CAMMSE

Center for Advanced Multimodal Mobility Solutions and Education

Project ID: 2017 Project 10

THE EFFECT OF COMPETITION OF TRANSPORT MODES ON MOBILITY

Final Report

by

Washington State University
Consortium Member

Jia Yan, Ph.D. (ORCID ID: https://orcid.org/0000-0003-4132-7504)
Associate Professor, School of Economic Sciences
Hulbert 101, Pullman, WA 99164
Phone: 1-509-335-7809; Email: jiay@wsu.edu

for

Center for Advanced Multimodal Mobility Solutions and Education
(CAMMSE @ UNC Charlotte)
The University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223

July 2018

ACKNOWLEDGEMENTS

This project was funded by the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE @ UNC Charlotte), one of the Tier I University Transportation Centers that were selected in this nationwide competition, by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT), under the FAST Act. The authors are also very grateful for all of the time and effort spent by DOT and industry professionals to provide project information that was critical for the successful completion of this study.

DISCLAIMER

The contents of this report reflect the views of the authors, who are solely responsible for the facts and the accuracy of the material and information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation University Transportation Centers Program in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The contents do not necessarily reflect the official views of the U.S. Government. This report does not constitute a standard, specification, or regulation.

Table of Contents

List of Figures

List of Tables

EXCUTIVE SUMMARY

Despite the ever-increasing demand, controversies have been surrounding the ride-hailing industry since the day of its rise. Tighter government regulation or even banning is called around the world. In this paper, we address the issue by designing a quasi-experiment and estimate how much Uber benefits consumers in a creative way. Using three datasets created before and after Uber service availability, and dividing San Francisco the studied area into grids of 4km^2 each, we are able to investigate consumer commuting behavior at an individual level and find out Uber brings out at least $0.76 gains per commuter per trip and generates an annual consumer surplus of $100 million in San Francisco.

The three datasets include the National Household Travel Survey Data from 2008 to 2009 when Uber service was not yet available, the origin-destination level Uber itinerary data and Google map data of 2017. We first use NHTS data to identify consumer preference in 2008 under a discrete choice framework. We then construct counterfactual scenarios in which Uber becomes an option with Uber and Google data, and find out the consumer surplus changes Uber brought.

Chapter 1. Introduction

1.1 Problem Statement

Uber is a startup company that operates a technology platform connecting driver-partners and riders. Although displacing traditional jobs on one hand, Uber generates new jobs and powers billions in economic impact in cities around the world. Previous studies focus only on Uber's influence on the efficiency of the transportation system

1.2 Objectives

Our paper focuses on passengers and investigates the consumer surplus that Uber brings out.

1.3 Expected Contributions

Using the three datasets of survey and real consumer travel itineraries, for the first time, we are able to calculate the consumer surplus and quantify how many benefits to individual consumers Uber generates. Methodologies employed in this paper can be applied to studies about other car-hailing platforms and findings can help policy makers make better decisions among all controversies.

1.4 Report Overview

The remainder of this paper is organized as follows. In section 2, we describe the details of data and our design of data collection. Section 3 is our empirical model. Section 4 presents the results of our estimations, demonstrates the way we calculate consumer surplus and the calibration of some unknown parameters. Last section is the conclusion.

Chapter 2. Literature Review

2.1 Introduction

In this report, we design a quasi-experiment with data on San Francisco. In order to estimate the mode choice model, we divide the entire San Francisco area on Google Map into grids of 4km2 each. Each grid can be an origin or a destination or both, therefore an origin and destination pair (thereafter OD pair) is defined by either two grids (origin and destination are in different grids) or one grid (origin and destination are in the same grid). There are a total of 67,280,000 OD pairs in San Francisco.

Our three datasets include the National Household Travel Survey Data from 2008 to 2009, the origin-destination level Uber itinerary data and Google map data of 2017. The NHTS data contains five files including information on households, commuters and vehicles when Uber service was not yet available. The most important one is the Travel Day Trip file which records trips occurred during a 24-hour period for a particular individual. This individual level data allows us to estimate commuter's preference in a discrete choice model. Uber Technologies provided the grid-based Uber itinerary data which contain trip attributes such as average trip duration and average fare between grids in September 2016.

We collect the Google data by requesting mode attributes for each OD pair from Google Map APIs during a period of two months from January 1st to February 28th 2017. As show in Appendix Table A9, with the input of the origin and destination, Google Map returns data on the trip duration and distance. In the whole process, we used three web servers and wrote a spider crawl program to help us collect the data.

Using aforementioned three datasets, we take the following steps to calculate the consumer surplus. First, we use the NHTS data and the conditional logit model to estimate commuters' heterogeneous preference in San Francisco. Second, assuming consumer preference doesn't change over time, we calculate consumer surplus in 2017 when consumer face mode choice set constructed by combining Uber data and Google data. Third, we re-calculate consumer surplus presuming that Uber is unavailable in 2017. Finally, we take the difference between step 2 and step 3, which gives the consumer surplus that each commuter gains per trip. To further investigate the heterogeneous consumer surplus, we calculate the surplus for sub-population of commuters.

We find that the average gain in consumer surplus due to the availability of Uber ranges from 0.76 to 2.85 dollars per person, depending on which conditional logit models to use to reveal consumer preference. Moreover, results show that weekend commuters gain more consumer surplus than weekday ones, so do the non-peak hour passengers than peak hour passengers. That is mainly because it takes longer time and costs more to travel in peak hours and weekdays, therefore decreasing the maximized utility. The overall consumer surplus for commuters in San Francisco is around $279,680 per day or $100 million per year.
Literature on consumer surplus is profound. The calculations of consumer surplus depend on consumer's 'willingness-to-pay' and the demand curve. Since individual level data are scarce and costly to access, recent literature is constrained to use market share level data to estimate demand on differentiated products (Berry, 1994; Berry et al.,1995; Nevo, 2000; Petrin 2002; , Eizenberg, 2011). In this paper, the way we collect individual level data from Google Map enables us to measure consumer surplus in a discrete choice framework using log-sum (Small

and Rosen, 1981; Börsch-Supan, 1990; Kalmanje and Kockelman, 2004; Dagsvik and Karlstrom, 2005; Small et al, 2005; Small et al, 2006).

There are a broad strand of literature discussing the impacts of Uber on transportation industry especially in taxi business (Hall, 2015; Hall et al.,2015 Buchholz, 2016; Cohen et al., 2016). With the accessibility to almost 50 million's individual data in Uber's trip service, Cohen et al. (2016) take advantage of Uber's pricing schema (surge price) and apply the discontinuity design to identify demand on UberX service. They found that overall consumer surplus generated by UberX was around $6.8billion in 2015.

Our work contributes to existing literature in mainly two ways. First of all, rather than considering demand only on Uber as in Cohen et al. (2016), we identify consumer's uncompensated demand under a discrete choice model taking into account all choice modes a commuter may have. Our ability to achieve this derives from the richness of our data. Second, this paper also contributes to economic research by using more open data. Our ability to crawl open data from google enables us to construct the complete choice set for commuters. One of the shortcomings of this paper is we have to assume consumer's preference on mode choice doesn't change over 2009-2017. We observe individual's choice decision in 2007 through the NHTS data, but the new choice set in the environment with the availability of Uber is constructed using data of 2017.

2.2 Data

We use three datasets in this paper: the National Household Travel Survey Data from 2008 to 2009, the orgin-destination level Uber itinerary data and Google Map data of 2017. The NHTS data are from a survey conducted by Federal Highway Administration (FHWA) from 2008 to 2009. Uber data are provided by Uber Company. Google data are requested through Google MAO APIs. All data are grid-based, which enables us to merge them.

2.2.1 NHTS Data

National Household Travel Survey Data were collected through random digit dialing (RDD) telephone survey from March 2008 to May 2009. We use the 2009 SanFrancisco components of NHTS data in this report. The survey interviewed 150,147 households across the U.S. Each household was randomly assigned a specific date to call. If a household agree to participate in the initial telephone interview, it would receive in amil a travel day diary with guidance and start recording the trips. Interviews were carried out nationwide.

2.2.2 Important Variables in NHTS Dataset

The data have five data files of different levels. The household files contain records for each household, the individual files record level information, and so on. The household file and the vehicle file are linked through an 8-digit ID. An individual is linked to a household through a 10-digit ID with the first 8 digits representing the household ID number and the last 2 digits representing household member. Travel Day Trip file is organized by individual-trip level and it is linked an individual by adding 2-digit to represent number of trips in one day to person ID.

The household file includes, but not limited to, data on household ID, number of household members, number of workers in a household, number of drivers in household, derived total household income, and home address in urbanized area. Details on important variables of the household file are shown in Appendix Table A1.

The individual file includes data on an individual such as interview date, job category, and distance home to work, respondent age, respondent gender, responder education, workplace address, zip code of work location, and so on. Details important variables of the individual file are shown in Appendix Table A2.

The Vehicle file contains, but not limited to, data on vehicle make name, vehicle model name, vehicle model year, vehicle make code, vehicle make name, vehicle model code, and other vehicle relevant information. Details on vehicle file important variables are shown in Appendix Table A3.

The Travel Day Trip file is one of the most important file in the survey which record trips taken during a 24-hour period, and includes information such as mode of public transit used, travel day, trip start time, trip distance in mile, time of entire trip, number of people with respondent on trip, the purpose of trip, travel day trip destination and so on. This individual level data allows use to reveal consumer's preference in a discrete choice model framework. Details on travel day trip file important variables are shown in Appendix Table A4.

The last component of dataset is location file, which contains geographical information such as home latitude and longitude, city name, purpose of trip, trip destination latitude and longitude, work latitude and longitude and so on. Details on geographical information file important variables are shown in Appendix Table A5.

To enable empirical analysis, we equally divide San Francisco areas on Google MAP into grids. Each grid is assigned a unique ID. An Origin and Destination pair (OD Pair) is one-way-based. The following figure illustrates the way we divide San Francisco. Based on latitude and longitude, we split the entire San Francisco into 5800 (58 X 100) grids. The horizontal axis represents longitude and the vertical axis indexes latitude. The step width of horizontal and vertical axis is 0.035 and 0.03 decimal degree, respectively. The corresponding physical distance is approximate 2 miles, separately. Therefore, all trips of NHTS data can be classified into the grids according to their longitude and latitude of origin and destination. For example, in this Figure 1, one color indexes at least one trip. The left-hand side panel highlights all the origins of the trips, while the right-hand side panel denotes the destinations.

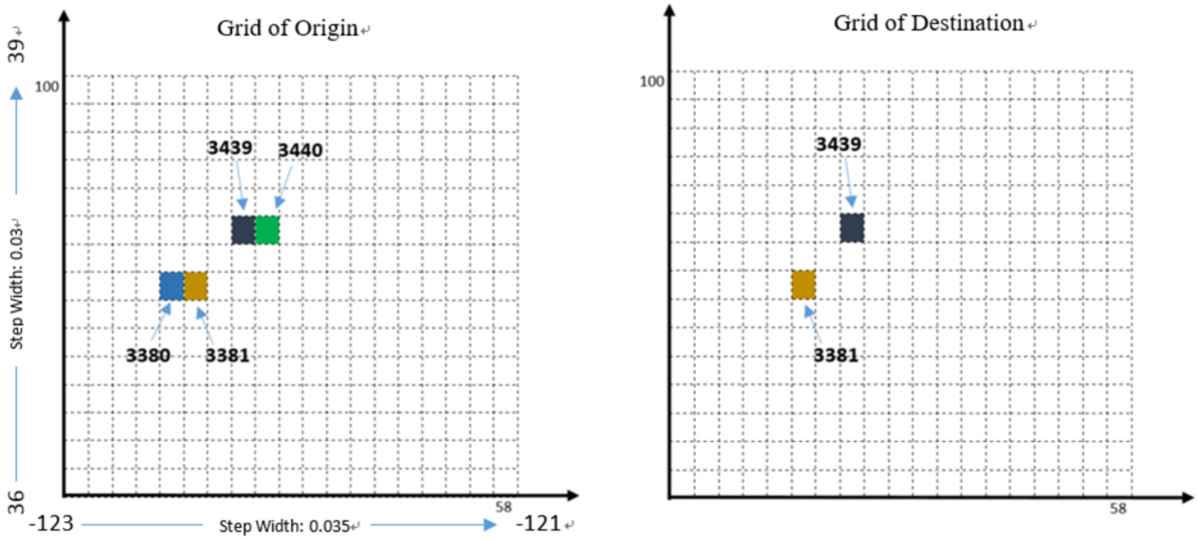
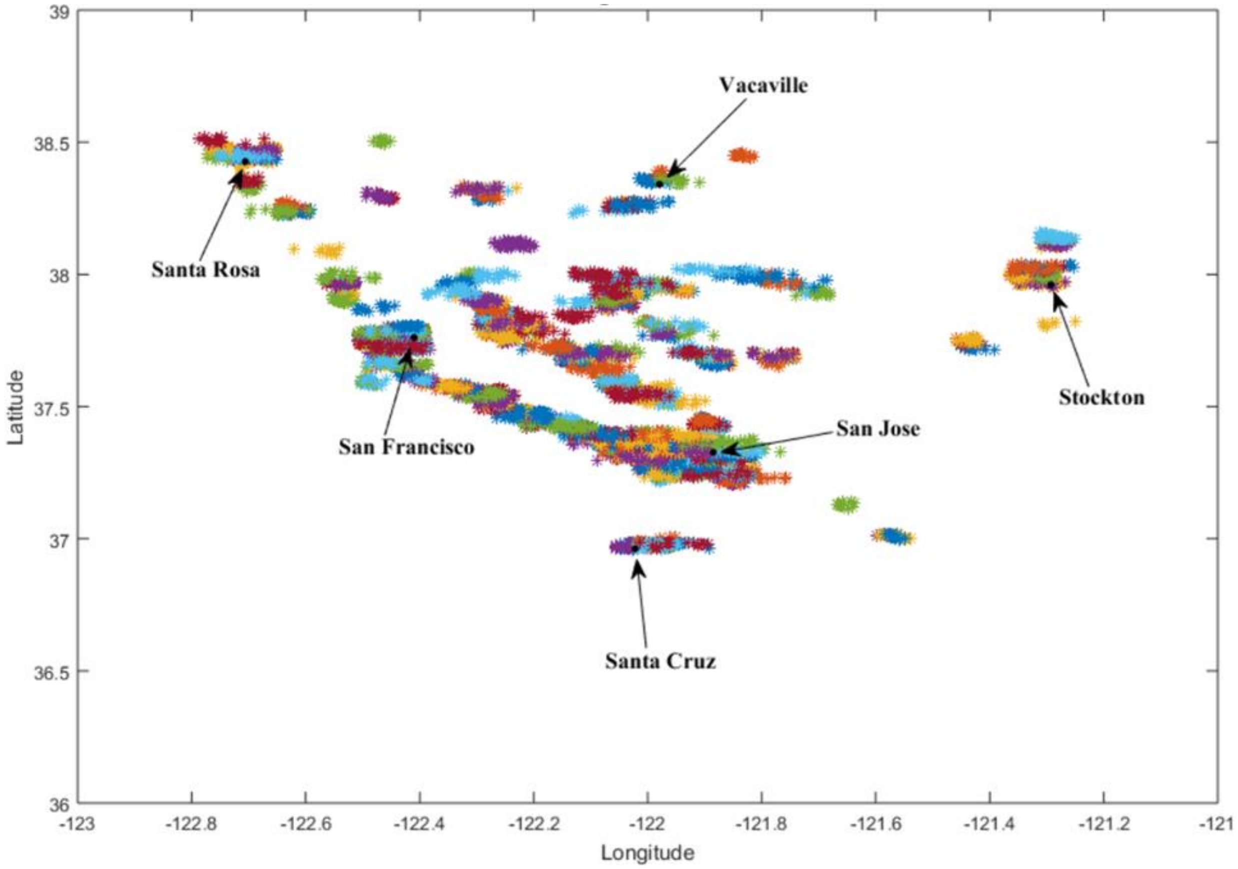


Figure 1. Grid of Origin and Destination in San Francisco

To take a glimpse of our data, we project all origin of trips into a graph by latitude and longitude, as shown in Figure 2.

According to our division rules, San Francisco are divided into 5800 grids. Theoretically, we could have 67,280,000(5800 X 5800 X 2) OD pairs. Excluding those without NHTS records and those covered by ocean or bay, we end up with 8772 OD pairs, as summarized in Table 1. Among those pairs, there are more than 20 trips in 108 OD pairs.. If the grid size is 4km^2, we have 5961 OD pairs and there are more than 20 trips in 163 of them. Table 1 also show the distribution of number of trips in 4KM^2 grids. As presented, most of OD pairs have less than 10 trips.



Note: This figure only uses the [1th-percentile, 99th-percentile] of latitude and longitude of trips.

Figure 2. The Origin of Trips in San Francisco

Table 1. Summary of NHTS Data

Total Number of OD Pairs[2]	8772
Total Number of 2 X 2 OD Pairs[3]	8772
Total Number of 2 X 2 OD Pairs with More Than 20 Trips	108
Total Number of 4 X 4 OD Pairs	5961
Total Number of 4 X 4 OD Pairs with More Than 20 Trips	163
Total Number of 8 X 8 OD Pairs	4329
Total Number of 8 X 8 OD Pairs with More Than 20 Trips	128
Total Number of OD Pairs with More Than 20 Trips	399
Distribution of Trips in OD Pairs defined by 4KM^2 Grids	
Smallest	1
1%	1
5%	1
10%	1
25%	1
50%	1
75%	2
90%	7
95%	25
99%	43
Largest	100
Mean	4.08
Std. Dev.	8.73
Number of Working Trips	1933
Number of Non-working Trips	22,878
Total Records	24,811

2.2.3 Uber Data

Uber Company provided Uber data according to our request. We illustrated the way we define our grids and expect Uber trip observation s data presented in our grid-based way. We use four points to define a grid. As illustrated in Figure 3, point A, B, C, D define a grid of origin. The ID number of this grid is #1. The interval of longitude is [-123,-121] and the step width is 0.035, which gives 58 slots. The interval of latitude is [36,39] and the step width is 0.03, which gives 100 slots. Therefore, the total number of grid of origin is 5800. Uber's database engineers take the following procedure to generate the Uber dataset:

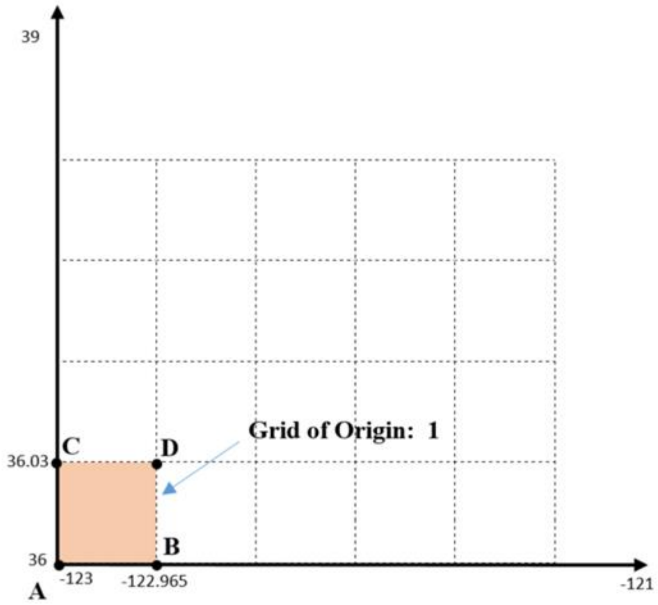


Figure 3. Points Used to Define a Grid

Step 1. Pick a grid of origin and check whether there is trip starting from the grid.

Step 2. If there is, find out the grid(s) of destination of those trip(s).

Step 3. This process continues after all trips are found out in this OD pair. Mode attributes such as travel distance, travel duration, fare of all trips are then averaged by OD pairs.

Sample Uber data are presented in Appendix Table A6.

The Uber data record trips occurred in San Francisco for one week and contains 60167 observations, which consists of 31,816 UberX service trips and 28,351 28,351 Uber Pool service. Summary statistics are shown in Table 2. The sample consists of 188 OD pairs defined by 4KM2 grid size. The mean traveling distance is 5.93 miles. The mean traveling duration is 18.05 minutes. Mean costs of commuting are 14.49 dollars.

Table 2. Summary of Uber Data

	Observation	**Mean**	**Std. Dev.**	**Minimal**	**Maximum**
Distance(Mile)	60,167	5.93	5.29	0.03	46.81
Duration(Minute)	60,167	18.05	9.52	1.33	92.33
Cost(Dollar)	60,167	14.49	9.86	4.27	110.82

2.2.4 Google Map Data

We use crawl technique to request data from Google Map API. The Google dataset is constructed by taking the following steps: First, based on the pre-defined grids as shown in Figure 1, we construct 67,280,000(5800 X 5800 X 2) OD pairs and assign each OD pair an unique ID. Second, for each OD pair, we input origin and destination (centroid of the gird) into Google Map APIs. Google Map APIs then return available mode choice and their attributes such as distance, duration and fare. If no available alternatives, APIs return none. We requested the data three times per day. The time we request data is during morning peak hour (7:00-10:00, utc), afternoon peak hour(16:00-19:00, utc) and other time. This process lasts for two month from Jan 01 2017 to Feb 28 2017.

Sample Google data are presented in Appendix Table A9Table 3. The sample consists of 179 OD pairs defined by 4KM2 grid size. Summary statistics are shown in Table 3. There are 422,092 observations of Google Map data, which consists of 278,062 peak hour observations and 144,030 non-peak hour observations. Among the data, 290,235 records occurred in weekday and 131,858 records occurred in weekends. Summary statistics of distance, duration and fare of all alternatives are also shown in Table 3. The mean of commuter distance ranges from 5.47 miles to 7.26 miles. Traveling duration ranges from half hour to 2 hours. We only have fare for bus and train. The mean costs of bus are 3.43 dollars while that of train are 3.82 dollars. It is noteworthy that the costs of Taxi ares are not returned by google map APIs, we thus calibrate the costs using data from yellow cap company. The pricing schema is as follow: $3.5 for the first 0.2 mile and after that 0.55 dollar for each additional 0.2 mile.

Table 3. Summary of Google Map Data

		Observation	Mean	Std. Dev.	Minimal	Maximum
Distance (Unit: Mile)						
	Drive	264,131	6.77	5.92	0.96	38.51
	Walk	245,436	6.01	5.68	0.93	44.62
	Bike	218,188	5.47	4.41	0.94	43.56
	Bus	238,737	7.26	6.68	0.93	76.03
	Train	233,333	6.92	6.74	0.85	89.19
Duration (Unit: Minute)						
	Drive	261,472	16.18	7.35	1.77	38.87
	Walk	242,916	114.79	96.87	18.43	606.52
	Bike	215,937	33.23	20.45	4.45	115.35
	Bus	236,286	49.00	26.52	0.25	174.15
	Train	230,931	45.13	19.44	0.18	132.05
Fare (Unit: Dollar)						
	Bus	189,215	3.43	1.49	0.70	8.50
	Train	136,454	3.82	2.29	0.70	12.15
Observations						
Peak Hour		278,062				
Non-Peak Hour		144,030				
Weekday		290,234				
Weekend		131,858				
Total		422,092				

Chapter 3. Solution and Methodologies

3.1 Economic Model

We use NHTS data to estimate commuter's mode choice. Data used in this analysis is the San Francisco component of NHTS dataset. We divide the entire San Francisco into grids by $4KM^2$ and keep OD pairs with more than 20 trips and more than 1 transportation modes, which finally gives 163 OD pairs.

The econometric model combines multinomial and conditional logit so that the utility depends not only on individual characteristics but also on alternative attributes. Specifically, the indirect utility of commuter i from choosing transport mode j is

$$u_{ij} = x_i'\alpha_j + z_{ij}'\beta + \varepsilon_{ij}$$

Where x_i is a vector of characteristics of a commuter and z_{ij}' is a vector of attributes of $j-$th alternative such as fare, time, and parking if applied, and so on. z_{ij}' is varying across choices.

Demographic variables used in the model are education dummy (1 if a commuter hold a Bachelor's degree or higher degree), family income dummy (1 if annual family income is greater than $80,000), household size dummy (1 if household has more than 2 members), age dummy (1 if age is greater than 30 and smaller than 50) and gender dummy (1 if male). The complete choice set contains four alternatives: driving, bus, taxi and train. However, the available choice set that faces a commuter varies across OD pair. The total number of available choice is the number of type of transportation modes chosen by commuters. For example, in a OD pair, if we observe commuters can only choose driving and public transit, then the choice set contains two modes. If all commuters choose driving, then driving is the only mode in the OD pair.

We only observe driving distance in NHTS data, not the costs of driving. The cost of driving is calculated based on AAA's study[1] on driving cost per mile, such as Small Sedan 46.4 cents, Medium Sedan 58.9 cents, Large Sedan 72.2 cents, Sedan Average 59.2 cents; SUV 4WD 73.6 cents; Minivan 65.0 cents. Fares of public transit and train are calibrated using google map, which are $2.25 and $2.5, separately[2].

When it comes to driving we means a commuter can choose Car, Van, SUV, Pickup truck, other truck, RV, Motorcycle or Light electric veh (golf cart). The number of trips for each transportation mode is shown in appendix Table A7. Bus contains categories of local public transit, commuter bus, school bus, charter/tour bus, city to city bus, shuttle bus, street car/trolley. Train is categorized by Amtrak/inters city train, Commuter train, Subway/elevated train.

In construction of all alternatives that a commuter might choose, the time of a mode is measured by the average value of traveling duration in a particular OD pair. For example, if a commuter's choice is driving, in constructing attributes of other available mode, such as bus, the time of bus is the average of all trips by bus in the OD pair.

3.2 Estimation of Mode Choice

Table 4 presents the results of the conditional logit model. We first estimate the model with all data. To investigate the potential different choice behaviors, we then re-estimate with only data of trip observations that occurred during 7:00AM-10:00AM and 16:00PM-21:00PM. In

[1] http://newsroom.aaa.com/tag/driving-cost-per-mile/

[2] This is the fare given by google map on 27 Dec. 2016.

both estimations, most coefficients have the expected signs. Commuters are less likely to, no matter in peak hours or not, choose bus, train and taxi, an observation which is consistent with our knowledge.

To study preference heterogeneity, we interact personal characteristic with particular alternative. Alternative dummies (Bus, Taxi, and Train) are interacted with education dummy (1 if the commuter holds a bachelor's degree), family income dummy (1 if household annual income is greater than $80,000), household size dummy (1 if household has more than 2 members), age dummy(1 if age is greater than 30 and smaller than 50) and gender dummy (1 if the commuter is male). Results are demonstrated in Table 5. Commuters from higher annual income family are less likely to take public transit. As expected, household with more than 2 members are also more likely to drive. Middle age commuters have higher probability to take bus. Although not statistically, we find household with more than 2 members are less likely to take taxi and train. The result stands when estimated with only peak hour data for estimation.

Table 4. Estimation of Mode Choice

	Model 1: All Data		**Model 2: Peak Hour Data**	
Variables	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
Bus	-2.331***	(0.116)	-2.275***	(0.144)
Taxi	-3.576***	(0.821)	-3.279***	(0.864)
Train	-3.424***	(0.278)	-3.591***	(0.335)
Trip Duration	0.010***	(0.003)	0.010**	(0.005)
Trip Cost	-0.058**	(0.024)	-0.103**	(0.048)
Bus × Education	-0.166	(0.129)	-0.137	(0.153)
Bus × Family income	-0.951***	(0.143)	-0.946***	(0.168)
Bus × Household size	-0.269*	(0.143)	-0.395**	(0.173)
Bus × Age	0.248*	(0.149)	0.498***	(0.175)
Bus × Gender	N.A.	N.A.	N.A.	N.A.
Taxi × Education	-0.819	(0.826)	-0.826	(0.826)
Taxi × Family income	1.167	(0.899)	1.145	(0.900)
Taxi × Household size	-0.116	(0.780)	-0.137	(0.780)
Taxi × Age	0.002	(0.787)	0.020	(0.787)
Taxi × Gender	.	.	.	.
Train × Education	-0.026	(0.293)	0.103	(0.347)
Train × Family income	0.158	(0.276)	0.155	(0.320)
Train × Household size	-0.352	(0.303)	-0.138	(0.342)
Train × Age	0.371	(0.291)	0.244	(0.338)
Train × Gender	N.A.	N.A.	N.A.	N.A.
Observations	11389		8270	
pseudo R-square	0.665		0.664	

Note: Driving is the comparison group. Age Dummy equals to 1 if age is greater than 30 and smaller than 50.Gender equals to 1 if commuter is Male. Household size dummy equals to 1 if the household has more than 2 members. Family income dummy equals to 1 if annual household income is greater than $80,000.

Individual heterogeneity generates significant variation in demand and thus has large effects of consumer surplus calculations (Hausman and Newey, 2016). To further investigate the heterogeneity preference of commuters, we interact trip duration and trip cost with demographic variables and re-estimate the conditional logit model. Results are shown in Table 5. Middle age commuters, when facing long distance travel, are less likely to drive. Household with more than 2 members are more likely than other family to drive. High income household with large family size are more likely to choose driving as travel mode.

Table 5. Mode Choice with More Interactions

	Model 3: All Data		**Model 4: Peak Hour Data**	
Variables	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
Bus	-2.206***	(0.121)	-2.200***	(0.152)
Taxi	-3.722***	(0.845)	-3.761***	(0.950)
Train	-3.290***	(0.281)	-3.493***	(0.341)
Trip Duration	0.000	(0.004)	-0.002	(0.006)
Trip Cost	-0.038	(0.027)	-0.046	(0.057)
Trip Duration× Age	-0.018**	(0.008)	-0.026**	(0.011)
Trip Duration× Gender	N.A.	N.A.	N.A.	N.A.
Trip Duration× Household size	0.036***	(0.007)	0.051***	(0.011)
Trip cost × Income[1] × Household size	-0.083	(0.070)	-0.172*	(0.097)
Bus × Education	-0.148	(0.129)	-0.096	(0.155)
Bus × Family income	-0.942***	(0.143)	-0.894***	(0.170)
Bus × Household size	-1.001***	(0.208)	-1.277***	(0.269)
Bus × Age	0.631***	(0.204)	0.952***	(0.253)
Bus × Gender	N.A.	N.A.	N.A.	N.A.
Taxi × Education	-0.811	(0.831)	-0.824	(0.834)
Taxi × Family income	1.283	(0.901)	1.429	(0.913)
Taxi × Household size	0.013	(0.830)	0.262	(0.869)
Taxi × Age	0.096	(0.783)	0.143	(0.779)
Taxi × Gender	N.A.	N.A.	N.A.	N.A.
Train × Education	-0.035	(0.293)	0.091	(0.347)
Train × Family income	0.136	(0.277)	0.161	(0.320)
Train × Household size	-0.753**	(0.313)	-0.708*	(0.367)
Train × Age	0.562*	(0.305)	0.560	(0.366)
Train × Gender	N.A.	N.A.	N.A.	N.A.
Observations	11389		8270	
pseudo R-square	0.669		0.668	

Note: Driving is the comparison group. Age Dummy equals to 1 if age is greater than 30 and smaller than 50.Gender equals to 1 if commuter is Male. Household size dummy equals to 1 if the household has more than 2 members. Family income dummy equals to 1 if annual household income is greater than $80,000.

3.3 Quantifying Consumer Surplus

To analyze the impacts of the availability of Uber on consumer surplus, we collect bother Uber data and google data. Assuming consumer's preference doesn't change over time, for each OD pair from NHTS data, we construct the choice set by combining Uber data and google data according to origin ID, destination ID, trip periods (morning peak hour, afternoon peak hour, and other time), weekday (or weekend). We have observations of Uber data for one week. Google data are two months. We thus average google data according to OD pair and travel time and date. It is noteworthy that commuters in different OD pair might have face different choice set, since some alternatives are unavailable there. This is taking into account in our calculation.

We calculate consumer's surplus by comparing surplus in environment with the existence of Uber and the scenario when Uber is unavailable. Our calculation is conditional on the assumption that consumer's preference doesn't change over time. Let $\Psi = \{1,2,3,4,5\}$ denotes the travel set that a commuter faces, where 1 represents driving, 2 represents taking bus, 3 represents Taxi, 4 represents Train, and 5 represents Uber. And we labeled -1 for other travel mode that we call non-travel choice. That the travel choice set and non-travel choice set are connected by a inclusive value as described in the nested logit model (Train, 2013). The strength of connected is described by the log-sum coefficientλ.

To calculate the consumer surplus, we need to calibrate two parameters, $\bar{\phi}_{-1}$ representing the constant expected utility of the nontravel option, λ representing the log-sum coefficient. Following Small et al.(2006), the calibration of parameters are shown in Table 6.

Table 6. Calibration of Parameters

Item	**Value**
Constant Utility ($\bar{\phi}_{-1}$)	-12.65
Log-sum Coefficient(λ)	-0.36

We take a few steps to get the results of consumer surplus. Frist of all, we calculate the log-sum for each particular OD pair with the availability of Uber.

$$I_n^1 = ln\left[\exp(\bar{\phi}_{-1}) + \sum\nolimits_j \exp(\lambda x_{jn}\hat{\beta}_n)\right], j = \{1,2,3,4,5\}$$

Where x_{jn} a vector of attributes of choice j and $\hat{\beta}_n$ is the estimated coefficient from the conditional logit.

Second, we calculate the log-sum for each particular OD pair when Uber is assumed to unavailable.

$$I_n^0 = ln\left[\exp(\bar{\phi}_{-1}) + \sum\nolimits_j \exp(\lambda x_{jn}\hat{\beta}_n)\right], j = \{1,2,3,4\}$$

Third, we calculate the average consumer surplus for the representative consumer.

$$\Delta CS_n = I_n^1 - I_n^0$$

We use the following approximation to convert the consumer surplus from utility to monetary units

$$\Delta CS_n' = \frac{1}{\alpha_c}\Delta CS_n$$

Where α_c is the marginal utility of income measured by taking the derivative of utility with respected to costs. Alone with Small et al.(2006), α_c is determined by using Roy's identity, i.e., $\alpha_c = -(-2.4042 + 1.3869 * HIF_n)$ where HIF_n equals to 1 if n is from a high income family.

Finally, we also calculate the consumer surplus for each segment of population and the total welfare for the entire san Francisco area.

$$\Delta TotalCS = \sum\nolimits_i \frac{N_i}{\alpha_c}\Delta CS_n'$$

Where i is the index of OD pair and N_i is the population of commuters in OD pair i.

The results of the impacts of Uber on consumer surplus are show on Table 7. The average consumer surplus that a commuter gains in each trip ranges from 0.76 to 2.85 dollars/person. When only considering the effects brought by Uber Pool service, consumer surplus is between 1.6 to 5.8 dollars/person. If only considering the impacts caused by UberX service, the values are between 0.9 to 3.6 dollars/person. Generally, Uber Pool Service increases consumer surplus 2 times more than UberX does. We also consider the consumer surplus in weekday and weekend travelers. It is interesting that weekend travelers gain more consumer surplus than weekday travelers. Actually, this is consistent with the theoretical prediction. As shown in Small and Rosen(1981), the choice probability can be considered as the uncompensated demand on an alternative. Consumer surplus is equivalent to the utility gains, in monetary units, from an alternative that a commuter chooses to maximize his/her utility. For a weekday travel, it takes more time and costs in each particular trip and thus gains less maximized utility. Therefore, the consumer surplus is lower. This also applies for peak hour commuters. The non-peak hour commuters gain 3-4 times higher consumer surplus than that of peak hour commuters. One of the most possible reasons is the serious traffic jam during peak hour in San Francisco area.

To investigate the heterogenous effects of segments of population, we calculate the effects of Uber on people with different income. As show column 1 and column 2 of in Table 7, household with income greater than $80,000 gain more than 3 times higher consumer surplus than that of household whose income is lower than $80,000. When considering heterogenous preference, as shown in column 3 and column 4, these effects are even larger.

Table 7. The Impact of the Availability of Uber on Consumer Surplus

	Model 1	**Model 2**	**Model 3**	**Model 4**
Consumer Surplus				
Average(dollars/person)	0.76	0.79	2.85	1.52
Uber Pool Service	1.60	1.66	5.80	3.15
UberX Service	0.90	0.95	3.66	2.00
Weekday	0.68	0.71	2.66	1.43
Weekend	0.99	1.03	3.34	1.78
Non-peak Hour	1.56	1.64	5.74	2.98
Peak Hour	0.46	0.48	1.77	0.98
Household Income(> $80,000)	1.06	1.10	4.33	2.42
Household Income(<= $80,000)	0.44	0.46	1.25	0.55

To calculate the overall consumer surplus in monetary units that Uber bring to san Francisco commuters, we need to know the number of commuters in each OD pair. Unfortunately, this data in not available in our dataset. We use the census data which shows that there are approximately 265,000 workers travel into the city and about 103,000 head out per day[3]. Under this context, the overall consumer surplus that Uber bring to consumer is around $279,680 per day or $100 million per year.

[3] https://www.census.gov/hhes/commuting/files/ACS/top20-commuter-adjusted-population.pdf

Chapter 4. Conclusion

Consumers are better off when different options are readily available. For a long time, great attentions have been paid to Uber's social and economic impacts, while its benefits to individual consumer are generally ignored. There have been controversies surrounding Uber and other ride-hailing platforms about their disruptions to the taxi industry. In this report we take a unique approach to collect big data and to develop a quasi-experiment to quantify for the first time the gain in consumer surplus brought out by Uber, which is around $0.76 to $2.85 per person per trip. The findings in this report will help policy-makers cut through controversies and make more informed decisions.

References

1. Alon, Eizenberg. “Upstream innovation and product variety in the united states home pc market”. Hebrew University Working Paper (2014)
2. Berry, Steven T. "Estimating discrete-choice models of product differentiation." The RAND Journal of Economics (1994): 242-262.
3. Berry, Steven, James Levinsohn, and Ariel Pakes. "Automobile prices in market equilibrium." Econometrica: Journal of the Econometric Society (1995): 841-890.
4. Buchholz, Nicholas. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. Technical report, University of Texas at Austin, 2015.
5. Börsch-Supan, Axel. "On the compatibility of nested logit models with utility maximization." Journal of Econometrics43.3 (1990): 373-388.
6. Camerer, Colin, et al. "Labor supply of New York City cabdrivers: One day at a time." The Quarterly Journal of Economics 112.2 (1997): 407-441.
7. Dagsvik, John K., and Anders Karlström. "Compensating variation and Hicksian choice probabilities in random utility models that are nonlinear in income." The Review of Economic Studies 72.1 (2005): 57-76.
8. De Jong, Gerard, et al. "The logsum as an evaluation measure: review of the literature and new results." Transportation Research Part A: Policy and Practice 41.9 (2007): 874-889.
9. Hall, Jonathan V., and Alan B. Krueger. "An analysis of the labor market for Uber’s driver-partners in the United States." ILR Review (2015): 0019793917717222.
10. Hall, Jonathan, Cory Kendrick, and Chris Nosko. "The effects of Uber’s surge pricing: A case study." The University of Chicago Booth School of Business (2015).
11. Hausman, Jerry A., and Whitney K. Newey. "Individual heterogeneity and average welfare." Econometrica 84.3 (2016): 1225-1248.
12. Kalmanje, Sukumar, and Kara Kockelman. "Credit-based congestion pricing: Travel, land value, and welfare impacts." Transportation Research Record: Journal of the Transportation Research Board 1864 (2004): 45-53.
13. Nevo, Aviv. "Mergers with differentiated products: The case of the ready-to-eat cereal industry." The RAND Journal of Economics (2000): 395-421.
14. Petrin, Amil. "Quantifying the benefits of new products: The case of the minivan." Journal of political Economy 110.4 (2002): 705-729.
15. Small, Kenneth A., and Harvey S. Rosen. "Applied welfare economics with discrete choice models." Econometrica: Journal of the Econometric Society (1981): 105-130.
16. Small, Kenneth A., Clifford Winston, and Jia Yan. "Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design [with Comments]." Brookings-Wharton Papers on Urban Affairs (2006): 53-96.
17. Small, Kenneth A., Clifford Winston, and Jia Yan. "Uncovering the distribution of motorists' preferences for travel time and reliability." Econometrica 73.4 (2005): 1367-1382.
18. Svoboda, Miroslav. "History and troubles of consumer surplus." Prague Economic Papers 2008.3 (2008): 230-242.
19. Train, Kenneth E. Discrete choice methods with simulation. Cambridge university press, 2009.

Appendix A: Data Summary

Table A1. Important Variables in Household File

Variable Name	Label
HOUSEID	HH eight-digit ID number
HHSIZE	Count of HH members
LIF_CYC	Life Cycle for the HH
NUMADLT	Count of adult HHMs at least 18 years old
WRKCOUNT	Number of workers in HH
HHZIP	Zipcode of the HH location
CBSACAT	CBSA category for the HH home address
CBSASIZE	CBSA population size for the HH home address
HHMETDIV	Metro Division FIPS code for HH address
HHCITYFP	City FIPS code for home address
HHCNTYFP	County FIPS code for home address
GSTRATUM	Stratum of HH location based on geocoded address
DRVRCNT	Number of drivers in HH
HHFAMINC	Derived total HH income
WRKCOUNT	Number of workers in HH
DRVRCNT	Number of drivers in HH
Rail	MSA heavy rail status for HH
URBAN	Home address in urbanized area

Table A2. Important Variables in Person File

Variable Name	Label
HOUSEID	HH eight-digit ID number
PERSONID	Person ID number
PERINDT2	Person interview date
DIFFDATE	Number of days between travel date and interview date
JOBCATEG	Job category
MCUSED	Times used motorcycle/moped on road in the past month
R_AGE	Respondent Age
R_SEX	Responder gender
EDUC	Highest grade completed
DISTTOSC	Distance home to school
DISTTOWK	One-way distance to workplace
TIMETOSC	Minutes to get to school
TIMETOWK	Minutes to go from home to work last week
SCHTRN1	Mode to school
SCHTRN2	Mode from school
SELF_EMP	Self-employed
WKCNTYA	Work county
WKCTFIPS	City FIPS for work address
WKFTPT	Work full or part-time
WKRMHM	Has option to work at home
WKSTFIPS	State FIPS code for work address
WORKCT	Work place Census Tract
WORKSTAT	Workplace address - state
WORKZIP	Zipcode of work location
WRKAMPM	Arrival work -AM/PM
WRKHR	Arrival work - hour
WRKMIN	Arrival work - minute

Table A3. Important variable in Vehicle File

Variable Name	Label
HOUSEID	HH eight-digit ID number
VEHID	HH vehicle number used for trip
HYBRID	Hybrid vehicle
L_MAKE	Vehicle make name
L_MODEL	Vehicle model name
L_VYEAR	Vehicle model year
MAKECODE	Vehicle make code
MAKENAME	Vehicle make name
MODLCODE	Vehicle model code
MODLNAME	Vehicle model name
OD_DAY	Day of odometer reading
OD_MONTH	Month of odometer reading
OD_YEAR	Year of odometer reading
OD_READ	Odometer reading
VEHOWNMO	How long vehicle owned - Months
VEHTYPE	Vehicle type
VEHYEAR	Vehicle year

Table A4. Important variables in Travel Day Trip File

Variable Name	Label
HOUSEID	HH eight-digit ID number
PERSONID	Person ID number
TDCASEID	Trip number
PERINDT2	Person interview date
PUBTYPE	Mode of public transit used
TRAVDAY	Travel day - day of week
STRTTIME	Trip START time in military
TREGRTM	How long to destination from transit - converted to minutes
TRIPTIME	Entire trip took you
TRPACCMP	Number of people with you on trip
TRPMILES	Trip distance in miles
WHERE	Travel day trip destination
WHEREOS	Travel date trip destination - Other
WHYFROM	Trip purpose for previous trip

Table A5. Important variables in Geographical Information

TDCASEID	Trip number
HOUSEID	HH eight-digit ID number
PERSONID	Person ID number
HOMELAT	Household latitude
HOMELONG	Household longitude
TRPENDLA	Trip end latitude
TRPENDLO	Trip end longitude
WHERE	Travel day trip destination
WORKLAT	Work latitude
WORKLONG	Work longitude

Table A6. The Distribution of Trip Mile

STATS	TRPMILES
MIN	0.11
P1	0.44
P5	1
P10	2
P25	4
P50	8
P75	17
P90	30
P95	41
P99	62
MAX	85
MEAN	12.81
SD	13.36

Table A7. The Number of Trips in Each Transportation Mode

01 = Car	14,671
02 = Van	2,669
03 = SUV	4,561
04 = Pickup truck	2,108
05 = Other truck	62
06 = RV	15
07 = Motorcycle	81
08 = Light electric veh (golf cart)	19
09 = Local public transit	359
10 = Commuter bus	43
12 = Charter/tour bus	10
14 = Shuttle bus	29
15 = Amtrak/inter city train	7
16 = Commuter train	54
17 = Subway/elevated train	89
18 = Street car/trolley	19
19 = Taxicab	15

Table A8. Uber Sample Data

Origin GridID	Des GridID	weekday	Period	service	Travel Duration	Distance	Fare
4092	4201	1	other	UberX	5.43	1.11	6.6
4092	4201	2	PM	UberX	4	0.97	6.55
4092	4201	2	other	UberX	6	1.38	7.01
4092	4201	3	other	UberX	4.29	1.06	6.55
4092	4201	4	other	UberX	7	1.39	7.06
4092	4201	5	PM	UberX	5.56	1.28	6.74
4092	4201	5	other	UberX	4.52	1.11	6.59
4092	4201	6	PM	UberX	5.22	1.29	6.98
4092	4201	6	other	UberX	5.29	1.26	8.25
4092	4201	7	PM	UberX	8.75	1.8	7.73
4092	4201	7	other	UberX	8.82	2.4	8.64

Table A9. Google Map Sample Data

From Origin To Destination	**From (37.787, -122.397) To (37.758, -122.424)**			
Mode	Distance (Mile)	Estimated Total Traveling Time(Min)	Estimated Waiting Time(Min)	Fee ($)
AUTO	3.4	16	N.A.	N.A.
Bus	3.4	39	11	2.25
Train	3.4	37	12	1.95
Rail	3.4	37	12	1.95
Taxi	3.4	16	N.A.	N.A.
From Origin To Destination	**From (37.759, -122.464) To (37.783, -122.419)**			
Mode	Distance (Mile)	Estimated Total Traveling Time(Min)	Estimated Waiting Time(Min)	Fee ($)
AUTO	3.8	14	N.A.	N.A.
Bus	3.8	58	20	2.25
Train	3.8	51	20	2.25
Rail	3.8	60	20	4.20
Taxi	3.8	14	N.A.	N.A.

Note: This table gives two examples of how we generate transportation mode choice information using Google map. The only input is latitude and longitude of origin and destination.