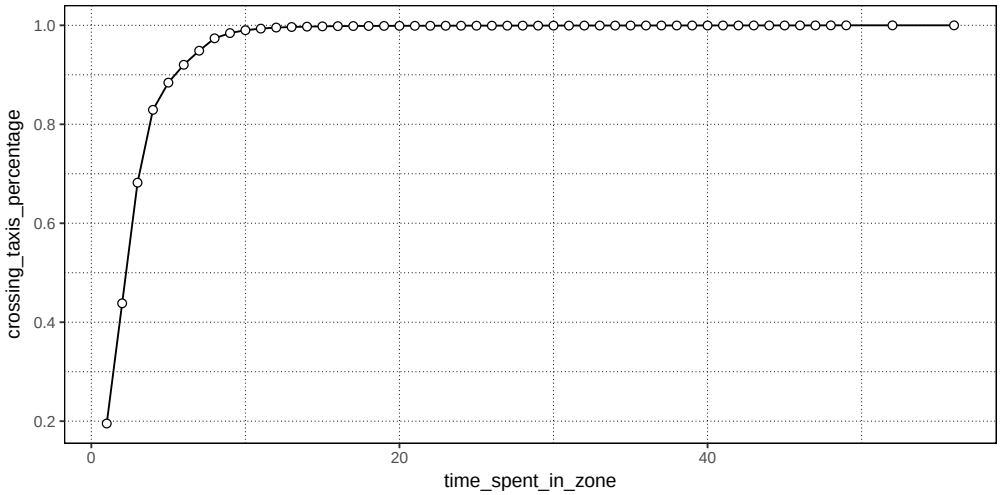

Figure 5.20: Zone-Specific Threshold Time Illustration.

means is that at this dwell time value, 95% of taxi would have crossed or moved out of the zone. Using figure 5.20 as a reference, we could observe that a significant number of taxis would have "crossed" zones even at 5 minutes.

Chapter 6

Macro-level, Partially Observable: Migration Domain

Overview

In this chapter, we look into development of high quality agent-based simulation model in the macro-level, partially observable domain. This is achieved through the use of a country-level agent-based dynamic network model to examine shifts in population given network relations among countries, which influences overall population change. Some of the networks considered include: alliance networks, shared language networks, economic influence networks, and proximity networks. Validation of model is done for migration probabilities between countries, as well as for country populations and distributions. The model developed provides a way to explore the interaction between climate change and policy factors at a global scale.

6.1 Introduction

Human migration is an important research topic, with major economic effects [59]. At the same time, the decision to migrate is also determined by economic factors [63]. This intertwining effect is best captured by agent-

based models, where agents interact with their environment, and changes in the environment affects the decisions of agents.

Previous studies on human migration generally make simplistic assumptions about the decision model of migration, or do not consider fluctuations in birth/death rates, age distributions as well as networks ties between countries. With recent economic trends [33], birth policy changes [13] and climate trends in mind, it is important to develop an agent-based model that is sensitive to these changes. In this work, we developed a country-level agent-based model which aims to mimic the agent's decision-making process for migration. This is done through consideration of a range of country networks, ranging from alliances to linguistic similarities to climate and migrant networks, just to name a few.

Additionally, we initialize the age distributions of countries according to actual data [76]. The age distribution is then shifted throughout the simulation through an aging process, as well as actual births and deaths in population. We then validate our model against data for migration probabilities, and country-level observations (population and age distributions). The results are promising, as we illustrate through performance measures such as average of prediction error.

6.2 Motivation

Our work is mostly motivated by recent trends. Given news on possible major shifts in population [2], this presents huge opportunities and risks [2]. Also, with international migration contributing to more than half of population growth in developed nations [64], it is important to understand the dynamics involved in migration, as well as build models that capture the complex interactions leading to these shifts in population. Building these models will in turn help policymakers to make better decisions. For

this purpose of this work, we do not model domestic migration, as it does not lead to shifts in population of countries, which is the main output of the model that we are developing. We do however, refer to previous work done on domestic migration, and incorporate some of the features into our model, including network externality effects.

The use of an agent-based and other models (e.g., flow-based models) for modeling migration is not a new research topic or area. However, many of them have in place simplifying assumptions, or do not consider a sufficient number of country networks, which we felt led to *less-than-realistic* models, especially if we are looking at migration on a global scale.

For example, [40] developed a comprehensive framework (including cost of moving, pay at different states) on the factors that affects domestic migration. The authors [40] cited data limitations as a reason for working on domestic migration. Such a framework could be used for international migration as well, where economic disparity network between countries would serve as a driver of migration. However, international migration would also require looking at other networks such as linguistic similarities (e.g., English speakers would prefer to move to English speaking countries), which does not apply for domestic migration.

[68] looks into rainfall as a predictor of migration decision, as well as network effects such as network peers affecting the propensity of migration. It works well in the area that the simulation is run, where livelihood of farmers is dependent on rainfall, which affects crops. However, we felt that more networks need to be considered if modeling migration at a global level, such as economic networks, where there are stylized facts on more developed nations drawing more immigrations [64].

[25] developed a flow-based, spatial-interaction model, which simulates domestic migration as a function of distance between origin and destinations. A distinction between the flow-based model and what we are working is

that migrants to a destination are not placed into a common homogeneous population pool. Instead, they are placed into separate sub-populations, in a Origin-Destination Migrant format, where they have a positive network externality effect for future migrants (further discussed in section 6.3). [3] came up with an interesting multi-evolutionary agent-based model, where agents undergo a cognitive procedure to migrate given certain age, happiness and wage criterions.

Therefore, considering the strengths and limitations of the previous works, we develop an country-level agent-based dynamic network model that takes into consideration a range of country networks (discussed in Section 6.3). It is also important to note here that our approach can be replaced with a system dynamics approach, although it would be cumbersome [32] to model the numerous population stocks involved (discussed in Section 6.3). Some characteristics of the ABM includes:

- Mesoscopic modeling: contrary to previous works on migration involving microscopic (agents as individuals) or macroscopic (homogeneous population pools within countries) models, we adopt a mesoscopic model approach, where agents are countries, and we model sub-populations within countries, where there are network externalities effect on potential migrants in future. Sub-population here refers to migrants networks, or migrants from country i residing in country j. These migrant networks would have a positive externality effect, thereby increasing the propensity of migrating. This will be further discussed in section 6.3.

- Heterogeneity amongst agents: Every country has their own migration policies, with differing "openness" towards immigrants [46]. While the policies are not publicly available, we used previous migration numbers as proxies for migration policies of countries. These are

further discussed in section 6.3. Every country also has their own age distribution, which is initialized using actual data [76].

- Interactions amongst agents: Decisions made by agents will affect other agents. For example, migration of population from country i to j will alter the age distribution and Origin-Destination Migrant Stock of country j (further discussed in section 6.3), which has an effect on future migration. For future implementations, we could even look at language distributions (e.g., number of Spanish speaking citizens in a country), and investigate shifts as a result of immigrants arriving at a country.

- Decision Model for Migration: We took into consideration previous works that were mentioned, and modeled the decision to be dependent on several country networks. We believe that this would make for a better decision model, as opposed to only considering climate [68] or economy [40] networks.

We believe that this would make for a better model which is sensitive to economic, climate trends, as well as considering other network ties between countries.

6.3 Agent-Based Model for Human Migration

We developed an Agent-Based Model (ABM) using NetLogo 3D 5.2.0. The interface of the ABM at initialization and during simulation is as shown in Figure 6.1. Various components of the simulation model are as mentioned in the sub-sections below.

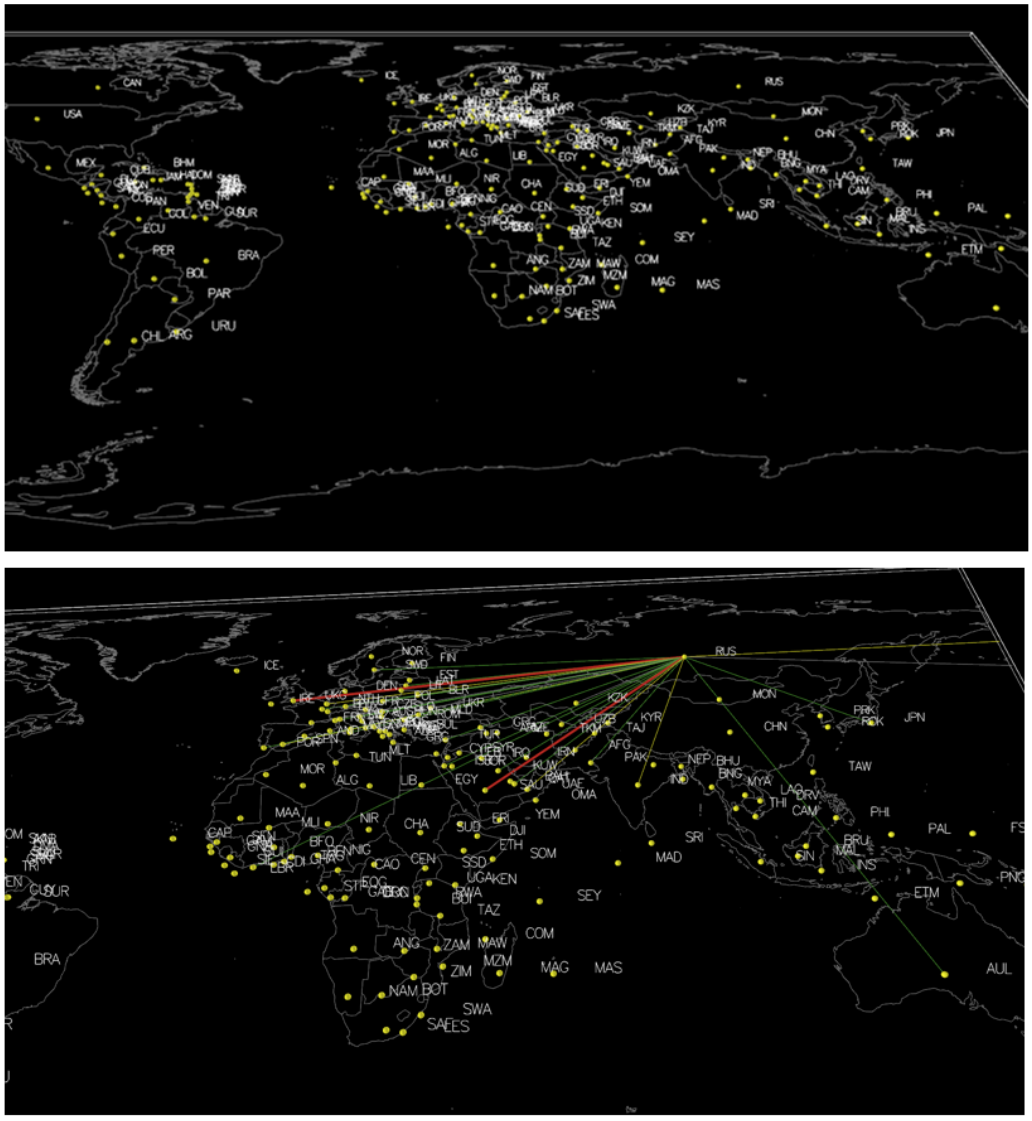

Figure 6.1: Interface of ABM at (left) initialization and (right) during simulation. Lines colored in red/yellow/green indicate high/medium/low migration numbers.

6.3.1 Data and Processing

Our core database consists of 194 countries, with static variables such as country name, code, as well as latitude and longitude. Yearly country statistics (e.g., GDP) is obtained from [80] [37], as well as official government [76] sites. When data is not available , we make approximations based on the indices of "similar" countries, as well as linear interpolation of data. We made use of data from years 2000 and 2010 (bilateral migration data available) for calibration of ABM, and years 2011 to 2013 for testing.

6.3.2 Countries Network

Several types of countries network are considered in this work. These are as discussed subsequently:

6.3.2.1 Alliance and Hostility Network

For constructing the alliance and hostility network between countries, we gathered data from the Global Database of Events, Language, and Tone (GDELT) [70]. Events that occurred between countries are mapped to a CAMEO (Conflict and Mediation Event Observations) scale ranging [-10, 10] specified by [15]. The average of these "interactions" between countries is then converted to [-1, 1] scale, where -1 indicates a negative (hostile) relationship, and 1 indicates a positive (alliance) relationship between two countries. A value of 0 represents a neutral relationship between two countries.

6.3.2.2 Linguistic Similarity Network

For the linguistic similarity network, countries are tied by their dominant or major languages, and individuals are more likely to migrate to those countries which are linguistically similar. Also, as observed in Table 6.1, having more than one language in common does not provide better prediction

Table 6.1: OLS regression results: common language (integer and binary) as independent variable; migration probability as dependent variable.

		Estimate	$Pr(<\|t\|)$
Integer	(Intercept)	0.023678	9.55e-07 ***
	Common.Language	0.104941	<2e-16 ***
Binarized	(Intercept)	0.027132	2.94e-08 ***
	Common.Language	0.100437	<2e-16 ***

accuracy in estimating the migration probability. The linguistic similarity ties between countries is defined as $CommonLang_{i,j}$, where a value of 1 refers to country i and j having at least one common language, and 0 otherwise.

6.3.2.3 Proximity Network

A proximity network is computed by taking the distance (computed through the haversine formula) Using the center latitude/longitude of the 194 countries in our core database, a proximity network is developed, with $Dist_{i,j}$ representing the distance (computed through haversine formula) between countries i and j.

6.3.2.4 Sea-Level Network

For modeling sea-level networks, we used the "population below 5 meters" data from [80], represented as $PopBelow5m_i^t$, which shows the percentage of population in country i that is below 5 meters in year t. Do note that this data is aggregated at country-level, so for some of the larger countries, there might be disparity in various regions of the nation (e.g., coastal and inland regions), which is not captured here. Each network tie between countries i and j represents the difference in this factor, which is defined as:

$$PopBelow5m_{i,j}^t = PopBelow5m_j^t - PopBelow5m_i^t$$

6.3.2.5 Economic Influence Network

We used the Gross Domestic Product per Capita of countries (obtained from [80]) as a measure of economic influence, or $GDP^t_{i,j}$, which is formally defined as:

$$GDP^t_{i,j} = \frac{(GDP^t_j - GDP^t_i)}{(GDP^t_j + GDP^t_i)}$$

Where GDP^t_i and GDP^t_j represents the GDP of countries i and j in year t. This economic influence network will then be used to model citizens of countries with lower GDP displaying a higher propensity of migrating to countries with higher GDP [63].

6.3.2.6 Migrant Network

Last but not least would be migrant network. In the work done by [68], the authors developed an ABM whereby an individual is connected to 10 other individuals, and the propensity of migration depends on their networked peers. In our work, we emulate a similar network externalities effect, where increase in migrants from a origin country i residing at a destination j will increase the propensity of others (from country i) migrating to country j. This is formally defined as:

$$ODProportion^t_{i,j} = \frac{OD^t_{i,j}}{Pop^t_i}$$

Where $OD^t_{i,j}$ refers the OD (origin-destination) migrant stock in year t, and Pop^t_i refers to population of country i in year t. $ODProportion^t_{i,j}$ is just the normalized version of $OD^t_{i,j}$. Thus, an example here would be more Mexicans residing in the US will have a positive externality effect on future migrants, increasing the propensity of migrating.

Model-wise, what this means is that when there are individuals from multiple origin countries migrating to a destination country, they will not be placed into a homogeneous population pool. Instead, they are placed

into separate sub-populations, and undergo births/deaths, similar to the native population. While the granularity of model is unlike that of [68] (with the networks of individuals), these "sub-populations" residing in the destination country create a similar positive externalities effect. The OD migrant stock is initialized with actual data obtained from [74], and updated at every time interval of simulation when migration occurs. Similar to the native population, the OD migrant stock undergoes births and aging.

6.3.3 Migration Decision Model

For modeling the decision to migrate, we considered various countries networks (as discussed previously). We define the migration decision as $P^t_{i,j}$, the probability of migrating from country i to j at year t, and is formally defined as:

$$P^t_{i,j} =$$
$$F(GDP^t_{i,j}, PopBelow5m^t_{i,j}, CommonLang_{i,j}, Dist_{i,j}, Alliance_{i,j}, Migrant^t_{i,j})$$

In this work, we take the migration decision to be a linear combination of the various indicators, although it could take on other forms as well (e.g., Logit model). For calibrating the migration decision model, we take the bilateral migration numbers between countries i and j in year t, and express it over the native (non-migrant) population of a country i at year t to obtain the migration probability $P^t_{i,j}$. We currently assume that migrants do not partake in future migration, which we feel is a reasonable assumption, given the time frame we are looking at. Formally, this is defined as:

$$P^t_{i,j} = \frac{Mig^t_{i,j}}{Pop^t_i}$$

Validation of the simulation results will be further discussed in section 6.4.

Table 6.2: Individual OLS regression result for migration decision model, with the individual network ties as independent variable, and the probability of migrating as dependent variable. *** indicates significance at the 0.1% level; ** indicates significance at the 1% level; * indicates significance at the 5% level

	Estimate	Pr (>|t|)
$GDP^t_{i,j}$	0.049680	<2e-16 ***
$PopBelow5m^t_{i,j}$	-0.09710	4.49e-16 ***
$Migrant^t_{i,j}$	1.003172	<2e-16 ***
$CommonLang_{i,j}$	0.116152	<2e-16 ***
$Dist_{i,j}$	-0.215794	<2e-16 ***
$Alliance_{i,j}$	0.069994	<2e-16 ***

Table 6.2 shows the individual OLS regression results of the decision model, which supports why the individual country networks were selected. For future works, we could perform network analysis such as Multicollinearity Robust QAP (MRQAP) to explore the network effects involved.

6.3.4 Age Distribution

This is one of the important features of the model. The purpose of the age distribution is two-fold; not only does it help in the validation process (discussed subsequently), it also serves as a way for policy makers to project possible shifts in age distributions of a country.

The age distribution of countries are initialized using actual population and international migrant stock numbers from [80], as well as age distribution figures from [76]. Shifts to the distribution are introduced through aging (shift to the right), births (addition), deaths (removal), as well as migration (moved from source to destination country).

The age distribution for all countries are separated into migrants (in blue) and native/non-migrants population (in orange), as illustrated in Figure 6.2. For citizens migrating to another country, they are placed into the migrant population of the new country, where they do not partake in future migration decision process. This ensures that every citizen migrates

only once, which we felt is a reasonable assumption given the time frame (for testing) that we are looking at.

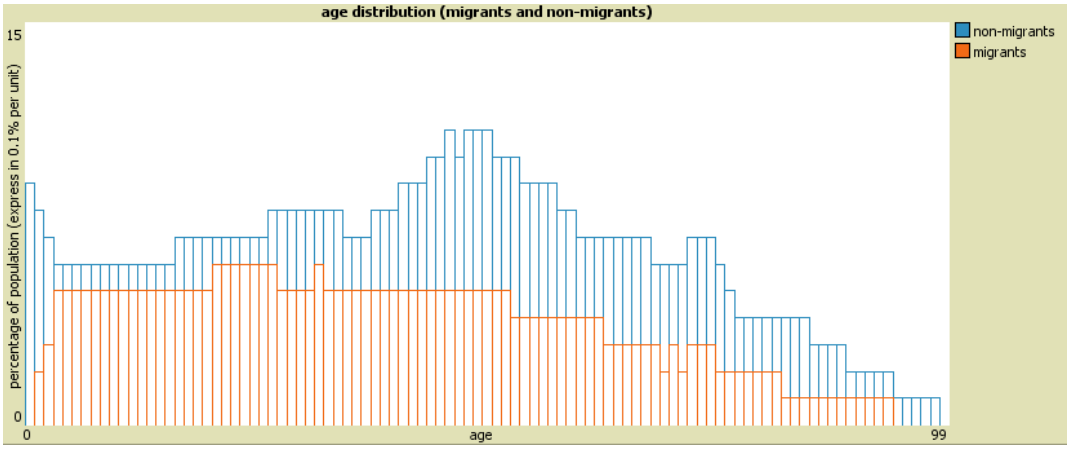

Figure 6.2: Age distribution for country at initialization. Distributions for migrants and native (non-migrants) are in orange and blue respectively.

6.3.5 Limit on Migration

As mentioned previously, migration policies vary greatly across different countries. For example, Canada has foreign-born population making up 21.3% of its population, while Japan has a much lower 1.7% foreign-born population [46]. While exact migration policies (i.e., exact numbers allowed for immigrants) are not publicly available, this is an important feature that would make for a realistic ABM. Thus, we used previous bilateral migration numbers and population data as a proxy for the openness of a country towards immigrants. The migration limit for a country j is formally defined as:

$$MigrationLimit_j = \max_t(MigrationProportion_j^t) \ , \forall j$$
$$MigrationProportion_j^t = \frac{Immigrants_j^t}{Pop_j^t} \ , \forall j, t$$

where $Immigrants_j^t$ refer to the number of immigrants arriving at country j in year t.

Input Variables	**Description / Values**
Country Networks Migration Limits Age Distribution	Described in section 6.3
Dependent Variables	**Description / Values**
Population	Pop_i^t
Age Distribution	Measured at end of simulation.
Number of Replications	30

6.4 Model Validation

Model validation is done by validating the simulated populations, age distributions, as well as the migration probabilities. This is done using average of errors as a performance measure. It is imperative to note that we are not developing a migration prediction model that perfectly fits against historical events. Instead, the focus here is on building an ABMS (Agent-Based Model and Simulation) that is sensitive to real-world trends and the evolution of country networks (mentioned in section 6.3). It is also worth noting that the data for the less developed countries might suffer from poorer data accuracies [64], which might affect the projections for these group of countries.

6.4.1 Migration Probabilities

For validating the migration probabilities, the actual migration probabilities is compared against simulated migration probabilities. The average of error (AE) for migration probabilities is formally defined as:

$$AE = \frac{\sum_{i,j,t} |P_{i,j}^t - \widehat{P_{i,j}^t}|}{|I| \times |J| \times |T|}$$

where $P_{i,j}^t$ and $\widehat{P_{i,j}^t}$ represents the actual and simulated probability of migration between country i and j at year t.

Table 6.3 shows the results of validation for migration probabilities between all countries, as well as between the top 50 most populous countries

Table 6.3: Validation Results: Flow Probabilities between all countries and populous countries. Populous countries refer to top 50 countries by population at initialization.

	Populous Countries	**All Countries**
Average Error	0.037 %	0.057 %
Standard Deviation	0.253 %	1.346 %

(at initialization). From the table, we can observe that the model is able to capture the migration movement of countries fairly well. Also, noting the better accuracy (lower average error) for the populous countries, this means that the model makes relatively better estimates for the populous countries, which is good, as poor estimates for the populous countries will have a bigger impact as opposed to the less populous countries.

6.4.2 Population Validation

For validating the simulated population numbers, we compare the simulated population numbers against actual population numbers obtained from [80], and normalize against the actual population numbers. The average of error here is defined as:

$$AE = \frac{\sum_{i,t} \frac{|Pop_i^t - \widehat{Pop_i^t}|}{Pop_i^t}}{|I| \times |T|}$$

where Pop_i^t and $\widehat{pop_i^t}$ represents the actual and simulated population of country i at year t.

Figure 6.3 shows the average error over time for (in orange) all countries and (in blue) top 50 populous countries. First observation here is that the error increases over time, which is not surprising, due to the compounding of errors [64]. Also, we can observe that the average error for the top 50 populous countries remains significantly lower than that of all countries over time. This shows the performance of the model in predicting the population, especially for the populous nations.

Figure 6.3: Average Error (Validation of Population) Over Time for (orange) all countries and (blue) top 50 populous countries.

Table 6.4: Validation Results: Age Distribution of Countries, for populations within the age range of 0 to 14 years.

	0 to 14 years
Average Error	2.91%
Standard Deviation	5.47%

6.4.3 Age Distribution Validation

Last but not least, we validated the age distribution of countries. As mentioned previously, we initialized the age distribution using actual data obtained from [76], and the age distribution is shifted as a confluence of migration, births and deaths in population. In this case, we validated against the proportion of population within the 0 to 14 years age group (data available on [80]). The average error is defined the same as the average error for country population, except that we now subset the population to those within the range of 0 to 14 years. Table 6.4 shows the results of validation. This shows that the model performs well in modeling the age distribution, given assumptions of shifts, migration and birth/death rates that we are introducing in the model.

6.5 Strengths and Limitations

In this section, we discuss some of the strengths and limitations of the model. Strengths of the model include:

- Modeling of network and ties amongst countries, ranging from economy to alliances to linguistic similarities. To the best of our knowledge, no previous work done in simulation of migration has considered such a wide range of country networks affecting migration.
- Simple linear decision model serves as a good starting point for rapid examination of factors affecting population movement, though this could be replaced with other models in future works.
- Age distribution added to allow policy makers to identify shifts in age distributions of countries as result of fertility, mortality, as well as international migration.

Limitations of the model include:

- Data for the less developed nations might suffer from poor data quality [64]. Thus this might affect the results obtained for those countries involved.
- The migration decision model currently does not consider major events, such as conflicts or war within or amongst countries, which could serve as a driver of migration out of a country.
- The model is more suited for short-term simulation of possible shifts in population instead of simulating over longer period of time. For longer term predictions, it could suffer from poorer accuracies as a

result of compounding of errors over time [64].

6.6 Conclusion

Agent-based models and simulation have increasingly become an important tool for explaining and generating human behaviors. In the areas of human migration, the agent-based paradigm has been used successfully to explain migration movements as a result of climate, economic and peer effects. Our work follows these trends, with a focus on international migration, as we attempt to explain migration and population shifts as a confluence of country networks, along with actual fertility and mortality rates.

The first major contribution of this chapter would be the development of country networks that were incorporated into the agent's migration decision model, ranging from alliances to linguistic similarity to economic disparity networks. To the best of our knowledge, previous works in migration do not consider such a wide range of country networks for the migration decision model. We believe that this would result in a more realistic model. The next major contribution would be incorporation of age distributions of countries (initialized with actual data), which is then shifted as a result of births, deaths and migration. The reason for incorporating this is two-pronged; not only does it allow policy makers to identify possible shifts in age distributions, it also allows for additional validation.

Migration is a complex decision which could incorporate various *hard-to-capture* components (e.g., emotional aspects). With the networks that were developed, as well as adoption of the agent-based paradigm, we provided a good explanation of country-level observations (population and age distribution) as a confluence of migration movements, as well as fertility and mortality. Improving the realism of the model remains one of the major research

directions in the future.

Chapter 7

Conclusion

In this chapter, we (1) summarize the contributions made in this dissertation, (2) highlight some research directions for future work, and (3) discussion of policy implications of each study.

7.1 Dissertation Summary

In summary, our work is mainly motivated by the need to develop *high-quality* agent-based simulation model that is able to mimic the movement dynamics of agents in a network environment. In this case, we do not look at development of high quality agent-based models as an end in itself. Rather, it's a *prerequisite* for subsequent usage of the agent-based model as an evaluation tool for various recommender systems and policies. To achieve this, we propose a methodological framework for modeling agents movement in a network environment, combining approaches such as data-driven modeling and discrete choice models. While the individual concepts are not new, we demonstrated through our framework a systematic approach for modeling under different problem domains as specified in figure 2.1. Additionally, we demonstrate the applicability of our proposed methodological framework in three concrete use cases, each representing a problem domain. This is the main contribution and novelty of our work - the combination

of various techniques in a micro (behavioral model) to macro (aggregate observations arising from agents decisions) multi-level modeling approach.

The first problem domain is described in chapter 4. In the leisure or theme park setting, we define the utility of going to an attraction based on that of the patron's preference and travel distance. Additionally, we included case studies to illustrate the use of agent-based models for evaluating (1) a dynamic route guidance application and (2) effect of spatial layout on crowding. With a model that is able to capture the aggregated movement dynamics through the use of a small sample data, this would allow policy makers to better come up with various crowd management strategies, through the use of such computational models as an evaluation testbed.

The second problem domain is described in chapter 5. In this case, we are faced with the problem of a huge amount of "noisy" data. Noise in this case refers to (i) inaccuracies in GPS readings of taxis and (ii) figuring out the true intention of drivers when making zone-level decisions. To deal with this, we first utilize MapMatching to map the raw lat-lon readings from GPS records to links via a viterbi-based algorithm. From there, we then define a minimum traversing time for traversing a zone, and filter out records where taxis are passing through "transient zones". A utility model for zone level movement is developed considering historical demand at zones, along with distance between zones. Our virtual experiments show that such an approach performs better than a traditional transition-matrix based (defined as single-tier model) one. We then utilized the agent-based model as an evaluative tool for a DGS system, where we looked into the performance of the DGS at (1) various times of the day and (2) under different penetration ratio of guided visitors. The model could be effectively utilized by policy makers to test out various supply and demand-side regulation policies.

The third problem domain is described in chapter 6. In this case, the time interval granularity is much coarse (in years) as opposed to the previous two areas, and we have a much more complex decision making process requiring far more contemplation. In this case, we define a utility model for migration that considers a wide range of networks, ranging from migrant to economic disparity. The model could be effectively utilized by policy makers looking into immigration and population projection, especially on impacts of such policies on various distributions, e.g., age distribution.

7.2 Future Work

We conclude this dissertation by highlighting several research directions that would further improve the current work. In chapter 4, we included a simple case study to demonstrate the usage of an agent-based model in evaluating the effect of spatial layout, in particular the placement of popular attractions proximity to the theme park entrance. For future work, it would be interesting to explore other findings (e.g., in the work by [82]), or a combination thereof.

Also, in chapter 5, we showed the usefulness of a tiered decision model in modelling the movement behaviors of taxi drivers in the roaming mode (i.e., in search of passengers). For future work, we look into hierarchies in terms of various levels of thinking and reasoning sophistication, instead of a zone-link level. Example of such a model would be the Cognitive Hierarchy Model by [14]. Given that we do not expect drivers to think at the same level of sophistication (as observed by differing revenues by different drivers), it would definitely be interesting to test this out on the existing data and model.

Bibliography

[1] Kau Ah-Keng. Evaluating the attractiveness of a new theme park: A cross-cultural comparison. *Tourism Management*, 14(3):202–210, 1993.

[2] Baer, D. 'the biggest change of our time' is happening right now in africa. Retrieved from: `http://www.techinsider.io/africas-population-explosion-will-change-humanity-2015-8`, 2015.

[3] Hugo S Barbosa, Fernando B de Lima Neto, and Wilson Fusco. Migration and social networks—an explanatory multi-evolutionary agent-based model. In *Intelligent Agent (IA), 2011 IEEE Symposium on*, pages 1–7. IEEE, 2011.

[4] Stuart J Barnes, Jan Mattsson, and Flemming Sørensen. Remembered experiences and revisit intentions: A longitudinal study of safari park visitors. *Tourism Management*, 57:286–294, 2016.

[5] Ana LC Bazzan. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems*, 18(3):342, 2009.

[6] Michael Behrisch, Laura Bieker, Jakob Erdmann, and Daniel Krajzewicz. SUMO–simulation of urban mobility: an overview. In *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. ThinkMind, 2011.

[7] Richard Bellman. *Dynamic programming*. Courier Corporation, 2013.

[8] Moshe Ben-Akiva, Michel Bierlaire, Haris Koutsopoulos, and Rabi Mishalani. Dynamit: A simulation-based system for traffic prediction. In *Proceedings of the DACCORD Short-Term Forecasting Workshop*, 1998.

[9] J Enrique Bigné, Luisa Andreu, and Juergen Gnoth. The theme park experience: An analysis of pleasure, arousal and satisfaction. *Tourism management*, 26(6):833–844, 2005.

[10] Joschka Bischoff and Michal Maciejewski. Agent-based simulation of electric taxicab fleets. *Transportation Research Procedia*, 4:191–198, 2014.

[11] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287, 2002.

[12] Amber Brown, Jacqueline Kappes, and Joe Marks. Mitigating theme park crowding with incentives and information on mobile devices. *Journal of Travel Research*, 52(4):426–436, 2013.

[13] Chris Buckley. China ends one-child policy, allowing families two children. The New York Times. Retrieved from: `http://www.nytimes.com/2015/10/30/world/asia/china-end-one-child-policy.html?_r=1`, 2015.

[14] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, pages 861–898, 2004.

[15] CEDS. Computational event data system: Cameo event data codebook. Retrieved from: `http://eventdata.parusanalytics.com/data.dir/cameo.html`, 2012.

[16] Shih-Fen Cheng, Larry Lin, Jiali Du, Hoong Chuin Lau, and Pradeep Varakantham. An agent-based simulation approach to experience management in theme parks. In *Simulation Conference (WSC), 2013 Winter*, pages 1527–1538. IEEE, 2013.

[17] Shih-Fen Cheng and Thi Duong Nguyen. Taxisim: A multiagent simulation platform for evaluating taxi fleet operations. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02*, pages 14–21. IEEE Computer Society, 2011.

[18] Suh-Wen Chiou. Optimization of area traffic control for equilibrium network flows. *Transportation Science*, 33(3):279–289, 1999.

[19] DARPA. Accelerating discovery with new tools and methods for next generation social science. `https://www.darpa.mil/news-events/2016-03-04`, 2016.

[20] DARPA. Next Generation Social Science. `https://www.darpa.mil/program/next-generation-social-science`, 2016.

[21] DARPA. Ground Truth. `https://www.darpa.mil/program/ground-truth`, 2017.

[22] DARPA. Putting social science modeling through its paces. `https://www.darpa.mil/news-events/2017-04-07`, 2017.

[23] Defense Advanced Research Projects Agency. Accelerating discovery with new tools and methods for next generation social science. Retrieved from: `https://www.darpa.mil/news-events/2016-03-04`, 2016.

[24] Defense Advanced Research Projects Agency. Putting social science modeling through its paces. Retrieved from: `https://www.darpa.mil/news-events/2017-04-07`, 2017.

[25] Adam Dennett. Estimating flows between geographical locations: 'get me started in' spatial interaction modelling. Technical report, Citeseer, 2012.

[26] Ping Dong and Noel Yee-Man Siu. Servicescape elements, customer predispositions and service experience: The case of theme park visitors. *Tourism Management*, 36:541–551, 2013.

[27] Jiali Du, Akshat Kumar, and Pradeep Varakantham. On understanding diffusion dynamics of patrons at a theme park. In *Extended abstract in the Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-14)*, 2014.

[28] Joshua M Epstein. *Generative social science: Studies in agent-based computational modeling*. Princeton University Press, 2006.

[29] Joshua M Epstein. *Agent_Zero: Toward neurocognitive foundations for generative social science*. Princeton University Press, 2014.

[30] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

[31] Martin Gardner. Mathematical games: The fantastic combinations of john conway's new solitaire game "life". *Scientific American*, 223(4):120–123, 1970.

[32] G Nigel Gilbert. *Agent-Based Models*. Number 153. Sage, 2008.

[33] Grant, Mark. Us recession looms – and here's why. Retrieved from: `http://www.cnbc.com/2016/02/04/mark-grant-us-recession-looms-and-heres-why.html`, 2016.

[34] GraphHopper. Map Matching Algorithm based on GraphHopper. `https://github.com/graphhopper/map-matching`, 2009.

[35] Josep Maria Salanova Grau and Miquel Angel Estrada Romeu. Agent based modelling for simulating taxi services. *Procedia Computer Science*, 52:902–907, 2015.

[36] Josep Maria Salanova Grau, Miquel Angel Estrada Romeu, Evangelos Mitsakis, and Iraklis Stamos. Agent based modeling for simulation of taxi services. *Journal of Traffic and Logistics Engineering*, 1, 2013.

[37] Index Mundi. Historical data graphs. Retrieved from: `http://www.indexmundi.com/g/`, 2015.

[38] Shashi Shekhar Jha, Shih-Fen Cheng, Meghna Lowalekar, Wai Hin Wong, Rajendram Rishikeshan Rajendram, Trong Khiem Tran, Pradeep Varakantham, Nghia Truong Trong, and Firmansyah Rahman. Upping the game of taxi driving in the age of Uber. 2018.

[39] Astrid Kemperman, Aloys Borgers, Harmen Oppewal, and Harry Timmermans. Predicting the duration of theme park visitors' activities: An ordered logit model using conjoint choice data. *Journal of Travel Research*, 41(4):375–384, 2003.

[40] John Kennan, James R Walker, et al. Modeling individual migration decisions. *International Handbook on the Economics of Migration, Edward Elgar, Cheltenham, UK, and Northampton, USA*, pages 39–54, 2013.

[41] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of SUMO-Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements*, 5(3&4):128–138, 2012.

[42] Jean-Jacques Laffont and David Martimort. *The theory of incentives: the principal-agent model*. Princeton university press, 2009.

[43] Hoong Chuin Lau, William Yeoh, Pradeep Varakantham, Duc Thien Nguyen, and Huaxing Chen. Dynamic stochastic orienteering problems for risk-aware applications. *UAI 2012: Proceedings of Conf. on Uncertainty in AI*, 2012.

[44] Kang Hoon Lee, Myung Geol Choi, Qyoun Hong, and Jehee Lee. Group behavior from video: a data-driven approach to crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 109–118. Eurographics Association, 2007.

[45] Larry Lin, Shih-Fen Cheng, and Hoong Chuin Lau. Building crowd movement model using sample-based mobility survey. In *Proceedings of the 2015 Winter Simulation Conference*, pages 139–150. IEEE Press, 2015.

[46] Line, B. and Poon, L. How other countries handle immigration. National Geographic. Retrieved from: http://news.nationalgeographic.com/news/2013/06/130630-immigration-reform-world-refugees-asylum-canada-japan-australia-sweden-denmark-united-kingdom-undocumented-immigrants/, 2013.

[47] Meghna Lowalekar, Pradeep Varakantham, Supriyo Ghosh, Sanjay Dominik Jena, and Patrick Jaillet. Online Repositioning in Bike Sharing Systems. 2017.

[48] Meghna Lowalekar, Pradeep Varakantham, and Patrick Jaillet. Online Spatio-Temporal Matching in Stochastic and Dynamic Domains. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI-16*, 2016.

[49] LTA. Singapore Land Transport Statistics in Brief 2015. 2015.

[50] Michal Maciejewski and Joschka Bischoff. Large-scale microscopic simulation of taxi services. *Procedia Computer Science*, 52:358–364, 2015.

[51] Luis M Martinez, Gonçalo HA Correia, and José M Viegas. An agent-based simulation model to assess the impacts of introducing a shared-taxi system: an application to Lisbon (Portugal). *Journal of Advanced Transportation*, 49(3):475–495, 2015.

[52] Luis M Martinez and José Manuel Viegas. Assessing the impacts of deploying a shared self-driving urban mobility system: An agent-based model applied to the city of Lisbon, Portugal. *International Journal of Transportation Science and Technology*, 6(1):13–27, 2017.

[53] Gordon W McClung. Theme park selection: Factors influencing attendance. *Tourism Management*, 12(2):132–140, 1991.

[54] Roland Mielke, Adham Zahralddin, Damanjit Padam, and Thomas Mastaglio. Simulation applied to theme park management. In *Simulation Conference Proceedings, 1998. Winter*, volume 2, pages 1199–1203. IEEE, 1998.

[55] Kazuo Miyashita. Asap: Agent-based simulator for amusement park. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 195–209. Springer, 2004.

[56] Luiz Moutinho. Amusement park visitor behaviour—scottish attitudes. *Tourism management*, 9(4):291–300, 1988.

[57] Paul Newson and John Krumm. Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 336–343. ACM, 2009.

[58] Sarah Nicholls, Bas Amelung, and Jillian Student. Agent-based modeling: A powerful tool for tourism researchers. *Journal of Travel Research*, 56(1):3–15, 2017.

[59] OECD. Is migration good for the economy? Migration Policy Debates. Retrieved from: `https://www.oecd.org/migration/OECD%20Migration%20Policy%20Debates%20Numero%202.pdf`, 2014.

[60] Johan Janson Olstam, Jan Lundgren, Mikael Adlers, and Pontus Matstoms. A framework for simulation of surrounding vehicles in driving simulators. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 18(3):9, 2008.

[61] OpenStreetMap. Open Street Map. `https://www.openstreetmap.org`, 2015.

[62] Xiaoshan Pan, Charles S Han, and Kincho H Law. A multi-agent based simulation framework for the study of human and social behavior in egress analysis. In *Proceedings of the ASCE International Conference on Computing in Civil Engineering*, volume 92, 2005.

[63] Pew Research Center. Changing patterns of global migration and remittances. Retrieved from: `http://www.pewsocialtrends.org/2013/12/17/changing-patterns-of-global-migration-and-remittances/`, 2013.

[64] Population Reference Bureau. Understanding and using population projections. Retrieved from: `http://www.prb.org/Publications/Reports/2001/UnderstandingandUsingPopulationProjections.aspx`, 2001.

[65] Neelam C Poudyal, Bamadev Paudel, and Michael A Tarrant. A time series analysis of the impact of recession on national park visitation in the united states. *Tourism Management*, 35:181–189, 2013.

[66] Bryan Raney, Nurhan Cetin, Andreas Völlmy, Milenko Vrtic, Kay Axhausen, and Kai Nagel. An agent-based microsimulation model of swiss travel: First results. *Networks and Spatial Economics*, 3(1):23–41, 2003.

[67] Thomas C Schelling. Models of segregation. *The American Economic Review*, 59(2):488–493, 1969.

[68] Christopher Smith, Sharon Wood, and Dominic Kniveton. Agent based modelling of migration decision-making. In *Proceedings of the European workshop on multi-agent systems (EUMAS-2010)*, 2010.

[69] Sharon V Thach and Catherine N Axinn. Patron assessments of amusement park attributes. *Journal of Travel Research*, 32(3):51–60, 1994.

[70] The GDELT Project. Global database of events, language, and tone (gdelt) : Data. Retrieved from: `http://www.gdeltproject.org/data.html`, 2014.

[71] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[72] Chieh-Yuan Tsai, Hui-Ting Chang, and Ren Jieh Kuo. An ant colony based optimization for rfid reader deployment in theme parks under service level consideration. *Tourism Management*, 58:1–14, 2017.

[73] Theodore Tsiligirides. Heuristic methods applied to orienteering. *Journal of the Operational Research Society*, 35(9):797–809, 1984.

[74] United Nations. International migration stock 2015. Retrieved from: http://www.un.org/en/development/desa/population/migration/data/estimates2/estimates15.shtml, 2015.

[75] URA. List of postal zone and districts. https://www.ura.gov.sg/realEstateIIWeb/resources/misc/list_of_postal_districts.htm.

[76] US Census Bureau. United states census bureau - migration/geographic mobility. Retrieved from: http://www.census.gov/hhes/migration/, 2016.

[77] Michael P Wellman. Putting the agent in agent-based modeling. *Autonomous Agents and Multi-Agent Systems*, 30(6):1175–1189, 2016.

[78] RCP Wong, WY Szeto, and SC Wong. Bi-level decisions of vacant taxi drivers traveling towards taxi stands in customer-search: Modeling methodology and policy implications. *Transport Policy*, 33:73–81, 2014.

[79] RCP Wong, WY Szeto, and SC Wong. A two-stage approach to modeling vacant taxi movements. *Transportation Research Procedia*, 7:254–275, 2015.

[80] World Bank. Open data. Retrieved from: http://data.worldbank.org/, 2016.

[81] Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford, and Kiran Lakkaraju. Data-driven agent-based modeling, with application to rooftop solar adoption. *Autonomous Agents and Multi-Agent Systems*, 30(6):1023–1049, 2016.

[82] Yingsha Zhang, Xiang Robert Li, and Qin Su. Does spatial layout matter to theme park tourism carrying capacity? *Tourism Management*, 61:82–95, 2017.

[83] Yingsha Zhang, Xiang Robert Li, Qin Su, and Xingbao Hu. Exploring a theme park's tourism carrying capacity: A demand-side analysis. *Tourism Management*, 59:564–578, 2017.