A National Hybrid Activity/Agent-Based Demand Model to Characterize the Mobility of the United States

Author: Kyle MAROCCHINI
Supervisor: Professor Alain KORNHAUSER

Independent work submitted in fulfillment of the requirements for the degree of Bachelor’s of Science and Engineering in the department of Operations Research and Financial Engineering

January 11, 2017
Abstract

A National Hybrid Activity/Agent-Based Demand Model to Characterize the Mobility of the United States

by Kyle MAROCCHINI

The rise of emerging passenger-oriented mobility services, ranging from public multi-modal transportation systems to privately operated fleets of driverless vehicles, have the potential to completely revolutionize the current state of transportation throughout the world. To effectively and efficiently develop and implement these systems necessitates understanding the demand for transportation from both spatial and temporal lenses. This paper presents a new activity-based model which generates a synthetic population complete with transportation needs, both spatial and temporal, disaggregated down to the individual level to provide a complete view of what a population’s daily transportation needs might look like on a typical day. The demand model presented improves upon earlier work done by Talal Mufti [1] in 2012, Jingkang Gao [2] in 2013, Chris Brownell [3] and Hill Wyrough [4] in 2014 and Matthew Garvey [5] in 2015. This work seeks to synthesize and expand upon the various concepts and methods presented by the aforementioned authors in one centralized paper. The model presented in this paper expands the scope of the population generated from New Jersey to the entire United States, producing a synthetic population of 308.7 million persons with total transportation demand spanning over 1 billion trips in a typical workday.
Acknowledgements

I would like to thank my advisor, Professor Alain Kornhauser, for his support and guidance throughout the semester. Professor Kornhauser provided enormous assistance in helping troubleshoot errors in the model’s implementation as well as suggesting alternative algorithms to assist in developing Module 7 and reducing overall Module runtime.

I would also like to thank Talal Mufti, Jingkang Gao, Chris Brownell, Penn Wyrough and Matthew Garvey whose work paved the way for this model’s current iteration. Special thanks also goes to ORF467 Fall 2016, who tirelessly unit-tested the travel demand generated and suggested additional improvements, as well as all previous ORF467 classes, who have continually assisted in refining the model.

This paper represents my own work in accordance with University Regulations.

Kyle Marocchini
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>ii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The Transportation Problem and Demand Modelling</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Methodologies for Travel Demand Modelling</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Trip-Based Modelling</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 Activity-Based Modelling</td>
<td>3</td>
</tr>
<tr>
<td>1.2.3 Agent-Based Modelling</td>
<td>4</td>
</tr>
<tr>
<td>1.2.4 Synthesizing the Three Models</td>
<td>4</td>
</tr>
<tr>
<td>1.3 The Temporal Unit - the US Workday</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Goals</td>
<td>5</td>
</tr>
<tr>
<td>1.4.1 Improvements over Prior Work</td>
<td>6</td>
</tr>
<tr>
<td>2 Methodology</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Objective and Motivation</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Module 1: Generation of a Synthetic Populace</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Module 2: Assignment of Workplace</td>
<td>11</td>
</tr>
<tr>
<td>2.4 Module 3: Assignment of School</td>
<td>14</td>
</tr>
<tr>
<td>2.5 Module 4: Assignment of Activity Patterns</td>
<td>16</td>
</tr>
<tr>
<td>2.6 Module 5: Assignment of Trip Destination</td>
<td>18</td>
</tr>
<tr>
<td>2.6.1 Submodule 5.5: Assignment of Trip Destination</td>
<td>19</td>
</tr>
<tr>
<td>2.7 Module 6: Assignment of Arrival and Departure Times</td>
<td>22</td>
</tr>
<tr>
<td>2.8 Module 7: Pixelization of Demand</td>
<td>23</td>
</tr>
<tr>
<td>2.8.1 Autonomous Taxi Networks</td>
<td>24</td>
</tr>
<tr>
<td>2.8.2 Pixelization of the United States</td>
<td>26</td>
</tr>
</tbody>
</table>
2.9 Fundamental Assumptions ................................. 28
  2.9.1 Geographical Assumptions ......................... 29
  2.9.2 Methodological Assumptions ....................... 30

Module 1 .................................................. 30
Module 2 .................................................. 31
Module 3 .................................................. 31
Module 4 .................................................. 31
Module 5 .................................................. 32
Module 6 .................................................. 32
Module 7 .................................................. 33

3 Data ....................................................... 34
  3.1 Module 1 ................................................ 34
    3.1.1 2010 United States Census Summary File 1 Data . 34
    3.1.2 2010 American Community Survey of Household Income 35
  3.2 Module 2 ................................................ 35
    3.2.1 2010 American Community Survey Journey-to-Work Census . 36
    3.2.2 2010 American Community Survey Industry Type Participation by Gender and Median Income . 36
    3.2.3 Employee and Patronage Data .................... 36
  3.3 Module 3 ................................................ 37
    3.3.1 National Center for Education Statistics .......... 37
    3.3.2 Post-Secondary School Data ..................... 38
  3.4 Module 4 ................................................ 38
  3.5 Module 5 ................................................ 39
    3.5.1 Federal Aviation Administration Facilities Data .. 39
  3.6 Module 6 ................................................ 39
  3.7 Module 7 ................................................ 39

4 Conclusions ............................................... 42
  4.1 Data Applicability ................................... 42
  4.2 Run Time ............................................ 44
  4.3 Memory .............................................. 46
4.4 Statewide Interdependency .............................................. 46
4.5 Next Steps ................................................................. 47
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>TT, HHT and Income Bracket Codes.</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Example output for Module 1.</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>Example output for Module 2.</td>
<td>14</td>
</tr>
<tr>
<td>2.4</td>
<td>Example output for Module 3.</td>
<td>16</td>
</tr>
<tr>
<td>2.5</td>
<td>Example output for Module 5.5.</td>
<td>19</td>
</tr>
<tr>
<td>2.6</td>
<td>Example output for Submodule 5.5.</td>
<td>21</td>
</tr>
<tr>
<td>2.7</td>
<td>Example output for Module 6.</td>
<td>23</td>
</tr>
<tr>
<td>2.8</td>
<td>A Comparison of two ATN Models.</td>
<td>26</td>
</tr>
<tr>
<td>2.9</td>
<td>Pixelization of Princeton, NJ.</td>
<td>28</td>
</tr>
<tr>
<td>2.10</td>
<td>Example output for Module 7.</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>NHTS Empirical Distribution of Daily Trips by Start Time.</td>
<td>40</td>
</tr>
<tr>
<td>3.2</td>
<td>Trip Tour Distributions.</td>
<td>41</td>
</tr>
<tr>
<td>3.3</td>
<td>Bell Times and Durations by NAISC Industry.</td>
<td>41</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Trip Tours used in Module 4 ............................................. 17

3.1 NAICS Industry Codes Used ............................................. 37
Chapter 1

Introduction

1.1 The Transportation Problem and Demand Modelling

Solving the problem of transportation is key to virtually every sector of the economy. Organizations and individuals are consistently faced with the dilemma of choosing how to quickly and cheaply travel from one location to another. The rise of “mobility as a service” solutions, particularly in the field of privately owned networks of autonomous vehicles, has the potential to become a revolutionary mode of transportation, one that may even rise to compete with the personally-owned automobile [3]. However, to implement any form of transportation network, it is critical for the planner to first understand the spatial and temporal distributions of demand for their mobility service in the existing transportation network. By knowing these distributions for a given population down to the individual level, it becomes far easier to craft operational structures and services that satisfy said population’s demands as the planner knows both where and when to route vehicles in the network. To accomplish this task, it is thus necessary to first construct these desired mobility distributions for a given population.

To construct these distributions is a nontrivial matter, as the spatial and temporal traits that characterize a given population include a variety of socioeconomic, demographic, familial and land use factors  at “a level of detail which no survey can provide” [3]. This paper presents a new model that provides transportation demand disaggregation down to the individual level within the context of a typical workday throughout the entire United States, expanding upon earlier work previously restricted to the state of New Jersey by Talal Mufti, Jinkgang Gao and Chris
1.2 Methodologies for Travel Demand Modelling

Travel demand modelling is an essential tool that allows planners to make more informed decisions on the ways in which network infrastructure and policy will impact existing and future travel behavior. The overarching goal of travel demand modelling is to “analyze the response of users to changes brought about by new services, [infrastructural] investments and changes in operating and pricing policies” as to more accurately forecast changes in travel demand. Initially, forecasting models were used to assist in analyzing the placement of federally funded transportation infrastructures such as highways and rail systems. However, as processing power has increased and cheaper memory has become available, modelling is used more prevalently as a tool for analyzing complex, multi-model transportation systems, especially in urban environments. The methodologies used in travel demand modelling reflect this evolution over time.

1.2.1 Trip-Based Modelling

The earliest transportation models were based off of Lowry’s Model of Metropolis, which established the first standardized model in the field of demand modelling: the Trip-Based approach. The Trip-Based methodology was structured around a traditional “four-step model” encompassing trip generation, trip distribution, modal split and network assignment respectively. Spatial distributions are accounted for by partitioning geographic regions into Traffic Analysis Zones (TAZs) according to their land use. For example, a given county might be partitioned into commercial, residential, industrial and recreational TAZs.

The Trip-Based models provided sufficient information to examine the impacts of broad land use policies and aggregated regional travel demand. While this simple model provided clear advantages in implementation and aggregated travel flows, the methodology had significant shortcomings. Trip-Based models provided output that “lacked richness of detail” and ignored fundamental travel assumptions. Correlations between "home-based and non-home based trips" are ignored. The
scheduling of trips also plays no role in Trip-Based models: single-stop home-based trips and multi-stop trips are not distinct [10].

Ultimately, Trip-Based models fail to provide enough detail to distinguish individual trips and, more generally, individual travel behaviors. The spatial, temporal and functional interdependence of individual trips must be considered to develop accurate autonomousTaxi (aTaxi) systems and analyze their implications on travel demand.

1.2.2 Activity-Based Modelling

More modern transportation models are based off of the Activity-Based approach, which realizes transportation as a demand that "derives from people’s needs and desires to participate in activities" [6] and as such routes transportation based off of individual’s needs, as opposed to landuse aggregation. In these models, trips are defined as a singular movement of a person from an origin to a destination, regardless of mode. Tours are defined as a sequence of temporally consecutive trips that encompass a person’s travel demand throughout a unit of time. These models fundamentally assume a core set of trips, which factor in predictable anchors in time and space [9], such as home, work and school in combination with other locations. With the demand and supply of travel demand determined, the model then attempts to sequence them to construct a tour which reflects that person’s personal attributes, including employer, income, familial relations and so on.

Activity-Based models offer significant improvements over Trip-Based models. While land-use was the primary determinant for travel demand in the Trip-Based modelling framework, Activity-Based modelling recognizes the complex reality of personal travel and instead uses personal need. While some assumptions must be made regarding travel behavior (e.g., the partitioning of travel demand into a set of pre-determined trip types) in order to achieve consistency, these models allow for multi-modal integration and the inclusion of large number of variables to define personal need. CEMDAP [11] and ALBATROSS [12] are recommend for further reference on Activity-Based models.
1.2.3 Agent-Based Modelling

Agent-Based modelling seeks to overcome the aggregation presented in Trip and Activity-Based models by using the individual as the central motivating factor. These models “incorporate the complexity of human behavior using ‘agents’ that are autonomous and interactive...in nature” to simulate their decisions over time [13]. These models adopt the same scheduling approach as the Activity-Based model and substitute population subsets with the individual as the unit of analysis. While the focus on the individual allows for more detailed analysis of travel demand as a function of personal characteristics, the model requires highly detailed estimates of individual-individual covariance in order to provide accurate descriptions of travel demand.

1.2.4 Synthesizing the Three Models

Each modelling framework presented has strengths and weaknesses that must be carefully considered when determining the best modelling framework for an application. The travel demand model used for aTaxi simulation needs to:

- Provide comprehensive disaggregation that accounts for dependence on spatial, temporal and individual-level personal factors.
- Generate trips on a schedule-based framework in a consistent manner that allows for spatial analysis.
- Minimize assumptions, when possible. When not possible, provide for a framework that allows for informed assumptions to be made based on observed data on a scale that is comprehensive enough to be applied uniformly across the nation.

In order to achieve these three goals, a hybrid Activity-Agent based model is used. From the Activity-based model, a trip-scheduling framework is adopted at a spatial scope great enough to be applied uniformly across the nation. From the Agent-based model, a focus on individual level behavior is adopted. However, in order to avoid the issue of individual-individual covariance estimation, individuals are aggregated at the household level. All trip tours begin and end at home and residents are grouped by households.
1.3 The Temporal Unit - the US Workday

In order to provide a framework amendable to further temporal analysis, travel demand tours are generated over a given US workday, in line with other travel demand models. Weekdays have significant variability, so assigning a weekday will not create issues of limited analysis. However, there are two implicit assumptions that must be made if one is to use the workday. The first implicit assumption is that weekdays (or better said, each simulation from the model presented) are identically distributed. Secondly, travel on the weekend is also assumed to be nearly identical to the first. While the latter assumption can be made without much loss of generality, the former does not appear to be correct. To correct for this, we state that the workday simulation generated from the model is merely a foundation for the actual temporal distributions of travel demand. Additional frameworks must be integrated into the presented model to explain additional complexities (e.g. Monday/Friday commute patterns, multi-day trip tours, etc).

1.4 Goals

The goal of the model presented is to generate the precise origin, destination, and arrival/departure time for every trip made by every individual on a typical workday when school is in session. More generally, the model provides the spatial and temporal distribution for every citizen in the United States. Every individual simulation produces a unique trip file that contains an individualized, probabilistic record of every person-trip on an average weekday, which is expected to total to just over 1 billion trips. Each record includes every trip the person makes including spatial coordinates of the origins and destinations as well as the exact departure and arrival times in seconds after midnight. Files also contain pointers that indicate individual characteristics about a given person, including income, household location, familial status, etc.
1.4.1 Improvements over Prior Work

The national model provides several improvements over the older model. While the issue of scope is certainly an improvement, more relevant to the implementation of the model are the following.

- **Data:** While the New Jersey model was generated from various data sources and fine-tuned to correct for errors, the national model requires a far greater, more comprehensive and more uniform set of data. The robustness of data sources is also corrected by deferring to US Census Bureau data when applicable.

- **Modeling:** The New Jersey model was hard-coded for use in New Jersey. The national model generalizes the simulation so that every state can be run independently in the same manner. International travel and out-of-state travel is refined as well.

- **Implementation:** The implementation of the model was completely hard-coded for New Jersey with little generalization. Moreover, several simplifying assumptions were made regarding spatial and temporal distributions in order to assist in the implementation given the limited scope of the New Jersey data. The demand synthesizer implemented removes assumptions when possible and provides generalization to handle every state.
Chapter 2

Methodology

2.1 Objective and Motivation

The overall objective of the model presented in this paper, summarized briefly, is to generate a synthetic listing detailing the personal trips taken by all residents of the U.S. in order to be able to somewhat accurately simulate how well various operational implementations of aTaxis might serve such demand. The objective is to gain an understanding as to the size, scope and operational/management structures that would be needed to best serve the mobility needs of today’s population and land use and not to try to address how land use and mobility needs might evolve and converge in response to the availability of such a mobility system.

To obtain such a trip data set, one could observe each person’s travel tour on some representative day, say a Tuesday or Wednesday in October. However, as the U.S. Census shows, simply counting the number of person’s every 10 years is an extremely difficult undertaking. Observing and recording everyone’s travel tour on one day only adds further complexity. To overcome this, the model presented synthesizes each individual person tour using an iterative procedure, with each step constructively building on the output of previous steps to build said listing.

First, a simulated population of U.S. Residents is created. General attributes such as age and gender are used in conjunction with information on personal income and household residency to exactly mimic the U.S. Population as it was in 2010. The next step in the process takes in workforce participation, school enrollment, employer size, customer patronage rates and industry sector data to assign working
residents to workplaces and students to schools. Trips are constructed between different anchor points, defined as school, home, work and other activity, which are then sequentially ordered in tours. With the demand and supply for travel created, assumptions regarding travel behavior are used to match the two to generate a daily trip tour for every individual. With home, work and school anchor points determined, other type trip points are created which are then sequentially ordered in tours and given specific, to the second, departure and arrival times from start end and hours of service time distributions of the work, school and activity location for a given person. The four steps collectively create a comprehensive data set containing the temporal and spatial characteristics of each synthesized trip taken throughout an average work/school day in the U.S..

2.2 Module 1: Generation of a Synthetic Populace

To begin the process, it is first necessary to generate the entire population of the U.S., roughly spanning over 300 million people. At its core, this step constructs the U.S.’s travel supply and provides persons with personal attributes that inform and control their demand for transportation, a fundamental aspect of all Activity-Based methodologies. However, to maintain compliance with the equally important aspect of disaggregation, it is paramount to ensure that the personal attributes generated have a level of specificity that allows one to characterize individuals. In order to balance these two needs, the U.S. Census Bureau’s 2010 Census Block-Level Data was used to generate the population.

Census Blocks are the smallest geographic unit used by the Census Bureau for collection of non-sample data, are bound by streets and are usually populated with fewer than 100 people. While some inherent aggregation is assumed in relying on Census-Block level data, namely that every individual lives in the centroid of a census block and the distributions implied by the census data are correct, no other widely available data source currently exists that would allow disaggregation beyond the geographic size of a typical census block. 11,078,297 Census Blocks, covering all 50 States and the District of Columbia, are iterated through to construct the population in Module 1.
Module 1 begins by constructing residents with age and gender from Census data so that the age brackets and genders match the population presented in the Census data exactly. To ensure this is the case, ages and genders are assigned by sampling from the distribution of residents in each age bracket by gender without replacement. As Census data only provides ages to specific brackets (e.g. 5-7, 8-11, 9-13, etc.) selection is done by uniformly sampling within each age bracket. Each person is assigned a 10 digit Person ID Number. The first two numbers identify the state, while the remaining eight identify the person sequentially as they were generated.

While Mufti’s original implementation included sampling with replacement, this choice proves troublesome when constructing households. Consider a hypothetical census block with two male residents, one 45 years of age, the other eight years of age, both living in one house. Using sampling with replacement, it is statistically possible to generate two eight year old residents. In the U.S. Census, every house has a householder, defined as the ‘head of the house’, whom are distributed by gender and household type. In this hypothetical Census block, the single householder would have to be one of the two male eight year olds. This result would not only be unrealistic, it would also directly contradict the data, as the Census data defines all householders to be over 16 years of age. To ensure that the simulation presented was coherent with the data, it was decided to use sampling without replacement for age and gender.

The next step is to place each resident within a household. Households are identified and distributed by Household Type (HHT) in Census data, seen in Figure 2.1. Our model follows the designations provided by the data by maintaining familial (HHT 0) and non-familial (HHT 1) households, but aggregating the remaining Household Types as group quarter housing (HHT 2-8). Those living in group quarter housing are assigned to households by sampling from the newly constructed population, using Census distributions on group quarter occupancy by age bracket. The total population minus those assigned into group quarter housing are then placed into familial and non-familial households. To do so, the model leverages two columns within the U.S. Census Block Level data. The first details the occupancies of households within a census block according to their occupancies, and the second presents
the distribution of residents within familial and non-familial households. In the familial case, these distributions are based on the relations household occupants have with the householder.

The next step borrows on Mufti’s original concept of Traveler Types (TTs) and assigns residents a specific Traveler Type. TTs are derived from specific population attributes and are used later in the model to generate various trip tour patterns for residents when constructing trips. TTs encode assumptions on unemployment, leave and sick-days, as well as workforce data, to account for the correct number of employees that go to work on a given day. Figure 2.1 details how Age and HHT are used to assign traveler types to individuals. Age group percentages are given as well as HHT categories that assign an individuals to a TT. TTs 0-4 are based off of percentages in the Longitudinal Employer-Household Dynamics (LEHD) survey [14], while traveler types 5-6 are based off on an assumption of 10% unemployment, work-at-home rates of roughly 8% [14] and sick days at 4% [15]. Mufti’s framework is adopted with one slight modification; Out-Of-State employees are removed, as no such traveler exists in the current model (although employees living outside the U.S. do exist).

The final step in Module 1 is to ascribe an income to every individual. To do so,

<table>
<thead>
<tr>
<th>Traveler Type Code</th>
<th>Traveler Type Name</th>
<th>Classification (Age; HHT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Does Not Travel</td>
<td>0-5, 7, 9+; 2, 3, 4, 5, 7</td>
</tr>
<tr>
<td>1</td>
<td>Student Non-Worker</td>
<td>5-15, 16-18 * 99.81%</td>
</tr>
<tr>
<td>2</td>
<td>Student Worker In County</td>
<td>16-18 * 0.193%</td>
</tr>
<tr>
<td>3</td>
<td>College No Commute</td>
<td>18-22 * 90.34%</td>
</tr>
<tr>
<td>4</td>
<td>College Worker In County</td>
<td>18-22 * 9.66%</td>
</tr>
<tr>
<td>5</td>
<td>Typical Traveler</td>
<td>22-64 * 78%</td>
</tr>
<tr>
<td>6</td>
<td>Home Worker Traveler</td>
<td>22-64 * 22%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HHT Code</th>
<th>HHT Name</th>
<th>Income Code</th>
<th>Income Bracket</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Family</td>
<td>0</td>
<td>&lt; $10,000</td>
</tr>
<tr>
<td>1</td>
<td>Non-Family</td>
<td>1</td>
<td>$10,000-14,999</td>
</tr>
<tr>
<td>2</td>
<td>Correctional Facility</td>
<td>2</td>
<td>$15,000-24,999</td>
</tr>
<tr>
<td>3</td>
<td>Juvenile Detention Center</td>
<td>3</td>
<td>$25,000-34,999</td>
</tr>
<tr>
<td>4</td>
<td>Nursing Home</td>
<td>4</td>
<td>$35,000-49,999</td>
</tr>
<tr>
<td>5</td>
<td>Other Institutional Quarters</td>
<td>5</td>
<td>$50,000-74,999</td>
</tr>
<tr>
<td>6</td>
<td>Dormitories</td>
<td>6</td>
<td>$75,000-99,999</td>
</tr>
<tr>
<td>7</td>
<td>Military Quarters</td>
<td>7</td>
<td>$100,000-149,999</td>
</tr>
<tr>
<td>8</td>
<td>Other Non-Institutionalized Quarters</td>
<td>8</td>
<td>$150,000-199,999</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>9</td>
<td>&gt; $200,000</td>
</tr>
</tbody>
</table>

**Figure 2.1:** Traveler Type, Household Type and Income Bracket Codes.
Chapter 2. Methodology

every household, familial or non-familial, is assigned an income and income code as in Figure 2.1 based on distributions of household income from the American Community Survey [15]. Each household’s income is then randomly distributed to the household members provided they should earn an income.

Module 1 produces 51 master files containing all of the residents and their personal attributes for each state, including the District of Columbia, where residents are grouped together by block and by household. These files are the foundation for the next tasks that assign each resident to their daily activities, such as work, school, and other recreational trips. The files also serve as a resident directory of every individual generated in the trip and are used for verifying distributions as well as analyzing results. An example of the output from Module 1 for Wyoming can be found in Figure 2.2.

2.3 Module 2: Assignment of Workplace

Module 2 begins the process of determining travel person-trips by determining where eligible residents work. This is done by sequentially determining an employee’s county of work, industry of work and place of work. In the Module, a departure is made in scope. Instead of relying on Block-level data, the model now refers to County-level data as input. As data often does not exist for business and school activities at any level lower than the county, this change was a necessary one. To accommodate this change, Federal Information Processing Standards (FIPS) county codes - 5 digit numbers that uniquely identify all 3,142 counties or county equivalents in the United States [16] - are used to identify specific counties and county-level
Chapter 2. Methodology

data throughout the model, as they permit easy searching and sorting of data.

Module 2 begins by ascertaining which residents are eligible to work. Only Traveler Types 2 (School-Work), 4 (College-Work) and 5 (Typical Traveler) are valid workers. Once this is completed, valid workers are assigned specific counties of work using the Journey-to-Work (JtW) Census [17] which details the flow of workers from their home county to their workplace county. For every county, a weighted distribution is created using a Gravity Model that quantitatively measures the degree of attraction of a work county given the residence county. The Gravity Model, much like its Newtonian counterpart in physics, is a way to ascribe the ‘gravitational’ pull to a location, which increases with the popularity of the destination and declines with distance squared. The model in Equation 2.1 is responsible for generating the weights of a cumulative distribution function (CDF) from which a random work county is drawn. For all qualified workers, Module 2 uses this distribution to assign each worker a county of work.

\[ w_{h,c} = \frac{x_{h,c}}{\sum_{j} x_{h,j} D_{h,j}^2} \quad \forall \ c \in C \] (2.1)

where:

- \( h \) = the index of the worker’s home county
- \( C \) = the set of commutable counties for workers in county \( h \)
- \( x_{h,c} \) = the number of works commuting from county \( h \) to \( c \)
- \( D_{h,c} \) = the distance between the centroids of counties \( h \) and \( c \)
- \( w_{h,c} \) = the weight of attraction for a worker commuting from county \( h \) to \( c \)

With the worker’s county assigned, the next step in the process is to assign the work to an industry sector. The North American Industry Classification System (NAICS) codes used by the U.S. Census Bureau are also used here to categorize all possible industries. To designate a worker’s industry, the worker’s income, gender and work county, as well as the distribution of employment by gender and industry within the worker’s work county, are used. The U.S. Census Bureau Industry by Sex and Median Earnings [18] survey, which relates industry participation for every county in the United States by gender and median income, provides the data necessary to assign a worker to an industry. As with Equation 2.1 a Gravity Model is
employed to relate a worker to the industries within their work county. Attractiveness weights are used to select an industry as a function of a worker’s income and gender. The mechanics used are described in Equation 2.2.

$$w_{i,k} = \frac{e_{i,k}}{\sum_k e_{i,k} \frac{(I - \text{med}(I)_k)^2}{(I - \text{med}(I)_k)^2}} \forall \ k \in K$$

(2.2)

where:

- $i$ = the index of the worker’s home county
- $I$ = the worker’s income
- $K$ = the set of NAISC industries, indexed by $k$
- $\text{med}(I)_k$ = the median income of workers in industry $k$ in county $i$ of the same gender as the worker
- $e_{i,k}$ = the number of workers in industry $k$ in county $i$ of the same gender as the worker
- $w_{i,k}$ = the weight of attraction for a worker to industry $k$

With the worker’s county and industry assigned, all that remains is to assign the worker a specific employer and more importantly, a workplace. To assign a worker a workplace a dataset describing every employer’s physical workplace throughout the United States partitioned by county and by NAISC industry was used. Once more, a Gravity Model was employed to assign a worker to a workplace. The model in Equation 2.3 is responsible for generating the weights of a CDF from which an employer is randomly drawn.

$$w_{h,j} = \frac{e_{h,k,j} \frac{D_{h,j}}{D_{h,j}}}{\sum_j e_{h,k,j} \frac{D_{h,j}}{D_{h,j}}} \forall \ j \in J$$

(2.3)

where:
\( i \) = the index of the worker’s home county

\( k \) = the index of the worker’s industry from the set of NAISC Industries \( K \)

\( J \) = the set of workplaces of industry \( k \) and within county \( i \), indexed by \( j \)

\( e_{i,k,j} \) = the number of workers in county \( i \) of industry \( k \) at workplace \( j \)

\( D_{h,j} \) = the distance between the home of the worker (defined as the centroid of the worker’s census block, \( h \)) and workplace \( j \)

\( w_{h,j} \) = the weight of attraction for a worker to workplace \( j \)

With employees assigned (or unassigned in the case of non-workers) to a workplace, Module 2 produces a revised file of residents that provides additional information about workplaces. A sample output of one of the files generated for the state of Delaware is seen in Figure 2.3. Only information added by Module 2 is shown in Figure 2.3 as the output from this Module is appended to the work from earlier modules. Workplaces form a crucial component of the trip tours taken by most U.S. residents throughout the workday. The next Module, Module 3, continues this process by determining the final crucial anchor, schools, for all U.S. residents.

**Figure 2.3: Output from Module 2 for state of Delaware.**

### 2.4 Module 3: Assignment of School

Module 3 sees residents that are aged from 6 to 22 years of age and not residing in institutional group quarters assigned to appropriate schools. Parallels can be drawn between Module 2 and Module 3 as they work in a similar fashion. As in Module 2, a series of steps are taken to narrow down each student to their correct school. To begin, students are partitioned into kindergarten, elementary, primary, secondary or post-secondary school by age. Due to insufficient data on pre-kindergarten education, students younger than six are not assigned to a school. Students from K-12 are divided between public and private schools, leveraging a dataset provided by the
Chapter 2. Methodology

National Center for Education Statistics (NCES) [19] detailing participation in school type by age for all counties. Students in post-secondary schools are assigned to four-year institutions or graduate programs, two-year associate degree programs or non-degree granting programs. Deficiencies in the NCES dataset for post-secondary education resulted in a poor coverage of post-secondary programs. To overcome this issue, post-secondary schools were instead drawn from the Business datasets used in Module 2. To weight each school appropriately, the number of employees at each post-secondary institution were used as a substitute metric for student enrollment, which was not available for all schools.

With all students partitioned into education groups (and their respective institutions), the next step is to select an individual school for that student. Once more, a Gravity Model is used to build a distribution of the possible schools a student can attend. The mechanics of this model are summarized in Equation 2.4.

\[
w_{i,j} = \frac{x_j}{\sum_j x_j D_{i,c}^{-2}} \quad \forall \quad c \in C
\]

where:

- \(i\) = the index of the student’s home county
- \(k\) = the type of school attended by the student
- \(C\) = the set of counties adjacent to county \(h\), indexed by \(c\)
- \(J\) = the set of schools of type \(k\) in county \(c\), indexed by \(j\)
- \(x_j\) = the enrollment of students in school \(j\)
- \(D_{i,c}\) = \[
\begin{cases} 
D_{i,c}, & \text{for } c \neq i \\
0.75(\min_c(D_{i,c})), & \text{for } c = i
\end{cases}
\]
  where \(D_{i,c}\) denotes the distance between the centroids of counties \(i\) and \(c\)
- \(w_{i,j}\) = the weight of attraction to school \(j\) for students in county \(i\)

One special caveat encoded in Equation is the spatial adjacency requirement. This simplicity was introduced as for most schools, this assumption on spatial proximity generally holds. For students in post-secondary schools, many often reside in dorms on campus and for these students, it is almost guaranteed that if they reside in dorms, their school should be within their home county or the adjacent county. For public schools from K-12, students are highly likely to attend school in their
Chapter 2. Methodology

2.4 Commonly, this statement may not hold as strongly for those in K-12 private schools, the geographic range encompassed by a home county and all its adjacent counties is large enough to reasonably assume this holds as well.

With students assigned (or unassigned in the case of non-students) to a school, Module 3 produces a revised file of residents that provides additional information about school enrollment. A sample output of one of the files generated for the state of Alaska is seen in Figure 2.4. Schools form the last crucial component of the trip tours taken by most U.S. residents throughout the workday in combination with workplaces and home residencies. The three anchors of our travel demand have thus been completely spatially determined. It now remains to determine the type of trip tours that our residents engage in and to determine those trips that are not encompassed as one of these anchor points, defined earlier as Other type trips.

![Figure 2.4: Output from Module 3 for state of Alaska.](image)

2.5 Module 4: Assignment of Activity Patterns

Modules 4 and 5 collectively complete the spatial distribution of a person’s daily tour. Module 4 begins this process by assigning a particular activity pattern for each person. An activity pattern consists of a combination of four of the main trip types, namely Home (H), School (S), Work (W) and Other (O), which collectively define every possible daily tour that begins and ends at home. All 21 activity patterns, displayed in Table 2.1 are determined for a person by their traveler type. Traveler types also constrain the set of trips within each particular activity pattern. For example, any non-worker traveler type cannot have a W trip located within their activity pattern. The distribution of activity patterns amongst all traveler types are constructed to match the mean number of trips taken daily in the U.S., between three and four
Activity patterns are modified from Mufti’s original implementation as Mufti allowed for far more flexible activity patterns, e.g. non-workers that had W trips. It was decided that to be more consistent with the traveler type designations generated in Module 1 and more importantly, to be realistic, activity patterns would be constrained by traveler type. The activity patterns presented are augmented from Mufti’s original set through the inclusion of patterns that do not involve W or S trips. While the distributions of activity patterns are artificial, they encode assumptions and observations on travel behavior datasets and reasonably approximate the types of tours people would be expected to make on a daily basis. With Module 4 complete, particular destinations can be now be assigned for each trip within a person’s given activity pattern.
2.6 Module 5: Assignment of Trip Destination

Module 5 takes the assigned activity patterns from Module 4 and determines the actual destinations for every W, S, H and O trip within a resident’s activity pattern. In doing so, the spatial distribution of daily tours is generated for all residents. Module 5 uses the same data set employed in Module 2, focusing now on the patronage aspect provided for every workplace. Patronage refers to the number of persons who visit an employer in the data set. Module 5 attempts to send enough trips to every workplace in order to match their patronage levels in the dataset. However, as the patronage dataset is far less comprehensive and more inaccurate than the data obtained from Census surveys, Module 5 will not construct additional trips to ensure that patronage data is matched exactly. For example, if a workplace had 12 patrons, Module 5 would attempt to send 12 persons to visit the workplace, but would not send residents on additional trips to ensure that exactly 12 people visit the workplace.

While previous Modules provided most of the information used to determine the locations of W, S, and H trip destinations, the difficult part of Module 5 is to determine the location of O trips. To determine these O trips, Module 5 uses a Gravity Model to determine the likelihood of a given O Trip destination. The mechanics of this model are described in Equation (2.5).

\[
w_{i,j} = \frac{p_j}{\sum_j \frac{p_j}{D_{i,j}}} \quad \forall \ j \in J
\]

where:

- \( i \) = the index of the resident, referencing the resident’s current location
- \( J \) = the set of valid O type destinations, indexed by \( j \)
- \( p_j \) = the patronage at workplace \( j \)
- \( D_{i,j} \) = the distance between the resident’s current location \( i \) and the location of the O type trip \( j \)
- \( w_{i,j} \) = the weight of attraction to O type destination \( j \) for a resident at location \( i \)

The Gravity Model comes with additional spatial restrictions for O trips to encode assumptions on travel behavior. Similar to Module 4, all O trips are restricted
to geographically adjacent counties. In a hypothetical H-W-O-H tour, the O trip’s county would need to be adjacent to the W trip’s county. Likewise, in a H-S-O tour, the O trip’s county would need to be adjacent to the W trip’s county. Tours that include the W-O-W trip sequence (i.e. lunch trips) are restricted within 5 miles of the workplace. Additionally, O type trips to a minimum distance of 0.5 miles to ensure all trips represented are done via motorized transportation.

With Module 5 complete, the spatial distributions of all trips throughout the U.S. have been determined as demand and supply of travel have been completely linked. A sample output of one of the files generated for the state of New York is seen in Figure 2.5. This resident, with an activity pattern of 9, starts the morning by leaving home and going to school. After finishing up their time at school, they leave and go to work their after-school job at a local restaurant. Once they finish up their shift, this resident makes a quick stop at a nearby Goodwill store to do some evening shopping before heading home and ending their day. What remains now is to determine the temporal distribution of this demand, which is done by Module 6.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Residence State} & \text{County Code} & \text{Tributary ID} & \text{Block Code} & \text{HH ID} & \text{Person ID Number} & \text{Activity Pattern} \\
\hline
\text{H} & \text{N} & \text{S} & \text{201} & \text{918038} & \text{2485973} & \text{9} \\
\hline
\text{S} & \text{W} & \text{W} & \text{36021} & \text{42.389704} & \text{-73.485207} & \text{NA} \\
\hline
\text{W} & \text{N} & \text{O} & \text{36021} & \text{42.415647} & \text{-73.406524} & \text{NA} \\
\hline
\text{O} & \text{W} & \text{H} & \text{GOODWILL} & \text{36021} & \text{42.27617} & \text{-73.758432} \\
\hline
\text{H} & \text{W} & \text{O} & \text{36021} & \text{42.27331} & \text{-73.755525} & \text{45} \\
\hline
\end{array}
\]

\[
\text{FIGURE 2.5: Output from Module 5 for state of New York.}
\]

2.6.1 Submodule 5.5: Assignment of Trip Destination

While the main goal of Module 5 is to determine the spatial distributions of all daily tours, another sub-goal of Module 5 is to ensure that the demand generated is not only realistic, but also useful for the overarching aim of the entire model: to generate travel demand data useful for emerging passenger mobility services. Modal split is used to assign travel demand that one can reasonably assume would not be satisfied by passenger mobility services to alternative modes.

The first modal split taken is with regards to travel that could be satisfied through the use of non-motorized transportation, e.g. by biking or walking. For this case,
the modal split submodule works in parallel with Gravity Model in Module 5 to determine a limited number of cases that could be satisfied by non-motorized transportation. O trip destinations that are less than 0.5 miles from the origin location are rejected.

The second modal split focuses on trips too long to be satisfied by autonomous vehicles. The modal split identifies exceptionally long trips to be those that are greater than 200 miles in length. These trips are far more likely to take an alternative mode of transportation, e.g. air, and are routed. As S and O trips are confined geographically to adjacent counties, they cannot be serviced by this modal split, so the trips serviced are W type and can be thought of as representing commuters who work out of state or are attending a business trip.

A comprehensive dataset provided by the Federal Aviation Administration [20] on U.S. Airports is used to identify airports. When long trips are generated, Submodule 5.5 identifies them and alters their activity pattern so that the long trip is routed from airports near the origin and destination. After the resident completes their long trip, they are routed to their destination of work and then to a qualified hotel from Module 2 to complete the tour.

A gravity model similar to Equation 2.5 is employed as described in Equation 2.6.

\[
w_{i,j} = \frac{A_j}{\sum_j A_j D_{i,j}^2} \quad \forall \quad j \in J
\]

where:

\(i\) = the index of the resident, referencing the resident’s current location

\(J\) = the set of valid airport, indexed by \(j\)

\(A_j\) = the patronage at workplace \(j\)

\(D_{i,j}\) = the distance between the resident’s current location \(i\) and the location of the O type trip \(j\)

\(w_{i,j}\) = the weight of attraction to airport \(j\) for a resident at location \(i\)

Land area was employed as the attraction metric for two reasons. First, no widely available and comprehensive dataset exists detailing the patronage of every airport, while the FAA dataset selected provided land area on a scale comprehensive enough
for this model. Secondly, land area intuitively corresponds to the relative popularity of an airport. As airports become more popular, it is reasonable to assume they also become larger in order to service the increased demand.

A sample output of one of the files generated for the state of New York is seen in Figure 2.6. A resident residing north of New York City is assigned to commute to a petroleum refinery in Los Angeles. Instead of having the worker commute the more than 2,000 mile distance, Submodule 5.5 diverts the worker to LaGuardia airport and then across the country to LAX. After working his shift at the Paramount Petroleum Corporation, the worker retires to a nearby Crowne Plaza hotel. Note that the worker’s activity pattern, 14 (H-W-O-W-H), would require them to travel across the country twice. Given the great amount of distance involved, Submodule 5.5’s redirection is far more likely to occur in reality. The modal-split used within Sub-

<table>
<thead>
<tr>
<th>County Code</th>
<th>Tract Code</th>
<th>Block Code</th>
<th>HH ID</th>
<th>Person ID Number</th>
<th>Activity Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>061</td>
<td>202</td>
<td>2011</td>
<td>918038</td>
<td>2458973</td>
<td>14</td>
</tr>
<tr>
<td>N</td>
<td>A</td>
<td>Node 1 Name</td>
<td>Node 1 County</td>
<td>Node 1 Lat</td>
<td>Node 1 Lon</td>
</tr>
<tr>
<td>H</td>
<td>A</td>
<td>Node 2 Name</td>
<td>Node 2 County</td>
<td>Node 2 Lat</td>
<td>Node 2 Lon</td>
</tr>
<tr>
<td>A</td>
<td>W</td>
<td>LAGUARDIA</td>
<td>36081</td>
<td>40.777250</td>
<td>-73.872610</td>
</tr>
<tr>
<td>A</td>
<td>O</td>
<td>LOS ANGELES INTL</td>
<td>06037</td>
<td>33.942495</td>
<td>-118.400689</td>
</tr>
<tr>
<td>W</td>
<td>N</td>
<td>PARAMOUNT PETROLEUM CORP</td>
<td>06037</td>
<td>33.768786</td>
<td>-118.260961</td>
</tr>
<tr>
<td>W</td>
<td>N</td>
<td>CROWNE PLAZA HARBOR</td>
<td>06037</td>
<td>33.738305</td>
<td>-118.281978</td>
</tr>
</tbody>
</table>

**Figure 2.6:** Output from Submodule 5.5 for state of New York.

module 5.5 is quite similar to any other possible modal-split. With this framework in place, it is conceivable to employ modal split with other modes of transit, provided the dataset supporting the mode is comprehensive enough for the geographic scope of the model. Modal split requires the usage of some key feature that allows the modeler to determine the mode a traveler would use. Ultimately, this decision making is done to imitate personal choice. In Submodule 5.5, distance was used as the key feature as travellers reasonably would switch to a faster mode of travel given the time required. However, alternative features, such as proximity to geographic locations (e.g. mass transit stations) could give way to a rail-based modal split. Secondly, the trip identified needs to be placed onto a network of modal options, simulating discrete choice on the behalf of the traveler. In Submodule 5.5, travelers were sent to airports in proximity to their destinations. However, one could further complicate the method of network placement by adding additional constraints in order to
more accurately imitate discrete choice by the traveler. While Submodule 5.5 is ultimately a proof-of-concept, the framework established will undoubtedly be useful in expanding alternative model split.

2.7 Module 6: Assignment of Arrival and Departure Times

Module 6 provides the final step in the model by creating the temporal distribution of trips. Specifically, Module 6 constructs arrival times, duration of stays and departure times for every node generated within Module 5. The goal of Module 6 is to match the temporal distribution of commuting, school, and errand trips throughout the day from the 2009 National Household Transportation Survey [21]. In terms of the daily tour, Module 6 determines the arrival, duration of stay and departure time for every node in the tour.

An artificial dataset of start times and end times, collectively defined as bell times, as well as durations, are constructed based on NAICS industry-wide assumptions on temporal distributions derived from the 2009 Survey. For the first trips of the day, arrival times are drawn from an exponential distribution whose expected arrival time is 5 minutes before the bell time. An arrival window is established from 10 minutes before the bell-time to the bell-time, effectively concluding that 10% of all arrivals will be outside of the window, i.e. late. The exponential distribution was selected as it often empirically models arrival times. Moreover, sequences of exponentials form a Poisson distribution whose scale parameter is the number of arrivals to a location. Departure times follow the same scheme in a reverse process, encoding the assumption that one is equally likely to arrive five minutes early as they are five minutes late.

Durations of stay are generated from a normal distribution, with mean from the NAICS Industry dataset and a 15% variance. For part-time work and part-time schooling, these durations are generated from a normal distribution with a three hour mean and 15% variance. For trips that go to home and then elsewhere, durations are sampled uniformly between 15 minutes to an hour.

Module 6 begins the time-scheduling process with the first trip and walks through
Chapter 2. Methodology

the daily tour computing duration of stay, departure time at the current node, arrival
time at the subsequent node, and so on. Trip times are scheduled independently of
all other travellers. As mentioned in the Introduction, very little is known about
individual-individual covariances, especially those amongst households. Moreover,
doing this allows for greater simplification. Trips generated are given times in sec-
onds past midnight.

Module 6 enhances the trip sequence file generated by Module 5 by adding the tem-
poral dimension to the tour schedule. With Module 6 complete, the travel demand
for the entire U.S. is both spatially and temporally distributed. 51 .csv files are con-
structed for the 50 States and the District of Columbia, constructed row-by-row. An
example output from Module 6 is seen in Figure 2.7. This sample output shows the
working day of a Taxidermist, who begins his day around 6:45 AM. He commutes
roughly 35 minutes to work and works until 4:30 PM and then heads to a local Olive
Garden nearby his home for dinner. After returning for a brief respite at home, he
heads out to the Beltone Hearing Aid Center. After quickly stopping by the Center,
he returns home again before heading out for an evening at the local Pratt Pub &
Oyster Bar. The final node, his return home from the bar, is omitted as for all trips,
less those handled by Submodule 5.5, it is the home of the resident.

<table>
<thead>
<tr>
<th>Residence State</th>
<th>County Code</th>
<th>Block Code</th>
<th>HH ID</th>
<th>Person ID Number</th>
<th>Activity Pattern</th>
<th>Arrival Time</th>
<th>Departure Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>001</td>
<td>20002</td>
<td>2039</td>
<td>13894</td>
<td>0100037250</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Node 1 Type</td>
<td>Node 1 Predecessor</td>
<td>Node 1 Successor</td>
<td>Node 1 Name</td>
<td>Node 1 County</td>
<td>Node 1 Lat</td>
<td>Node 1 Lon</td>
<td>01001</td>
</tr>
<tr>
<td>Node 2 Type</td>
<td>Node 2 Predecessor</td>
<td>Node 2 Successor</td>
<td>Node 2 Name</td>
<td>Node 2 County</td>
<td>Node 2 Lat</td>
<td>Node 2 Lon</td>
<td>01051</td>
</tr>
<tr>
<td>Node 3 Type</td>
<td>Node 3 Predecessor</td>
<td>Node 3 Successor</td>
<td>Node 3 Name</td>
<td>Node 3 County</td>
<td>Node 3 Lat</td>
<td>Node 3 Lon</td>
<td>01001</td>
</tr>
<tr>
<td>Node 4 Type</td>
<td>Node 4 Predecessor</td>
<td>Node 4 Successor</td>
<td>Node 4 Name</td>
<td>Node 4 County</td>
<td>Node 4 Lat</td>
<td>Node 4 Lon</td>
<td>01001</td>
</tr>
<tr>
<td>Node 5 Type</td>
<td>Node 5 Predecessor</td>
<td>Node 5 Successor</td>
<td>Node 5 Name</td>
<td>Node 5 County</td>
<td>Node 5 Lat</td>
<td>Node 5 Lon</td>
<td>01001</td>
</tr>
<tr>
<td>Node 6 Type</td>
<td>Node 6 Predecessor</td>
<td>Node 6 Successor</td>
<td>Node 6 Name</td>
<td>Node 6 County</td>
<td>Node 6 Lat</td>
<td>Node 6 Lon</td>
<td>01001</td>
</tr>
<tr>
<td>Node 7 Type</td>
<td>Node 7 Predecessor</td>
<td>Node 7 Successor</td>
<td>Node 7 Name</td>
<td>Node 7 County</td>
<td>Node 7 Lat</td>
<td>Node 7 Lon</td>
<td>01001</td>
</tr>
<tr>
<td>Node 8 Type</td>
<td>Node 8 Predecessor</td>
<td>Node 8 Successor</td>
<td>Node 8 Name</td>
<td>Node 8 County</td>
<td>Node 8 Lat</td>
<td>Node 8 Lon</td>
<td>01001</td>
</tr>
</tbody>
</table>

Figure 2.7: Output from Module 6 for state of Alabama.

2.8 Module 7: Pixelization of Demand

With Module 6 complete, the non-commercial travel demand of the entire United
States is determined both spatially and temporally. While determining this travel
demand is undoubtedly the main goal of the model presented in this paper, there
is more to be done. As mentioned in the introduction, the overarching goal of the model was to provide travel demand for the explicit purpose of analyzing and developing implementations of aTaxi transportation systems. In order to understand how Module 7 furthers this goal, a brief discussion of Autonomous Taxi Networks (ATNs) is necessary.

2.8.1 Autonomous Taxi Networks

In Brownell et al [22], ATNs are presented as a next generation solution to the problems of personally owned automobiles. Brownell defines an ATN by two characteristics. First, an ATN consists of fully autonomous, constantly communicating vehicles - the taxis that comprise the network - which pick up passengers and deliver them to their requested destinations. Second, the taxis are built completely around the need of the user to satisfy travel demand. They do not operate on a regular schedule such as a bus or a train, but are only deployed when a passenger indicates demand. While the infrastructural needs of ATNs vary by model, the parameter that is most relevant to the usage of this model is the mechanism by which riders are picked up and delivered to their destinations by the autonomous taxis. Two ATN models developed by Kornhauser et al and Mark Gorton will be described in order to determine how the demand generated can be altered as to better serve the goal of analyzing ATNs.

In Kornhauser et al [23], an ATN system is designed that borrows its layout from the classic implementation of a Personal Rapid Transit (PRT) network. This model establishes stations across the state of New Jersey in a uniform grid spaced 0.5 mile apart from one another. The PRT model assumes that passengers will walk to their closest station, roughly 0.35 miles away at greatest. This corresponds to a maximum travel of roughly seven minutes by the traveller (assuming a 3mph travel speed) though the majority of passengers would take far less time to travel. Similarly, at the end of their trip, passengers would disembark at a station and walk to their destination, again a maximum distance of 0.35 miles away. Ridesharing is also implemented in the model. Two or more riders will share the same vehicle if their origin and destinations are within the same 0.5-by-0.5 mile square, corresponding to the same arriving
and departing taxi stand. Moreover, a temporal constraint must be satisfied, namely that they both arrive within $t_{\text{max}}$ seconds of one another at the originating taxi station.

Mark Gorton [24] introduces an alternative ATN called Smart Para-Transit (SPT). In the SPT system, individual people request a trip to a given destination, at which point they are picked up by the SPT vehicle at a “central transit point.” Along the way, the vehicle may stop at one or two other “central transit points” to pick up more passengers. This provides a significant reduction in autonomous taxis needed to satisfy demand as well as reduced effort on behalf of the traveller. While Kornhauser’s ATN required individuals to move to the same ATN station within 0.35 miles of each other, Gorton’s system allows for autonomous vehicles to pick up passengers over wide areas. For example, while two individuals in the PRT system would have to walk at most 0.7 miles collectively to an ATN station, under the SPT system, individuals can be picked up by the autonomous vehicle. Assuming an average speed of 25 mph, an autonomous taxi in the SPT model would be able to pick up two travellers within 2.5 miles over each other in the same time as the PRT system, a significant increase. The SPT system follows the spatial constraint imposed by the PRT model by adopting a ‘station’ approach, with an increase in grid length from 0.5 miles to 1.5 miles. This corresponds to a maximum distance of roughly 1.05 miles from the center node. Figure 2.8 further explains the methodology used by illustrating the pick up and delivery of twelve different passengers from Manhattan to North-Eastern New Jersey.

While both ATN models have similarities and differences, the greatest similarity within the passenger pick-up and delivery methodology is the spatial aggregation of travel demand. While real-life implementations of these models might attempt to disaggregate demand or aggregate it on a different mode, like the temporal, Gorton and Kornhauser both make use of spatial aggregation into different sized pixels in order to route vehicles. Brownell’s spatial analysis of the travel demand is fundamentally based on this simplification as with demand pixelized, it becomes far easier to develop generalized cost functions that can provide upper bounds on fleet sizes and determine the cost of travel. The pixelization process also greatly assists
in simplifying computation. To assist in this process, Module 7 advances Brownell’s initial New Jersey pixelization process to the entire nation.

### 2.8.2 Pixelization of the United States

Module 6 outputs the travel demand of this model in the form of 51 .csv files for all 50 states and the District of Columbia. Each row contains a daily trip tour completed by a person, with every individual trip chain comprised within the trip tour. As trip tours aggregate trips by resident, it is necessary to first go through each trip tour and disaggregate the trips. This is done by walking through the trip tours and separating each node into an arrival trip and a departure trip depending on the characteristics of the node.

With the trips disaggregated, it is now possible to begin the process of pixelization. To discretize the entire U.S., we overlay a pixel grid with pixels of dimension 0.5 x 0.5 miles in side length. Following earlier analysis by Kornhauset et al [23], we use Equations 2.7 and 2.8 to map every \((Lon_o, Lat_o)\) pair to a pixel.

\[
X_P = \frac{69.1}{d_s} \cos\left(\frac{Lat_o}{57.3}\right)(Lon_p - Lon_o)
\]  

(2.7)
\[ Y_p = \frac{69.1}{d_s} (Lat_p - Lat_o) \quad (2.8) \]

where:

\( d_s \) = the length of each pixel side

\((Lon_o, Lat_o)\) = the longitude-latitude pair of the trip

\((Lon_p, Lat_p)\) = the longitude-latitude pair of the bottom left corner of the grid

\((X_p, Y_p)\) = the x-y coordinates of the pixel for that trip’s longitude-latitude pair

As the majority of the analysis used from this dataset is based on Kornhauser’s PRT ATN implementation, we let \( D_s = 0.5 \) miles. The bottom left corner of the grid, \((0, 0)\) or \((Lon_p, Lat_p)\), corresponds to \((-75.6^\circ E, 38.0^\circ N)\) in the gridding of the United States. With these variables, we can simplify Equations 2.7 and 2.8 as follows.

\[ X_p = \text{floor}(108.907(Lon_p - 75.6)) \quad (2.9) \]

\[ Y_p = \text{floor}(138.2(Lat_p - 38.0)) \quad (2.10) \]

Equations 2.9 and 2.10 simplify the pixelization process by redefining the nature of the grid. All longitude-latitude pairs are assigned to the bottom left corner of each pixel. Additionally, each pixel is denoted by this bottom left numbering system. Figure 2.9 shows an example of this grid overlaid on the Princeton, New Jersey area.

With all trips pixelized, all that remains is to restructure the data as to make it readily available for pixel-based analysis. To do this, we first take each file and split the state files into files partitioned by County. After partitioning the state files by county, we assign every pixel to a county file. This ensures that every pixel is unique to a county file and will only be found in that county file. After this, we sort each county file in descending order by \( X_p, Y_p \) and Departure Time. Once the files have been sorted, each file is split into a subset of county files if more than 1,000,000 rows exist in the file. Each county subset file has less than 1,000,000 rows and the subset files collectively define the original main county file. For example, a file from New Jersey’s Mercer County with 4,500,000 rows (corresponding to 4,500,000 trips) would be split into five files, four of which would be roughly 1,000,000 rows in length and one of roughly 500,000 rows in length. Files are split as to ensure pixel-uniqueness.
is transferred to the county subset files as well. With Module 7 complete, the travel demand for the entire U.S. is completely pixelated and ready for analysis. Over 3,500 .csv files are created, each 1,000,000 rows in length or less, detailing the trips for every county in the United States. An example output for Conecuh County, Alabama is shown in Figure 2.10.

![Figure 2.9: Pixelization of the Princeton, NJ Area from Kornhauser et. al [23].](image)

Figure 2.9: Pixelization of the Princeton, NJ Area from Kornhauser et. al [23].

2.9 Fundamental Assumptions

A model of real world phenomena is only as good as the assumptions it is based on. To assist in the critical evaluation and future improvement of this model, fundamental assumptions taken by the model are detailed. The discussion of these assumptions are broken up by Module. Most of the assumptions listed are relevant to the
scope of the data provided in addition to standard assumptions with computational processing time and memory limitations.

2.9.1 Geographical Assumptions

The methodology described in this chapter uses various tiers of geographical aggregation used by the U.S. Census Bureau. This model uses, from largest to smallest, the following geographical classifications: states, counties, tracts and blocks. While most readers will undoubtedly be familiar with the state level classification, the county, block and tract level classifications may be more unfamiliar. To gain a better appreciation for the geographical scope of those model, the geographical assumptions are briefly covered here.

The demand simulations are performed on the 50 states and the District of Columbia. Amongst all 51 states or state equivalents, there are 3,143 county or county equivalents. On average, each of these counties is composed of 23 Census Tracts. Each of these tracts in turn can be composed of 151 Census Blocks on average. Module 1 begins by simulating every resident from the 11,078,297 Census blocks that decompose the entire United States. Each block is simulated independently and all of the data used, aside from household income, is done at the block level. The second level of geographical aggregation is the Census tract level. Census tracts are composed of several block groups (groupings containing many blocks). The average number of groups within a tract is approximately three. This geographical hierarchy determines the method by which tract level data is applied to the underlying blocks. In the case of household income, tract level data is assumed to apply uniformly to the blocks - that is, every block within a tract is assumed to be uniformly distributed with the same household income distribution as the entire tract. While such decisions are motivated entirely by data constraints, there is some accuracy to the assumption that a relatively large area is uniformly representative of its parts. However, in diverse urban areas, e.g. Manhattan or Seattle, the make-up of blocks contained within a tract may be diverse enough to refute this assumption.
2.9.2 Methodological Assumptions

While all of the methodological assumptions have been stated directly or implicitly in the earlier sections of this chapter, future readers may find it helpful to identify key assumptions in the aims of advancing or correcting the model. These key assumptions are listed by Module below.

**Module 1**

- Every household and therefore every resident geographically resides at the centroid of the Census block they are in.

- The distribution of residents by age and sex is known at the Block level, but ages are divided in the census into intervals, e.g. 0-4, 5-9, and so on. Ages within these intervals are assumed to be uniformly distributed and are sampled as such.

- The population is divided into households and group quarters such as dormitories and nursing homes. All are represented as households however and have a household type from 0 to 8. 0 and 1 refer to actual households and the rest refer to group quarters. A full listing can be found in Figure 2.1.

- Households are built by first choosing a household size and a female or male householder. The rest are filled based on household relations distributions from the Census Summary File 1. While householders are sampled without replacement in the generation of the population, all sampling done after is sampled with replacement.

- Residents are assigned a traveler type from 0-6, which helps are later used to specify their sequences of daily activities. These traveler types are based on age and household type.

- Incomes are assigned to each household. Once a household has an income assigned to it, incomes are randomly distributed amongst workers that are qualified to receive one.
• All distances are calculated precisely using Great Circle Distance for every module.

Module 2

• All workers are categorized under traveler types 2, 4 and 5. Aside from these categories, no other types are assigned an employer. Those not within these categories are assumed to be retired, too young, are currently not travelling to work today or are unemployed.

• County level data used from this point and for all other points throughout the modules are assumed to be representative of the block level population data.

• Workers are assigned an industry, a county of work and an employer in this order. The attractiveness of these categories is appropriately measured using Gravity Models.

Module 3

• Students at or below six years of age are assumed to not go to school, despite the opportunity for pre-kindergarten education.

• Students are restricted to their home county or adjacent counties for school enrollment.

• Alternative proxies for school enrollment are used when school enrollment data is not available, as is the case in post-secondary institutions. More information is available in Chapter 3.

• Students are assigned a school given the type of school they are attending and the counties in which they can attend valid schools. Once these two criterion have been determined, attractiveness weights are assigned to a school using a Gravity Model.

Module 4

• All possible travel tours taken by every resident through the entire United States are completely described by Table 2.1.
• The joint distribution of travel tours by traveler type is completely described by Figure 3.2.

• Every tour used begins and ends at home.

Module 5

• O Type trips made from work during lunch must be located within the work county.

• All other O Type trips can be in the county itself or any county that is adjacent to the resident’s current county.

• The geographic location of O Type Trips are sampled with replacement from a distribution that is proportional to the daily patronage at that location over the Euclidean distance from the resident’s location to that place.

• Trips under a quarter mile in distance are ignored for O Type trips and all W-O-W Lunch Trip chains must be within 5 miles of the W node’s physical location.

Module 6

• NAISC Bell times are used for modelling departure and arrival times. Patronage durations are also used. See Figure 3.3 for further information.

• Arrival and departure times are represented as exponential distributions. Sequences of these arrival and departure times are distributed as Poisson processes with rate parameter equal to the number of arrivals at the location.

• The expected arrival and departure times are 5 minutes before the location’s bell times.

• Patronage durations are distributed normally, with mean from the NAISC Patronage Duration and variance of 15%.

• For part-time work and part-time schooling, durations of stay are distributed normally with a mean of three hours and a variance of 15%.
• For trips that go home and then elsewhere, durations are sampled uniformly between 15 minutes to an hour.

• All times given are in seconds from midnight.

• An average speed of 30 MPH is used to calculate time for all trips.

Module 7

• In the pixelization process, each side of the grid is represented by some measure of distance as a fraction of a degree of latitude or longitude. As the Earth is not a perfect sphere, these square pixels become more and more flattened as one moves towards the equator. This flattening is assumed to be negligible for the purposes of ATN analysis.

• Some pixels will cross county borders and contain residents of multiple counties. In order to preserve uniqueness in the county-pixel assignment process, pixels are assigned to a county based on the county of the first originating trip to come from that pixel in sequential order. To further illustrate this assumption, imagine a pixel laid across four counties A, B, C and D with four travellers from Counties A, B, C and D. If the pixelization implementation reads through the trip file containing this pixel and the trip from traveller B is the first read, this pixel is assigned to County B, regardless of the order or contents of any subsequent trips in the trip file.
Chapter 3

Data

The contents of this chapter present the input data used in the model to generate the simulated national population and the travel demand of this population. The demand synthesizer requires extensive data on demographics, business activities and school enrollment across the entire U.S. to construct the travel demand profile of an average U.S. workday. Data is presented in the order that they are used by each Module in the model.

3.1 Module 1

Module 1 creates 51 .csv files listing all of the residents, grouped by household, for each state or state equivalent. Each resident has information on their personal attributes, like age, gender, income and household location. The data used to generate these files is found throughout multiple sources affiliated with the U.S. Census Bureau.

3.1.1 2010 United States Census Summary File 1 Data

The demographic and housing data used in the synthesizer comes from the 2010 U.S. Census Summary File 1 (SF1). This data set contains the most comprehensive set of summary statistics compiled from the 2010 U.S. Census, with a particular focus on population items such as “sex, age, ..., household relationship, household size, family type, family size, and group quarters.” [25] The data from SF1 is the only source comprehensively available at the block level and provides distributions on population. It is used to populate each census block as they were surveyed in
2010. The relevant data files were obtained from the Census Bureau’s FTP site. The specific data used are found in data segments 4, 5, 6 and 44 which provide information on demographics, group quarters and population distributions of families within households [25]. Mufti’s original Microsoft Access queries and three separate SQL queries were used to create three separate text files for each of the states: ‘##DemographicQuery’, ‘##GroupQuartersQuery’ and ‘##FamilyQuery’ (where ## represents the two character state abbreviation). Each row in each file represents a Census block within that particular state. ‘##DemographicQuery’ uses SF1 Data Segments 4 and 44 and provides information on resident’s geographic attributes, sex by age by gender and households by size. ‘##GroupQuartersQuery’ uses SF1 Data Segment 6 and provides information on group quarter population distributions by type by age. ‘##FamilyQuery’ uses SF1 Data Segment 5 and provides information on household population distributions by household type by familial relation.

3.1.2 2010 American Community Survey of Household Income

The income data used in the synthesizer come from the 2010 American Community Survey (ACS). The ACS is conducted every year and provides 1, 3 and 5 year estimates of various population attributes. The ACS’s data table labeled ‘2008-2012 Household Income in the past 12 Months (in 2012 Inflation-Adjusted Dollars)’ [26] provides 5 year estimates of household income by household type for every census tract throughout all of the state or state equivalent. This data set, obtained through the American FactFinder Database, is used to provide estimates of household income by applying tract level income distribution to the blocks that compose a tract.

3.2 Module 2

Module 2 updates the 51 .csv files from Module 1 by assigning eligible workers a workplace. To accomplish this, workers are assigned a county of work, then an industry of work within that county and finally a workplace. The data used for Module 2 is a combination of nation-wide public and private employment. NAICS Codes, as employed by the U.S. Census Bureau, are used extensively in this Module and are detailed in Table 3.1.
3.2.1 2010 American Community Survey Journey-to-Work Census

In order to assign workers to a county, data from the U.S. Census Bureau on commuting with regards to employment were used. The Census Bureau’s dataset labeled ‘Residence County to Workplace County Flows for the United States and Puerto Rico Sorted by Residence Geography: 2006-2010’ [17] was used. For every U.S. county, this dataset provides information on every county commuted to by workers from this county as well as estimates of the number of commuters taking these specific routes. Only data related to the 50 States and the District of Columbia was used. This data was obtained from the Metropolitan and Micropolitan research group of the Census Bureau.

3.2.2 2010 American Community Survey Industry Type Participation by Gender and Median Income

In order to assign a particular worker to an industry, given their gender, income level and working county, data from the 2010 ACS was used that described rates of participation in each NAICS industry by gender. The specific data table used was labeled ‘Industry by Sex and Median Earnings in the Past 12 Months (in 2012 Inflation-Adjusted Dollars) for the Civilian Employed Population 16 Years and Over’ [18]. As with earlier ACS datasets, this data was obtained from the American FactFinder database.

3.2.3 Employee and Patronage Data

To assign a worker to an employer, a comprehensive list of every workplace within the U.S., including information regarding the workplaces physical location, industry type, employment rate and patronage rate, is needed. Unfortunately, no such publicly available dataset exists. However, quasi-comprehensive datasets exist in the private sector. InfoGroup, a private data and marketing company, maintains a U.S. Business Database [27] which provided the information needed to assign workers to employers. The ORF467 Class of Fall 2013 identified this database and helped
3.3 Module 3

Module 3 finishes the selection of the fundamental anchor points in the trip tour chain by determining schools for eligible students. Given the residing county and the type of school a student was eligible for, a student is assigned a school. In order to assign this school, enrollment information is needed for every school.

3.3.1 National Center for Education Statistics

To assign eligible students to appropriate schools, data from the United States Department of Education was used that provided information on school enrollment. The National Center for Education Statistics (NCES), a division of the Department of Education, maintains a "Common Core" dataset [19] that details school location,
type (a combination of public/private and elementary/middle/high) and enrollment for every pre-secondary school in the U.S. To divide students into private and public schools, national rates provided by the NCES for participation in each type of school for each age level are used.

### 3.3.2 Post-Secondary School Data

For post-secondary schools, there exists no publicly available dataset that is comprehensive enough for the purposes of this model. While it is a relatively trivial matter to obtain the names and locations of all post-secondary schools, enrollment is not often provided and, if provided, is not guaranteed to be correct. The NCES does not provide the same level of detail for post-secondary schools as it does for secondary schools as these institutions do not fall under the purview of the Department of Education. In order to overcome this data, a proxy to enrollment is used: workers employed by a university. As all post-secondary schools have employees (teachers, administrators, staff, etc), they are typically well-represented in the InfoGroup Business dataset.

### 3.4 Module 4

Activity pattern distributions, pioneered by Mufti, are artificial distributions used to simulate the average number of trips taken by a type of traveller by assigning likelihoods to tours, given a resident’s traveller type. These artificial distributions are driven by the average number of trips taken each day and the distribution of each trip by purpose. According to the 2009 National Household Transportation Survey, the average number of daily trips taken was 3.79. Of these 3.79 trips, 10% are from school or church, 15% are from work and 75% are personal/errands/other [21]. The activity pattern distributions approximate these general empirical distributions. These distributions are also logically consistent with the traveler type. For example, non-workers and non-students do not visit workplaces or schools respectively. These distributions can be seen in Table 3.2.
3.5 Module 5

To assign all Other type trips, the same data used from Module 2 is used. However, patronage data is used in place of employment data.

3.5.1 Federal Aviation Administration Facilities Data

In order to route long trips onto airport flows, a comprehensive dataset is needed to describe all available airports throughout the 51 states or state equivalents with some key feature that is correlated with the amount of travel going through each facility. While the Federal Aviation Administration maintains a comprehensive dataset describing all airport related facilities throughout the nation, there is no publicly available information describing the number of passengers that flow domestically through these airports at this comprehensive of a level. In order to accomplish this, a proxy for travel demand at an airport is used: land area. It is reasonable to assume that larger airports tend to be correlated with more runways and thus service larger demand. Moreover, this was the only feature that was uniformly present for all aerial facilities throughout the U.S. for this dataset.

3.6 Module 6

Module 6 determines the temporal distribution of all trips by providing information on trip start times, durations of stay and modelling for special trip anchors, like school and work. The 2009 National Household Transportation Survey provides a surveyed temporal distribution of trips through the day, by trip purpose, in Figure 3.1. However, there exists little, if any, information on comprehensive transportation sequencing throughout the workday. As a result, guesswork is used to determine these trip sequences. Information on timing sequences is provided in Figures 3.2 and 3.3.

3.7 Module 7

The pixelization done by Module 7 requires no external data as all calculations used output from the earlier Modules to perform pixelization. However, to correct for
errors in incorrect latitude/longitude assignment to county, the FIPS code associated with every county is checked with the associated longitude-latitude pair of a trip originating from this. In order to do so, polygons of every county or county-equivalent within the United States are needed. With these polygons, the raycasting application of the Jordan Curve Theorem is used to determine whether or not that point is contained within the polygon. The U.S. Census Bureau provides a comprehensive dataset of Topologically Integrated Geographic Encoding and Referencing (TIGER)\textsuperscript{28} products which include polygons for every county.
### Chapter 3. Data

#### Figure 3.2: Trip Tour distributions adopted from Wyrough [4].

<table>
<thead>
<tr>
<th>Traveler Type</th>
<th>Do-Not-Travel</th>
<th>School-No-Work</th>
<th>School-Work</th>
<th>College-No-Work</th>
<th>College-Work</th>
<th>Typical Worker</th>
<th>Home-Worker</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>1.0000</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0040</td>
<td>0.1000</td>
</tr>
<tr>
<td>H-W-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0075</td>
<td>0.0500</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-H</td>
<td>0.0000</td>
<td>0.1250</td>
<td>0.0500</td>
<td>0.1500</td>
<td>0.0075</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.4050</td>
<td>0.0000</td>
<td>0.2000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-W-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-W-S-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1960</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-O-H</td>
<td>0.0000</td>
<td>0.3500</td>
<td>0.0850</td>
<td>0.3000</td>
<td>0.0075</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-O-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.3500</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-W-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1500</td>
<td>0.0000</td>
<td>0.2600</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-W-S-O-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2600</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-W-H-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2600</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-H-O-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.4500</td>
<td>0.0000</td>
<td>0.0075</td>
<td>0.1500</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-O-H-O-W-H</td>
<td>0.0000</td>
<td>0.3250</td>
<td>0.0000</td>
<td>0.4500</td>
<td>0.0075</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-W-O-H-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1500</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-O-H-O-O-H</td>
<td>0.0000</td>
<td>0.1500</td>
<td>0.0450</td>
<td>0.0000</td>
<td>0.0075</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-W-H-O-O-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0050</td>
<td>0.1500</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-H-O-O-O-H</td>
<td>0.0000</td>
<td>0.0250</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1500</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-W-O-H-O-O-O-H</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1500</td>
<td>0.0000</td>
</tr>
<tr>
<td>H-S-O-O-H-O-O-O-H</td>
<td>0.0000</td>
<td>0.0150</td>
<td>0.0500</td>
<td>0.0000</td>
<td>0.0010</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>1.0000</strong></td>
<td><strong>1.0000</strong></td>
<td><strong>1.0000</strong></td>
<td><strong>1.0000</strong></td>
<td><strong>1.0000</strong></td>
<td><strong>1.0000</strong></td>
<td><strong>1.0000</strong></td>
</tr>
<tr>
<td><strong>Expected No. of Trips</strong></td>
<td><strong>0.00</strong></td>
<td><strong>3.58</strong></td>
<td><strong>3.37</strong></td>
<td><strong>3.58</strong></td>
<td><strong>3.59</strong></td>
<td><strong>4.59</strong></td>
<td><strong>2.55</strong></td>
</tr>
</tbody>
</table>

---

**Figure 3.3: Bell Times and Durations by NAISC Industry adopted from Wyrough [4].**

<table>
<thead>
<tr>
<th>Industry</th>
<th>Employee Times</th>
<th>Patronage Times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Start Bell</td>
<td>End Bell</td>
</tr>
<tr>
<td>Agriculture, Forestry, etc.</td>
<td>7:30AM</td>
<td>4:30PM</td>
</tr>
<tr>
<td>Mining, Quarrying, etc.</td>
<td>7:30AM</td>
<td>4:30PM</td>
</tr>
<tr>
<td>Utilities</td>
<td>8:30AM</td>
<td>5:00PM</td>
</tr>
<tr>
<td>Construction</td>
<td>7:30AM</td>
<td>4:30PM</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>8:00AM</td>
<td>4:30PM</td>
</tr>
<tr>
<td>Wholesale Trade</td>
<td>9:00AM</td>
<td>5:00PM</td>
</tr>
<tr>
<td>Retail Trade</td>
<td>9:00AM</td>
<td>6:00PM</td>
</tr>
<tr>
<td>Transportation and Warehousing</td>
<td>8:00AM</td>
<td>5:00PM</td>
</tr>
<tr>
<td>Information</td>
<td>10:00AM</td>
<td>7:00PM</td>
</tr>
<tr>
<td>Finance and Insurance</td>
<td>9:00AM</td>
<td>5:00PM</td>
</tr>
<tr>
<td>Real Estate, etc.</td>
<td>9:00AM</td>
<td>5:00PM</td>
</tr>
<tr>
<td>Professional Services</td>
<td>9:00AM</td>
<td>6:30PM</td>
</tr>
<tr>
<td>Management of Companies, etc.</td>
<td>9:00AM</td>
<td>6:00PM</td>
</tr>
<tr>
<td>Administrative Services, etc.</td>
<td>8:30AM</td>
<td>5:00PM</td>
</tr>
<tr>
<td>Education Services</td>
<td>8:30AM</td>
<td>4:00AM</td>
</tr>
<tr>
<td>Health Care, etc.</td>
<td>8:30AM</td>
<td>4:30AM</td>
</tr>
<tr>
<td>Arts, Entertainment, Recreation</td>
<td>11:00AM</td>
<td>8:00PM</td>
</tr>
<tr>
<td>Accommodation and Food Services</td>
<td>11:00AM</td>
<td>8:00PM</td>
</tr>
<tr>
<td>Other Services</td>
<td>9:00AM</td>
<td>5:00PM</td>
</tr>
<tr>
<td>Public Administration</td>
<td>9:00AM</td>
<td>5:00PM</td>
</tr>
</tbody>
</table>
Chapter 4

Conclusions

4.1 Data Applicability

There are some issues with the model regarding the applicability of data. The first assumption of the model, that everyone resides at the centroid of a census block, is greatly limiting in that demand is aggregated geographically to these centroids, as opposed to being disaggregated across houses. An alternative data source to look to might be statewide voter registration databases, which provide a far more detailed level of data for individuals, at the expense of reduced coverage.

The generation of the populace is ultimately determined by the U.S. Census, which is conducted only every 10 years. However, there is significant growth in population between year to year. The applicability of 2010’s population to the current year becomes less and less reasonable as time progresses. To overcome this issue, the American Community Survey, an alternative survey to the U.S. Census conducted every year, is used to adopt the 2010 population to current population levels. To ensure that every iteration of the model is accurate, updated information on unemployment rates and school enrollment must be obtained and assumptions made on these criterion must also be updated. However, it is important to keep in mind that the goal of the model is not to obtain exact realistic individual trip distributions for every person, but rather, to create a dataset that matches the real spatial and temporal distribution of travel demand throughout the nation for every resident. While data on individual trip behavior is useful in enhancing the realism of these distributions, it should not be expected that every resident matches exactly with their true workplace and patronage locations.
The model makes basic assumptions that personal attributes such as gender, age, income and residence location explain daily movement, workplaces and patronage locations. While this assumption appears to be reasonable, using these sets of features greatly restricts the scope and methodologies by which available data can be applied to better match a person to their daily travel needs. More work needs to be done to find alternative variables that are correlated with predictable activity patterns. Gender and race appear as plausible explanatory variables. However, the U.S. Census Bureau maintains hundreds of comprehensive datasets detailing personal attributes. While gender and race would be ideal next steps, the best step would be to search through these datasets and determine what population attributes can be used to enhance the prediction of activity patterns.

There exist some issues with the applicability of data in Employment and Patronage. While information from the ACS regarding median income for industries given county and gender is quite comprehensive, the actual dataset of workplaces is only quasi-comprehensive. In some cases, there are industries which are represented in the ACS but not in the workplace dataset. Two solutions exist for this scenario: either route workers to the ACS industry and create new workplaces for them or prevent workers from entering this industry and assign them to the next best alternative. The latter solution, which is employed in this scenario throughout the model, highlights the need for a more comprehensive dataset of businesses. Moreover, information on employee and patronage ratios was tailored from the New Jersey state model to apply to workplaces in our scenario. While the New Jersey employee and patronage ratios worked fine in New Jersey, their applicability to a national dataset will need further fine-tuning to ensure that the assignment of workplaces to individuals is accurate. These same concerns exist in the case of assignment to schools. While information on pre-secondary institutions was generally comprehensive and accurate, the proxy substitution of workers for student enrollment for post-secondary schools is not the most accurate or reasonable proxy. An ideal next step would be to find information on the true enrollment within these post-secondary institutions.
The usage of Traveler Type designations date back to Mufti’s original implementation of the model. Mufti’s original use of the categories was to allocate the population into distinct subsets that behave in somewhat similar manners. While these distinctions do maintain similar characteristics, they are made based off of national levels of data, as opposed to county levels. Moreover, the usage of activity patterns in combination with traveler types provides an strict limit on the diversity of trip patterns that can be taken by the population. Expanding the traveler types as well as activity patterns would allow for more realistic representations of travel movements, especially for subgroups of the populations like college students, workers, the elderly, and so on. Other type trips are also quite limited in that they are not segregated by purpose. Doing so would also enhance the realism of the model.

4.2 Run Time

One significant issue with the model is the amount of time required to calculate each model in a given state. The first issue with run time is the number of files is the sheer volume of files needed to be read. In Module 2, information on the number of employers within an industry and a county is needed to assign a specific employer, which can be up to 20-30MB in size for larger counties. To assign an employer, the Journey-to-Work census provides information on commuting that lists anywhere from 1 to over a hundred different commutable counties from a given residence county. In the model’s current implementation, Module 2 reads through each resident and assigns them a workplace, but there is no guarantee on grouping for given residents. One resident read in from the Alabama state file could reside and be assigned a workplace in Mobile County, AL, but the next resident could work in New York County, NY. As such, there is significant amounts of re-reading and caching of files for a given state simulation. These same issues exist in Module 3 (although to a lesser extent) as well as Module 5. An ideal solution to this would be to perform some sort of intermediary assignment and regrouping of residents by their county of work, as opposed to their county of residence, in order to minimize re-reading of files.

Another significant issue with the model is the amount of time spent constructing
distributions for selection of workplaces, schools and Other type trips. While determining the distance between two longitude-latitude pairs takes nearly a millionth of a second using the Great Circle distance formula, each resident can require thousands of distance calculations. Creating the CDF of likelihoods for a thousand employers for a given worker requires a thousand calculations. While this does not mean much for areas with few employers, the run time greatly increases in urban areas. For many dense counties, there are several thousands of employers within an industry with many hundreds of thousands of residents. This issue is exacerbated in large states (in terms of population density). While Montana took only 4 CPU hours in Module 5, California and Texas each took over 150 CPU hours. To overcome this issue, several solutions have been tested. The first involved reducing the number of employers sampled at maximum. While this certainly improved run time, this solution severely impaired the integrity of the model and was rejected. The second solution, pioneered by Matthew Garvey, involved using an alternative K-D Tree data structure to perform quick computations on the nearest k - employers. At high k, this solution provided a good balance between the current implementation - examining all of the employers - as well as limiting run time. Other possible solutions included pre-building distributions and grouping residents as to prevent unnecessary generation of distributions.

One final avenue of approach for Run Time issues involves refining the implementation of the Modules. Mufti’s earlier implementations were written in Microsoft Access and Python and were generally written in a procedural manner. Gao furthered Mufti’s original implementation by extending modal-split to the Amtrak system within New Jersey and prototyping pixelization. Wyrough and Garvey updated Mufti’s implementation by introducing Object-Oriented Design when possible to simplify computation and provide further generalization, as well as prototyping alternative data structures to help decrease run time. This iteration of the model rewrote ambiguities in the earlier Python iteration by updating from Python 2.7 to Python 3.5 as well as providing further Object Oriented Design. Additionally, R scripts were written to offload ‘hot’ function calls in Python and improve run time.
Further experimentation in parallelizing code was done in Module 1, where the entire Python implementation was rewritten in MATLAB and parallelized, providing run time increases of 3-4x over the Python predecessor. Further iterations should examine the alternative data structures proposed by Wyrough and Garvey in addition to continuing the parallelization of Modules started in this iteration.

4.3 Memory

The size of the data generated by the Module also poses certain challenges. On average, the final trip files created from Module 7 are sized at roughly 154 bytes per traveller and is approximately 150 GB in size. Each Module also produces a smaller subset of this data and care must be taken at each step to ensure proper space exists to continue with the trip generation process. While steps were taken to increase ease-of-use for the end user, the vast size of the entire dataset makes handling a challenge. Pointers are created within the final dataset that reference back to the original synthetic population as to link people to their daily travel. To properly analyze and interface with this data, SQL or alternative data warehousing techniques should be used to ensure quick data access.

4.4 Statewide Interdependency

As mentioned in Chapter 2, the modules are run on a state by state basis. Essentially, this means that every state is run independently of every other state. While doing this provides great simplification in caching memory and computations necessary, it also creates discontinuities in interstate travel by ignoring the travel that occurs with respect rates of employment, patronage and school enrollment. On state borders, workplaces that are full in one state are regenerated in the next state as empty. This creates overlapping demand for many employment, patronage and school places that is unrealistic, given the data. While doing every state in parallel is unlikely to be achievable, running states in clusters around borders would be an effective way of solving this problem.
4.5 Next Steps

There are several avenues of approach for improving the model used. The first (and most obvious) approach would be to fix the limitations and assumptions noted above. Many of these areas will require non-trivial solutions that will undoubtedly provide greater accuracy and faster generation of demand. However, beyond that, additional features could be added to improve the model. The most readily available addition would be the expansion of Submodule 5.5. While airports are the only alternative mode to aTaxis, alternative modes of transport such as public bus or rail transport could be readily added. Expansion onto the Amtrak rail network would prove an interesting case study as the data for such a modal flow is already available. An additional avenue of approach would be the introduction of network flows. The demand currently generated is not selected to run on any road network. While one can generalize distance calculations when analyzing the cost of aTaxi implementations, placing the aTaxis onto the existing U.S. road network would allow further analysis to be conducted, especially with regards to traffic flows. Kornhauser’s CoPi-lot network could prove useful for implementing such a solution. Finally, further refinements of the pixelization system would allow for more accurate pixel-to-pixel aTaxi rebalancing analysis. As mentioned in the assumptions, the current pixelization system uses pixels that are skewed as one moves further from the origin point. While the current origin point is centered in Southern New Jersey, origin points near the center of the U.S. proved problematic as well. Further improvements might be to take a grid-based approach as in the Universal Transverse Mercator projection co-ordinate system. Work would need to be done to assist in dealing with long-range trip rebalancing but this approach appears to be promising for shorter trips.
Bibliography


[17] U.S. Census Bureau. Table 1: Residence County to Workplace County Flows for the United States and Puerto Rico Sorted by Residence Geography. URL [http://www.census.gov/population/metro/data/other.html](http://www.census.gov/population/metro/data/other.html)


[27] InfoGroup. ReferenceUSA Business Database. URL http://resource.referenceusa.com/