SYNTHETIC GENERATION OF INDIVIDUAL VEHICLE-BORNE PERSON TRIPS THAT CHARACTERIZE THE INDIVIDUAL MOBILITY ACROSS THE UNITED STATES ON A TYPICAL DAY

Kyle Marocchini, Corresponding Author
Department of Operations Research and Financial Engineering
Princeton University
229 Sherrerd Hall (ORFE Building)
Princeton, NJ 08544
T: +1 850 737 0956
Email: kyletm@princeton.edu

Alain Kornhauser, Ph.D.
Professor, Department of Operations Research and Financial Engineering
Princeton University
229 Sherrerd Hall (ORFE Building)
Princeton, NJ 08544
T: +1 609 258 4657 F: +1 609 258 1563
Email: alaink@princeton.edu

Word count: 5,974 words text + 5 tables/figures x 250 words (each) = 7,474 words

Submission date: August 1, 2016

ABSTRACT

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. Our demand model improves upon earlier work done by Talal Mufti in 2012, Jingkang Gao in 2013 and Chris Brownell in 2014 to expand 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.

INTRODUCTION

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 (*1*).

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 (*2*) at "a level of detail which no survey can provide" (*1*). 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 Brownell.

A Brief History of 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" (*4*) as to more accurately forecast changes in travel demand.

The earliest transportation models were based off of Lowry's Model of Metropolis (*5*), 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 (*2*). While these early models provided sufficient information to examine the impacts of broad landuse policies and aggregated regional travel demand, these models provided output that "lacked richness of detail" (*6*) and ignored fundamental travel assumptions (*2*).

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" (*2*) 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 (*6*), such as home, work and school in combination with other locations and attempts to sequence them to construct a tour which reflects that person's personal attributes, including employer, income, familial relations and so on. CEMDAP (*7*) and ALBATROSS (*8*) are recommend for further reference on Activity-Based models.

METHODOLOGY

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 US in order to be able to somewhat accurately simulate how well various operational implementations of autonomousTaxis (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 "simply" observe each person's travel tour on some representative day, say a Tuesday or Wednesday in October. Unfortunately, just counting everyone every 10 years is a substantial undertaking. Observing and recording everyone's travel tour on one day adds substantial additional 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. The first step in the process constructs a population of individuals whose ensemble reflects the spatial and demographic characteristics of the entire US population as depicted in the latest Census. Individuals are generated with traits ranging from age and gender to household location, personal income and employment sector. The next step takes in workforce participation, school enrollment and employer size and sector data to assign workers to workplaces and school aged children to schools. Using behavioral patterns such as truancy, vacation and illness distributions, a mobility tour is assigned to each member of this virtual population. Finally specific personal activity destinations that seek to deliver sufficient clientele to the existing distribution of activities/land uses (movies, restaurants, shops, etc.) complete each individual tour stop. Trips are constructed between different anchor points, defined as school, home, work and other activity, 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 US.

Module 1: Generation of a Synthetic Populace

To begin the process, it is first necessary to generate the entire population of the US, roughly spanning over 300 million people. At its core, this step constructs the US'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. (*9*) 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, 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 (Figure 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 data sources. 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 and assigns residents a specific Traveler Type. Traveler Types are derived from specific population attributes and are used later in the model to generate various trip tour patterns for residents when constructing trips. Traveler types 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. Readers with further interest on these assumptions are referred to Mufti (*3*). 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 US do exist).

The final step in Module 1 is to ascribe an income to every individual. To do so, every household, familial or non-familial, is assigned an income and income code (Figure 1) based on distributions of household income from the American Community Survey (*10*). Each household's income is then appropriately distributed to the household members, provided they should earn an income.

Traveler Type Code	Traveler Type Name	Classification (Age; HHT)
0	Does Not Travel	0-5,79+; 2, 3, 4, 5, 7
1	Student Non-Worker	5-15,16-18 * 99.81%
2	Student Worker In County	16-18 * 0.193%
3	College No Commute	18-22 * 90.34%
4	College Worker In County	18-22 * 9.66%
5	Typical Traveler	22-64 * 78%
6	Home Worker Traveler	22-64 * 22%

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

FIGURE 1 Traveler Type, HHT and Income Codes used in Module 1.

Module 2: Workplace Assignment

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 are used to identify specific counties and county-level 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, 4 and 5 are valid workers. Once this is completed, valid workers are assigned specific counties of work using the Journey-To-Work (JtW) Census (*11*). The relative likelihood that a resident travels to a given county of work from their home county is constructed from a

normalization of the JtW. With the likelihoods generated, a cumulate distribution function (CDF) was constructed and the county was sampled without replacement.

The next step is to assign a worker to an industry. To categorize industries, the U.S. Census Bureau's 2012 North American Industry Classification System (NAICS) was adopted. To determine the industry of work, it is necessary to know the gender, income and county of work for a given worker, in addition to the distribution of employment by gender by industry for the given county of work. For the distribution, a U.S. Census Table on "Industry by Sex and Median Earnings" (*13*) was used, relating industry participation rate by gender and median income within an industry for every county. A Gravity Model (1) was employed to generate the likelihoods that a worker would be employed within a given industry for a given working county. As the name implies, the Gravity Model presents a measure of attraction to a location, increasing with the popularity of a location and decreasing with distance squared. In (1), i is the index of the worker's work county. Inc represents the income of this worker. K is the set of all 20 NAISC industries, with k representing one of the industries in this set. $MedInc_k$ represents the median income of workers who have the same gender as the worker specified in the industry k. $E_{i,k}$ represents the number of employees in county i of the same gender as the worker specified who work in industry k. $W_{i,k}$ represents the likelihood that a worker will select a given industry k in the worker's work county i and is determined by (1). With the likelihoods generated, a CDF was constructed and the industry was sampled with replacement.

$$W_{i,k} = \frac{\frac{E_{i,k}}{(Inc - MedInc_k)^2}}{\sum_k \frac{E_{i,k}}{(Inc - MedInc_k)^2}} \qquad \forall\, k \in K \tag{1}$$

The final step of Module 2 is to associate each worker, to a specific employer. Datasets listing all employers and respective patronage data for each state were assembled from the 2012 ReferenceUSA Businesses dataset (*14*) based off of earlier assessments made by Gao (*15*). While not complete, the dataset was comprehensive enough to cover the majority of businesses within the US. The relative likelihood that a resident is employed at a given workplace is constructed from a normalization of the ReferenceUSA employment ratios. With the likelihoods generated, a CDF was constructed and the workplace was sampled with replacement.

With Module 2 complete, every eligible worker in the population generated has an assigned workplace. These workplaces play a key role as one of the three essential anchor points in trip tours. With the first anchor, housing, already complete and the second anchor, workplaces, now complete, the third anchor, schools, can now be determined.

Module 3: School Assignment

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.

Students before kindergarten are not represented as no publicly available comprehensive dataset exists documenting their distribution on a county-level. Students from K-12 are divided

between public and private schools, leveraging a dataset provided by the National Center for Education Statistics (NCES) (*15*) detailing participation in school type by age. Students involved in post-secondary education are divided amongst non-degree granting programs, two year programs and four-year institutions. 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.

As mentioned previously, a Gravity Model (2) was used to generate the likelihoods that a student would attend a school in a given county. In (2), *h* is the index of the student's home county. *C* is the set of all counties, indexed by *c*, that are geographically adjacent to the student's home county, *h*. $X_{c,j}$ indicates the enrollment in school *j*, in county *c*. $D_{h,c}$ represents the distance between the geographical centroid of county *h* and *c*. In the special case where $c = h$, $D_{h,c}$ is taken to be $0.75 * min_{c}\{D_{h,c}\}$. $W_{h,c,j}$ represents the likelihood that a student will attend a given school *j* in county *c* and is determined by (2).

$$W_{h,j} = \frac{\frac{X_{c,j}}{D_{h,c}^2}}{\sum_j \frac{X_{c,j}}{D_{h,c}^2}} \qquad \forall\, c \in C \tag{2}$$

One special caveat encoded in (2) 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 own county. While 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.

Individual schools for each student are determined by first assigning a school type to each student. School types are a combination of private/public and elementary/middle/high for K-12 and 4 year/2 year/non degree for post-secondary education. This is done similarly as in Module 2, where type is established by normalizing the number of students in NCES/Business datasets. In this case, sampling is done with replacement. Once a type has been established, the nearest school to the student's home is selected.

With Module 3 complete, the entirety of the simulated population has been placed in households and workplaces and schools, if eligible. The three anchors of our travel demand have thus been completely spatially determined. It now remains to determine those trips that are not encompassed as one of these anchor points, defined earlier as Other type trips.

Module 4: Tour Construction and Activity Pattern Assignment

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. Activity patterns (Table 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 US, between three and four trips.

TABLE 1 Activity Patterns for Module 4

Tour	Trip Breakdown	Total Trips
0	H	0
1	H-W-H	2
2	H-S-H	2
3	H-O-H	2
4	H-S-W-H	3
5	H-W-S-H	3
6	H-W-O-H	3
7	H-S-O-H	3
8	H-O-O-H	3
9	H-S-W-O-H	4
10	H-W-S-O-H	4
11	H-W-H-O-H	4
12	H-S-H-O-H	4
13	H-O-H-O-H	4
14	H-W-O-W-H	4
15	H-W-O-H-O-H	5
16	H-S-O-H-O-H	5
17	H-W-H-O-O-H	5
18	H-S-H-O-O-H	5
19	H-W-O-H-O-H-O-H	7
20	H-S-O-H-O-H-O-H	7

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.

Module 5: Trip Destination Assignment

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 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 this end, Module 5 employs a Gravity Model (3) to determine the likelihoods that a traveler visits an O trip destination. J represents the set of all O trip destinations, indexed by j. $P_{h,j}$ represents the patronage of destination j. $P_{h,j}$ represents the distance between the geographical centroid of the worker's census block, h, and the destination j. $W_{h,j}$ represents the likelihood that a worker will select a given destination j and is determined by (3). With the likelihoods generated, a CDF was constructed and the destination was sampled without replacement.

$$W_{h,j} = \frac{\frac{P_{h,j}}{D_{h,j}^2}}{\sum_k \frac{P_{h,j}}{D_{h,j}^2}} \quad \forall\, j \in J \tag{3}$$

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. Tours that include the W-O-W trip sequence are restricted within 5 miles of the workplace.

With Module 5 complete, the actual destinations of every W, S, O and H trip are determined. A node-based data structure is used to output the daily tour for a given person. Each W, S, W and O trip is considered to be a node within the daily trip tour. Module 5 outputs a latitude-longitude tuple to identify the location of each node, its type, its name, a NAISC Industry Code and pointers to each node's predecessor and successor.

Submodule 5.5: Modal Split

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 a 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 FAA on US 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 (5) is employed, where $P_{h,j}$ becomes A_j and represents the land area of airport j.

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.

The modal-split used within submodule 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.

Module 6: Arrival and Departure Time Assignment

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 (*6*).

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.

With Module 6 complete, the travel demand for the entire US is both spatially and temporally distributed. 51 .csv files are constructed for the 50 States and the District of

Columbia, constructed row-by-row (Figure 2). The final node is omitted as for all trips, less those handled by Submodule 5.5, it is the home of the resident.

Residence State	County Code	Tract Code	Block Code	HH ID	Person ID	Activity Pattern
01	001	20802	2039	13894	0100037350	19

Node 1 Type	Node 1 Predecessor	Node 1 Successor	Node 1 Name	Node 1 County	Node 1 Latitude	Node 1 Longitude	Node 1 Arrival Time	Node 1 Departure Time
H	H	W	Home	01001	32.49587	-86.47866	0	24047.47
Node 2 Type	Node 2 Predecessor	Node 2 Successor	Node 2 Name	Node 2 County	Node 2 Latitude	Node 2 Longitude	Node 2 Arrival Time	Node 2 Departure Time
W	H	O	AL HOLMES TAXIDERMY STUDIO	01051	32.44509	-86.14609	26413.94	59703.96
Node 3 Type	Node 3 Predecessor	Node 3 Successor	Node 3 Name	Node 3 County	Node 3 Latitude	Node 3 Longitude	Node 3 Arrival Time	Node 3 Departure Time
O	W	H	OLIVE GARDEN ITALIAN RSTRNT	01001	32.45872	-86.47866	62070.43	68938.06
Node 4 Type	Node 4 Predecessor	Node 4 Successor	Node 4 Name	Node 4 County	Node 4 Latitude	Node 4 Longitude	Node 4 Arrival Time	Node 4 Departure Time
H	O	O	Home	01001	32.49587	-86.47866	69591.72	70497.46
Node 5 Type	Node 5 Predecessor	Node 5 Successor	Node 5 Name	Node 5 County	Node 5 Latitude	Node 5 Longitude	Node 5 Arrival Time	Node 5 Departure Time
O	H	H	BELTONE HEARDING AID CTR	01001	32.46793	-86.45132	70795.67	73167.01
Node 6 Type	Node 6 Predecessor	Node 6 Successor	Node 6 Name	Node 6 County	Node 6 Latitude	Node 6 Longitude	Node 6 Arrival Time	Node 6 Departure Time
H	O	O	Home	01001	32.49587	-86.47866	73465.22	74483.25
Node 7 Type	Node 7 Predecessor	Node 7 Successor	Node 7 Name	Node 7 County	Node 7 Latitude	Node 7 Longitude	Node 7 Arrival Time	Node 7 Departure Time
O	H	H	PRATT PUB & OYSTER BAR	01001	32.46039	-86.42431	74962.48	80830.22

FIGURE 2 Example output from Module 6, sequentially ordered by index.

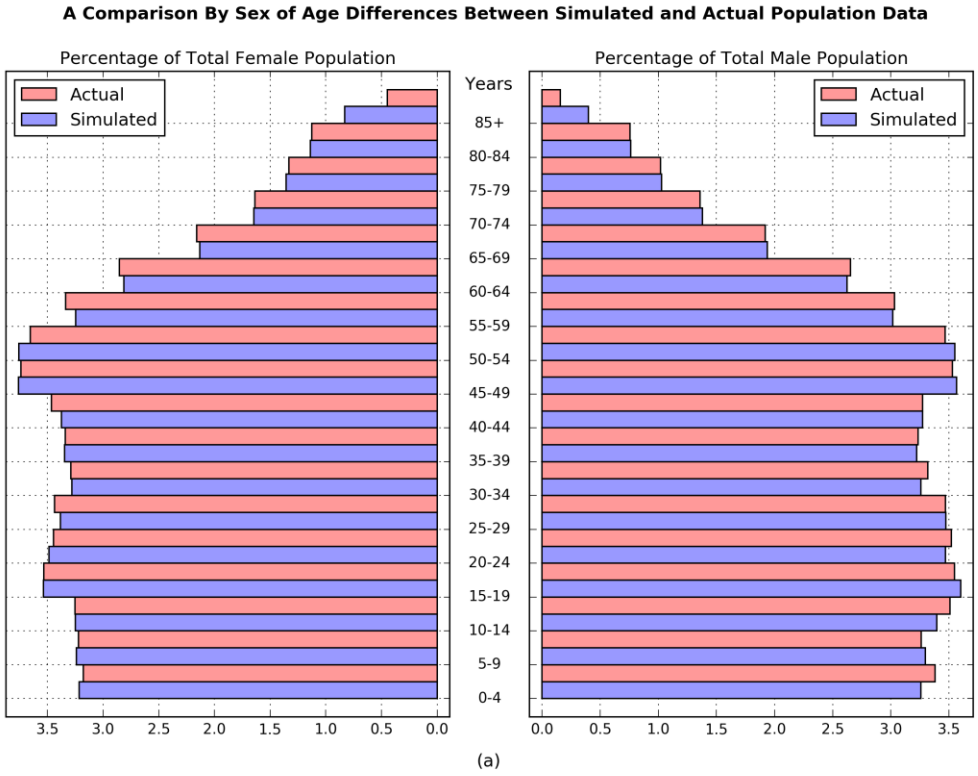
RESULTS

Demographics

Population generation is a task that is applied uniformly to every state. The total population generated, 308,745,538, is the same as it was in 2010. However, randomness is necessary to overcome aggregated information in the Census, including age and income distributions. To verify the results, it is necessary to briefly compare simulated and actual demographics with respect to age and income. An Age Pyramid is used in Figure 3(a) to compare the national simulated and actual 2010 age distributions. The average absolute percent difference between each respective age bracket is 2.73% for males and 2.91% for females. While the Age Pyramid demonstrates the effectiveness of the model is simulating 2010 data, the model's reliance on 2010 Census Data for population generation makes it difficult to extend the population to future years.

Household and personal income assignment is an exceptionally important task as income plays a fundamental role in determining a worker's industry. Figure 3(b) demonstrates the effectiveness of the model in representing incomes. As income data was aggregated to Tract-level, as opposed to Block-level, Figure 3(b) clearly shows that this necessary departure in scope did not significantly harm the generation of incomes. The mean and median household incomes of Alaska (*11*) are $82,091 and $66,521 respectively, while the synthetic Alaskan population have mean and median incomes of $98,774 and $67,532. The disparity in mean household

incomes can be explained by the highest bracket in the data (>$200,000). Incomes were constructed by sampling uniformly between income brackets, and for the last bracket, incomes were sampled from $200,000 to $1,000,000, which was clearly not representative of the actual distribution of incomes within this bracket. However, the synthetic median's close proximity to its actual counterpart reaffirms the success of the model. Moreover, as the population generation is applied uniformly state-by-state, the model's success in creating a realistic Alaskan population can be generalized to the 308,745,538 persons throughout the 50 States and D.C.



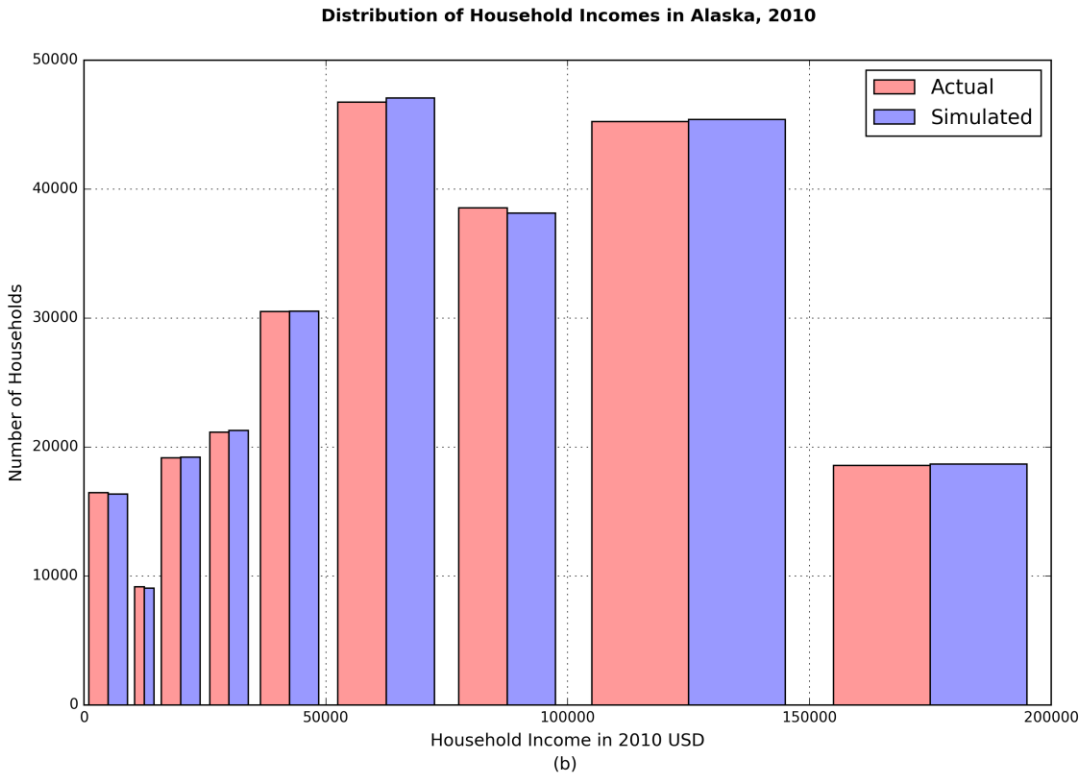


FIGURE 3 Comparisons for (*a*) U.S. Age Distributions and (*b*) State Income Distributions.

Home-Work Trips

Home-Work (H-W) trips comprise nearly 16% of trips taken by the synthetic population and allow us to examine trip length and spatial distributions. Figure 4 presents a cumulative distribution of trip length for all H-W trips taken in Florida, Georgia and Hawaii. The respective trip means in were 15.12, 13.43 and 19.29 miles, while the respective trip medians were 12.13, 9.08 and 15.23 miles. The cumulative distributions show patterns typical of gravity model

attraction and also reflect the infrastructure and geographies of each state, especially for Hawaii.

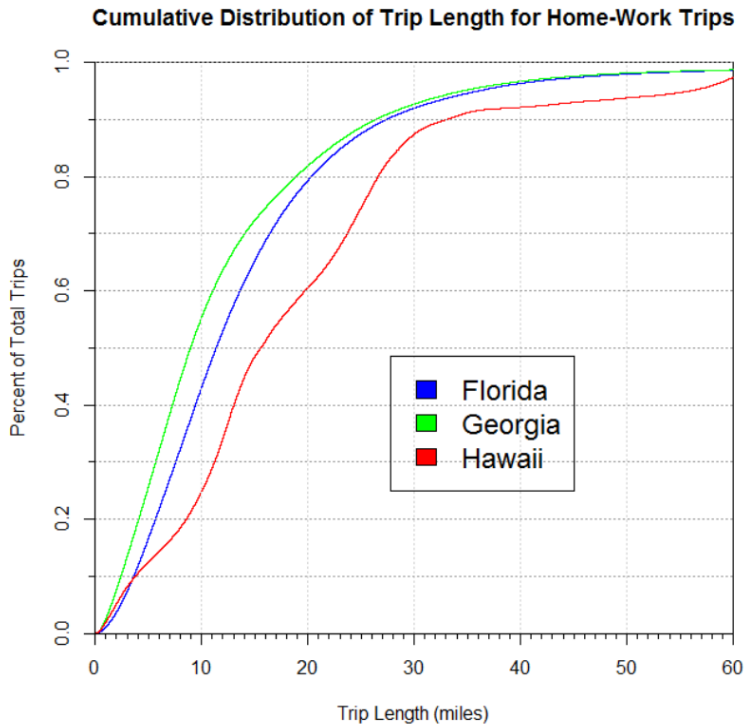


FIGURE 4 Cumulative Distribution of Trip Lengths for Florida, Georgia and Hawaii.

As commute time is the standard metric analyzed by the U.S. Census, trip length must be translated to trip time to validate model findings. To do so, trip lengths are roughly converted by assuming a constant 30 mile per hour commute speed, or 2 minutes per mile. The 2009 national average commute time was 25.1 minutes, with an average commute trip length of 12.09 miles (*11*), while the simulated national average is 14.03 miles. As this model does not incorporate road circuitry in distance, this average is certainly higher than the survey. However, the simulated national median commute is 12.68 miles, with a translated commute time of 25.42 minutes, which fits well with the data and suggests that the distribution of trip lengths is distorted by outliers that would otherwise be satisfied by alternative modes of transportation. Moreover, as Census estimates indicate that nearly 10% of trips take over 60 minutes, one might expect that the length of such a commute would be around roughly 30 miles. The 90% percentile of trips for Florida, Georgia and Hawaii are all within 25-35 miles, as are the rest of the states, verifying the

empirical data.

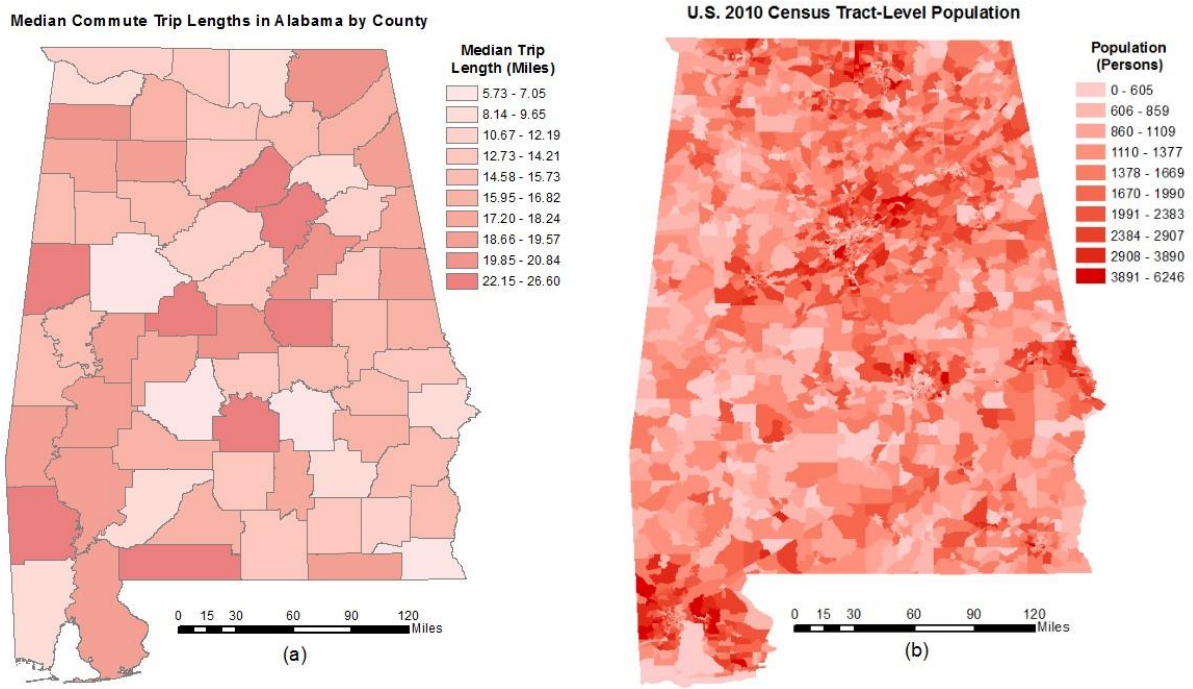


FIGURE 5 A comparison of county-level median trip length (*a*) to tract-level population distribution (*b*) in Alabama

However, state-level statistics do not reveal the whole story; even amongst states, there exists a great deal of variation in trip lengths (Figure 5). Counties with shorter commutes tend to be near large urban centers, including Mobile, Birmingham and Montgomery. As gravity models are employed, this bias towards short trips is expected for high attraction counties. Moreover, those counties adjacent to urban centers tend to have longer commutes, as can be seen most acutely with Huntsville and Montgomery. This attraction towards large cities is not only consistent with what the model should produce, given the assumptions, it is also consistent with the reality of commuting throughout the nation.

CONCLUSIONS

While more thorough analyses must be conducted on the output of the model presented to completely verify the efficacy of the assumptions and attraction metrics employed to generate trips, early results appear promising. The model's ability to comprehensively simulate national travel demand for a given workday provides planners with a new tool to forecast demand. For emerging mobility services like aTaxis, travel demand forecasting will play a key role in determining the operational structures necessary to satisfy the nation's travel demand. A failure to bring travel demand forecasting and operational implementations forwards risks stifling the enormous potential of autonomous vehicles. The data set produced by the model enables the first, albeit simplified, insight into the individual travel demand of US Residents and provides a framework to examine how one might use aTaxis to satisfy this demand.

REFERENCES

1. Brownell, C. and A. Kornhauser. A Driverless Alternative: Fleet Size and Cost Requirements for a Statewide Autonomous Taxi Network in New Jersey. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2416*, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 73–81.
2. Castiglione, J., M. Bradley, and J. Gliebe. Activity-Based Travel Demand Models: A Primer. Publication FHWA-RD-01-113. Transportation Research Board of the National Academies, Washington, D.C., 2014.
3. Mufti, T. Synthesis of Spatially & Temporally Disaggregate Person Trip Demand. MS Thesis. Princeton University, January 2012.
4. Ben-Akiva, M. E and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, Cambridge, MA, 1985.
5. Lowry, I. S. *A Model of Metropolis*. Publication RM-4035-RC. RAND Corporation, 1964.
6. Lee Jr., D. B. A Requiem for Large-Scale Models. In *Journal of the American Institute of Planners*. Vol. 39, No. 3, 1973, pp 163-178.
7. Santos, A., N. McGuckin, H.Y. Nakamoto, D. Gray and S. Liss. *Summary of Travel Trends: 2009 National Household Travel Survey*. Publication FHWA-PL-11-022. U.S. Department of Transportation, 2011.
8. Bhat, C.R., J.Y. Guo, S. Srinivasan, and A. Sivakumar. A Comprehensive Econometric Microsimulator for Daily Activity-Travel Pattern. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1894*, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 57–66.
9. Arentze, T.A., H.J.P. Timmermans. Albatross: A Learning-based Transportation Oriented Simulation System. EIRASS, Eindhoven University of Technology, Eindhoven. 2000.
10. *2010 Census Summary File Segment 1 Technical Documentation*. U.S. Census Bureau, Washington, D.C., 2012.
11. *ACS 2008-2012 5-Year Summary File Technical Documentation*. American Community Survey (ACS) Office, Washington, D.C., 2012.
12. U.S. Census Bureau. *Table 1: Residence County to Workplace County Flows for the United States and Puerto Rico Sorted by Residence Geography*. http://www.census.gov/population/metro/data/other.html. Accessed June 5, 2016.
13. U.S. Census Bureau Population Estimates Program. *Industry by Sex and Median Earnings in the Past 12 Months (in 2012 Inflation-adjusted Dollars) for the Full-time, Year-Round Civilian Employed Population 16 Years and Over*. http://factfinder.census.gov/. Accessed June 10, 2016.
14. ReferenceUSA. *U.S. Businesses Database*. http://www.referenceusa.com. Accessed June 17, 2016.
15. Gao, J. A Disaggregate Transportation Demand Model for the Analysis of an Autonomous Taxi System in New Jersey. Princeton University, January 2013.
16. *Documentation to the NCES Common Core of Data Public Elementary/Secondary School Universe Survey: School Year 2010-2011*. U.S. Department of Education, Washington, D.C., 2012.