C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION
A USDOT University Transportation Center

New York University
Rutgers University
University of Washington
The University of Texas at El Paso
City College of New York

Impact of Ridesharing in New York City

February 2020

Impact of Ridesharing in New York City

Stanislav Sobolevsky
New York University
ORC-ID: 0000-0001-6281-0656

Co PI: Kaan Ozbay
New York University
ORC-ID: 0000-0001-7909-6532

Devashish Khulbe
New York University
ORC-ID: 0000-0002-6436-8969

Chaogui Kang
New York University
ORC-ID: 0000-0002-0122-9419

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Acknowledgements

We appreciate the support from the US Department of Transportation, the NYU Tandon School of Engineering faculty startup funds, and the NYU Undergraduate Summer Research Internship program. Undergraduate students including Aiqi Zhou, Martin Buceta, Ziyan An, Eric Gan, and Weiyao Xie also contributed to this project. We also appreciate the discussion with Prof. Kaan Ozbay and the assistance from Shri Iyer, Joseph C. Williams, and John Petinos.

Executive Summary

Understanding holistic impact of planned transportation solutions and interventions on urban systems is challenged by their complexity but critical for decision making. The cornerstone for such impact assessments is estimating the transportation mode-shift resulting from the intervention. The project developed a citywide data-driven simulation modeling framework for probabilistic assessment of the mode-shift and resulting environmental, social, and economic impacts of transportation interventions. The framework simulates individual commuter choices under baseline and intervention scenarios accounting for general travel time and cost of each mode as well as individual preferences of the commuter and probabilistically inferring the unknown choice parameters given available data observations. The framework can deal with incomplete ground-truth travel demand data (e.g. historic commute data on major transportation modes available from transportation surveys and up-to-date taxi+FHV trip data from TLC with respect to inconsistencies between the two). It incorporates uncertainties from both - data inaccuracy/incompleteness as well as the model fit/parameter estimates – enabling assessment of statistical significance for the resulting impacts.

Considered intervention cases include introducing ride-sharing solutions (e.g. UberPOOL, Lyft shared etc.) in NYC and Manhattan Congestion Surcharge. As a result ride-sharing solutions cause overall commute time savings ranging between 1-2% for different categories of riders (more for the wealthier), however this comes with a slight mileage increase of up to 1.5% (more for the low-income commuters; as increased affordability of the service causes increased usage, largely offsetting miles cut by ride-sharing for most of the categories, except of the highest income commuters). This way shared FHV is the most efficient for the high-income commuters in terms of the time benefit vs traffic footprint tradeoff. Manhattan congestion surcharge on the other hand leads to an overall mileage in traffic decrease of up to 1% for different categories of commuters, which however comes at a cost of increasing their average travel time. The travel patterns of high-income commuters are impacted the most, as they, expectedly, use to rely on FHV more often. Travel mileage impacts for both scenarios are further translated into environmental impacts such as traffic congestions, gas consumption and emissions as well as the revenue impact for the public transit. The assessed travel time and cost savings constitute economic impact for urban population.

Being successfully evaluated on the cases above, the framework can be used for assessing mode-shift and resulting economic, social, and environmental implications for any future urban transportation solutions and policies being considered by decision-makers or transportation companies. It can work with diverse partially available historic or real-time data to provide statistically significant projections even in absence of the consistent ground-truth up-to-date observations for all the mobility modes involved.

Table of Contents

List of Figures

List of Tables

Section 1: Introduction

Text The vast scale of NYC can magnify even a slight improvement in the efficiency of the transportation solutions translating it into huge cumulative economic, environmental, and societal impacts. The rapidly growing for-hire vehicles (FHV) service is one area which can realize such optimization of drastically improving the efficiency of car and taxi transportation, as intended to cut traffic, congestion, and energy consumption [1]. Many evidences have demonstrated that ridesharing mitigated the impacts and it starts playing increasingly important role in NYC transportation growing at a fast pace (over 2.5 times from mid-2017 till end of 2018 with over 25 million miles traveled monthly by the end of 2018 on the shared rides according to NYC TLC open data) [2]. Such potential has been further unleashed with an in-depth understanding of the basic urban quantities/parameters (such as city size and driving speed) that affect the fraction of individual trips that can be shared [3]. Unfortunately, modal shifts resulting from increased affordability of the FHV service can easily offset those positive impacts, contributing to a substantial proportion of the overwhelmed road traffic and energy emission. Meanwhile, a discrimination of the impacts against different transport alternatives and for different population groups with distinct demographics is essential [4]. Urban stakeholders and municipal managers need to make informed decisions while considering policies and adopting solutions based on the travel behavior simulations driven by such knowledge, ideally, in a social petri dish.

The behavioral framework and a model for the set of complete and inter-related choices undertaken by travelers and potential travelers in the travel market is required. Both aggregate and disaggregate approaches have been developed to estimate travel demand and to split modal choices [5]. Those popular and widely used include the gravitational models [6], the Probit models [7], the Logit models [8] and many others. The explanatory variables included in the models often involve demographic, socioeconomic character, trip characters and mode attributes [8,9]. Many empirical evidences have demonstrated that, among the developed models, the Logit model often has more analytical advantages and offers more accurate results [10]. It has been taken as a reference model for urban travel mode choice simulations for a long time. Given a set of alternatives available, the probability of selecting a mode is determined by the Logit model (or its variants) as a function of the systematic portion of the utility of all the alternatives [11], and the proportion of modal shift under intervention scenarios can also be determined [12]. The most common model structures are the multinomial logit and nested logit models, which assume that the alternatives are grouped in nests (or combined modes) and the alternative (within each combined mode or between different combined modes) are independent from each other [13]. For the estimation and evaluation of a practical mode choice model, traveler and trip related data including the actual mode choice of the traveler are required, which should be obtained by surveying a sample of travelers from the population of interest.

A citywide synthesis at a mega-scale like New York City (NYC) requires enormous amount of detailed travel information for the modeling to understand the factors that affect travel-related choices and to predict how people travel in time and space. For decades, transportation researchers have largely used data of active solicitation [14], which are detailed but limited by a relatively small sample size (small data). The rapid rise and prevalence of mobile technologies have enabled the collection of a massive amount of passive data (big data) very different from data of active solicitation (small data) that are familiar to most transportation researchers and requires different methods and techniques for processing and modeling [15-17]. In recent years, data on human mobility and interactions in the city space saw an increasing number of applications. Data sources such as anonymized cell phone connections [18-21], credit card transactions [22,23], GPS readings [1], geo-tagged social media [24,25] as well as various sensor data [26] were leveraged as proxies for human mobility.

A critical drawback lies in having the available data either not including any user demographic information for individual trips or providing travel statistics with demographic information at the aggregate level only, as a response to alleviate privacy and surveillance concerns [27]. A synergy of disclosed (small and big) travel data from different data providers and departments is often required [28-30]: *to represent the resultant complete travel information (such as number of trips, travel time, and monetary cost) at a certain aggregate level, and it, subsequently, is not as accurate and detailed as the unsynergized incomplete data*. Such compromise imposes uncertainties onto both the data reliability and the modeling process [31-33], suggesting that the point estimates of modelled modal choices only represent one of the possible outputs generated by the models and, instead, anticipated modal choices are better expressed as a central estimate and an overall range of uncertainty margins articulated in terms of output values and likelihood of occurrence [34].

In this article, we seek to explore the modal choice behaviors in NYC using a data-driven framework based on partial ground-truth data and with consideration of both data and model uncertainties. By evaluating the synthesized transportation choices under scoping scenarios as well as the actual up-to-date taxi and FHV ridership, we train the mode-choice simulation model capable of simulating further mode-shift on the individual level under intervention scenarios of interest. Once quantified the mode-shift can be translated into the economic, environmental, societal impacts of the considered scenarios. The framework is illustrated on the two applied use cases - impact of ridesharing FHV in NYC as well as the Manhattan Congestion Surcharge, aiming to quantitatively inform stakeholders and policymakers of the implications of shared mobility and congestion pricing on the entire city as well as specific populations and neighborhoods.

Section 2: Data Description

Subsection 2.1 C2SMART simulation test bed data

The data provided by C2SMART simulation test bed based on [35] has approximately 27.3 million trips for travel modes - taxi, transit, walking and driving. In terms of wage categories, there are 16 sub-groups of wages for which number of commuters are present. Originally, the trips are aggregated on a Traffic Analysis Zones (TAZ) which are further aggregated into Taxi zones level for our models.

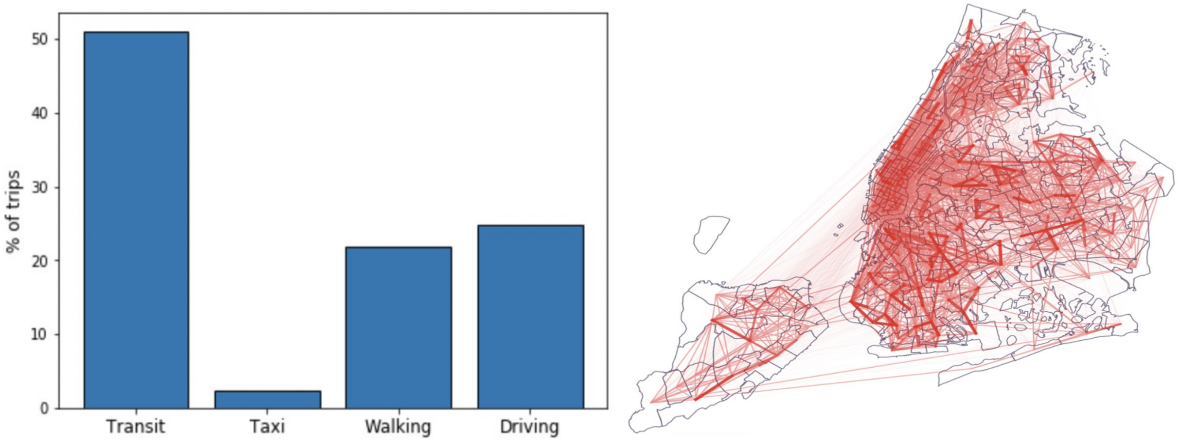

Figure 1: Distribution of trips by mode (left) and taxi zone coverage of data (right) in C2SMART data

In total, the data covers trips from 20,834 unique origin-destination pairs, covering almost 250 taxi zones out of 263 zones in New York City (Fig. 1). The data also has trips from modes like bike, carpool, shared bikes etc. which we do not include in our analyses. Nearly half of the trips constitute transit, followed by driving, walking and taxi (Fig. 1).

Subsection 2.2 LEHD/ACS data

Another data source is from LEHD (Longitudinal Employer-Household Dynamics). It has commuter information for 11 wage groups on a taxi zone level. Also present is the population choices of transportation for taxi, walking, transit, driving, biking, carpool etc. from the American Community Survey (ACS) data. Information regarding for-hire vehicles is missing, so only the four transport modes of our interest are present. Also, since only the origin-based commuter information is present, we can't explicitly have any true choices for a origin-destination pair. We use LEHD as alternative source of mobility demand distribution for impact stability evaluation purposes.

Subsection 2.3 Mobile GPS data

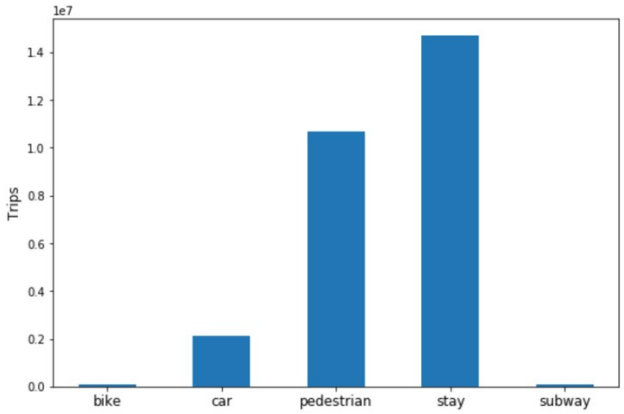

Figure 2: Trips distribution by mode in mobile GPS travel patterns

The mobility patterns provided by the project co-sponsor Arcadis are based on mobile GPS tracks and consist of 27.6 million unique trips in total. In terms of spatial distribution, the data has around 1.2 million unique point locations across the city which translates to 255 unique taxi zones out of the total 263 zones across NYC. Although the data consists of 5 unique travel modes- subway, bike, car, pedestrian and stay, we exclude the 'stay' mode from the analysis as it does not represent any mobility. The 'stay' mode classification also represents highest percentage of trips in the data. Excluding 'stay' mode, the data distribution represents 82% trips as pedestrian (walking), ~16% driving, and 1% transit and bike modes (Fig. 2). In the project we use the data as an alternative source of mobility demand distribution for impact stability evaluation purposes, while not relying on the mode choice inference.

Subsection 2.4 Travel times and fares

Travel times and cost information were obtained from publicly available APIs (HereMaps and Google API) along with information from TLC for taxi and for-hire vehicles (Appendix A.2). Necessary checks were made to remove outliers and data as well as comparisons were done across APIs to make sure information is coherent. To account for missing data for some origin-destination (O-D) pairs from for taxi and for-hire vehicles, we interpolated travel times and fares from known pairs. To incorporate uncertainty in the travel times and fares, we make sure to get information from at least 5 randomly selected O-D point pairs for each query of a taxi zone pair, both with API and TLC data.

Section 3: Modeling Frameworks

Our objective is to prototype a simulation modeling framework suitable for assessment of city-scale impacts of transportation innovations and policies on urban transportation systems along with the

associated environmental, economic, and social implications. The assessment will be evaluated on two pilot use cases of introducing ridesharing in New York City (offered through UberPOOL, Lyft Shared and other FHV companies) as well as Manhattan Congestion Surcharge. The impacts in question include: travel time and cost for passengers, traffic and congestion, gas consumption/vehicular emissions. Particular focus will be made on the equitably of the impact across populations.

Traditional counterfactual impact assessment is challenged by 1) the fact that spatial counterfactual does not seem feasible (ride-sharing is implemented city-wide and there is no comparable territory without deployment to be considered as control area), while 2) utility of the temporal counterfactual (comparing the same urban system before and after the deployment) is limited by multiple major trends and transformations happening within a complex urban system simultaneously with the deployment in question, 3) many target quantities of interest, such as overall urban traffic, gas consumption, emissions are hardly measurable with the available data and are again affected by multiple urban transformations happening simultaneously.

As an alternative, the present project proposes a methodology of constructing a data-driven integrated transportation simulation modelling framework, involving nested mode choice between private and public transportation, taxi and for-hire-vehicles, including ride-share modes as well as innovation adoption dynamics. For an estimated transportation demand, an agent-based choice model will be simulated, fitting unknown parameters of the individual utility of considered transportation modes as well as the agent characteristics (such as individual time value, particular mode preferences, innovation acceptance etc.) to the available partial observations of urban mobility.

Subsection 3.1 Multinomial Logit Model

We first consider the broadly used Multinomial Logit Model as the baseline approach for estimating the mode-choice for the regular commute. The model as well as its nested version (which we can use in case of related modes like taxi and FHV) offers an advantage of estimating the mode-choice probabilities using closed-form analytic formulas representing the aggregate-level choices of a simulation model. However, the parameters of the nested model lack the direct connection with the underlying simulation parameters and this way limit utility of the model for individual-level mode-shift assessment. However, it can still serve as a baseline to assess efficiency of the proposed simulation model, so we are including it to our study as such.

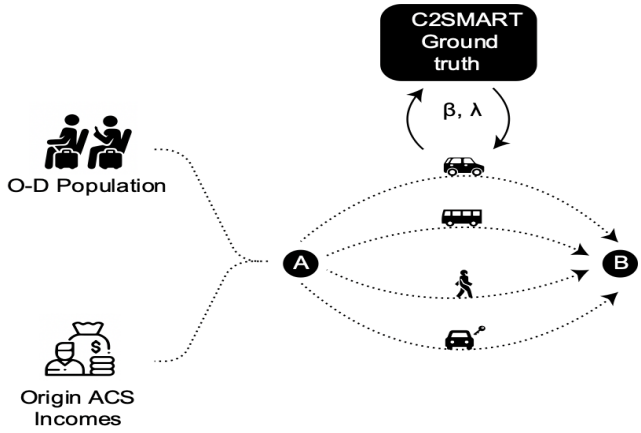

Figure 3: Multinomial Logit model framework

A Multinomial Logit (MNL) discrete choice model (Fig. 3) and its nested version with a nest for taxi+FHV and sub-nest for shared and non-shared FHV based are trained based on the two available datasets: (1) Number of trips between each O-D pair by wage group and 4 transport modes (Taxi, Transit, Walk, Driving) from C2SMART; and (2) Number of trips between each O-D pair by 3 transport modes (Taxi, FHV, shared FHV) from TLC. The models depend on a set of parameters - first of all λ, controlling the impact of the mode utility differences on the mode choice probability, β adjusting the objective value of time (time multiplied by individual wage rate) to anticipated monetary cost incorporating possible irrationality of individual decisions while combining it with the direct monetary cost in order to assess the overall utility. The nested model would further include $\tau_{\text{taxi+FHV}}$, τ_{FHV} controlling the choices between nests and within each nest [36]. Mathematically, the utility score Uj for alternative j depends on the time taken Tj between the O-D pair in consideration, the monetary cost Pj for choosing the alternative, the hourly income W of the commuter, and a random component of error ϵ_j, yielding a utility function

$$U_j = -\left(\beta W T_j + P_j + \varepsilon_j\right) \qquad (1)$$

and the individual utility of $U_j + \epsilon_j$. If ϵ_j follows a Gumbel distribution it can be seen that and the probabilities for each of the four major transportation modes to be chosen as having the highest utility is defined as

$$Pmode = \frac{e^{\lambda Umode}}{e^{\lambda Utaxi} + e^{\lambda Utransit} + e^{\lambda Uwalk} + e^{\lambda Udriving}} \qquad (2)$$

We further consider another version of the MNL with log-utilities (logMNL), corresponding to having multiplicative random factor applied to original utilities. Specifically adjust (1) as

$$U_j = -\left(ln\beta WT_j + P_j\right)$$

considering log-utilities and assuming individual log-utility to be $\mathrm{U_j} + \epsilon_\mathrm{j}$ with a random term again following Gumbel distribution. This will correspond to choosing a mode with a minimal inverse utility $\mathrm{e}^{-\mathrm{U_j}} = \left(\beta \mathrm{WT_j} + \mathrm{P_j}\right)\mathrm{e}^{-\epsilon_\mathrm{j}}$ rather than a minimal negative utility $-\mathrm{U_j} = \beta \mathrm{WT_j} + \mathrm{P_j} - \epsilon_\mathrm{j}$ in the classical setup, i.e. having a multiplicative exp-Gumbel individual random factor instead of an additive Gumbel random term.

When considering FHV and shared FHV modes one needs to acknowledge the relation with the taxi mode and corresponding correlations between individual preferences. This can be accounted by introducing a nest of taxi and modes along with a sub-nest of FHV and shared FHV modes to the model. For the nested model, the marginal probability of the outcome j is calculated based on the deterministic part Vj of the utility (i.e., $V_j = -\lambda(\beta WT_j + P_j)$), and the inclusive value IVk which signifies how inclusive each nest is based on its dissimilarity parameters (i.e., $IV_k = ln\ ln\ \sum_{l \in N_k} \quad e^{\frac{1}{\tau_k} V_l}$), yielding a chosen mode

$$Pr\ (y = j)\ = \frac{e^{\frac{1}{\tau_k} V_j}}{e^{IV_k}} \cdot \frac{e^{\tau_k V_j}}{\sum_m \quad e^{\tau_m IV_k}} \qquad (3)$$

The parameter τk cancels itself out for the nests containing a single transport mode. Eventually, the dissimilarity parameters τ1, τ2 for the taxi, non-shared FHV, shared FHV (sub-)nests together with the utility parameters λ, β determine the shift between each alternative, while τ1, τ2 largely control the balance within the taxi+FHV nest and FHV subnest.

The baseline model parameters were estimated through estimating λ, β of the utility function based on C2SMART data by minimizing the Weighted Root Mean Squared Error (WRMSE) between the number of trips from model prediction and real data for Taxi, Public Transit, Walk and Driving. The tested models measure the goodness of fit between model prediction and ground-truth data based on several metrics and search a wide range of parameters for the optimal fitting in reasonable time. The final nested model splits people's regular mobility between origins and destinations across the city (from LEHD data) and predicts aggregated transportation mode choices for each origin in high consistence with ACS data. The model also provides wage distribution for each transport mode to be used while assessing preferred transport mode choice for the commuters from the given wage group. In the next evolution of the model it will enable further direct simulation of their future choices under changing conditions according to the scenarios of interest.

Subsection 3.2: Individual choice-based simulation model

This approach is based on agent-based simulations of individual choices. In fact, so does the multinomial logit model representing one scenario when the individual preferences are represented by an additive random term following Gumbel distribution. This enables a closed form representation of the resulting probabilities, however not relying on that allows further flexibility in choosing the modeling framework. Besides direct control of the original simulation parameters will enable direct individual-level assessment of the mode-shift consistent with individual preferences.

In general, this model simulates mode choices for each individual origin-destination pair and the specific passenger of the given income category. We will use the two-stage Bayesian inference framework based on the the data of individual simulated trips generated by C2SMART simulation test bed as well taxi+FHV data available from TLC in order to estimate the model (Fig. 4). Specifically, for each set of the model parameters we can simulate the choices for each O-D pair based on available transport modes. For each given pair of origin-destination, passenger wage w and transportation mode, we assess utility based on travel time and cost estimates as well as a random factor, representing individual preferences towards each mode. Then log-utility is defined as

$$\ln U = \ln(\beta * t * w + c) + \epsilon$$

where β is the rationality adjustment for cost of time estimate as before, t is the travel time estimate, c is the travel fare/cost estimate and w is the wage of the commuter, while $\epsilon \sim N(0, \sigma^2)$ is the random component representing individual preference to the given mode. We assume ϵ terms generally independent across transport modes except of taxi, FHV and shared FHV, which are of course related - if one has increased preference towards taxi, it is likely that FHV will be also preferred and even more so between FHV and shared FHV which one can see as even more closely related, as while offering a slightly different type of service they are facilitated by the same provider/app. This way the model parameters to be estimated are: β, σ, corTFS - correlation coefficient between random factors ϵ of taxi and FHV or SharedFHV (SFHV) modes and corFS: correlation coefficient between random factors of FHV and SFHV.

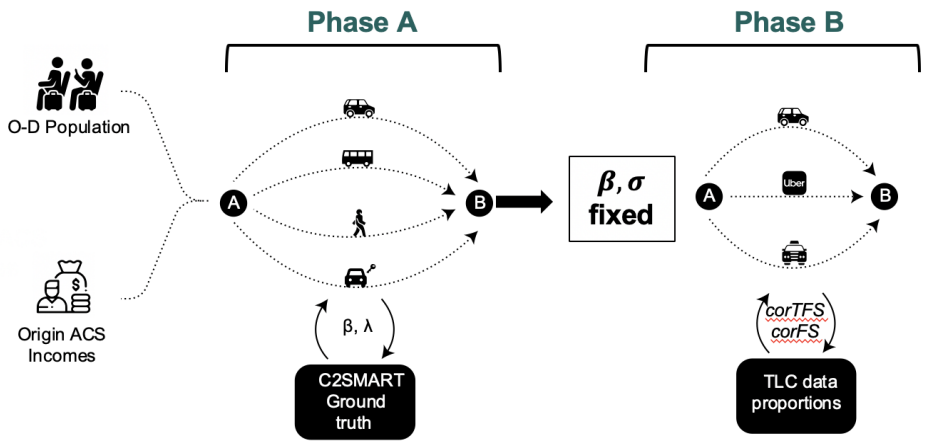

Figure 4: Individual choice-based simulation model framework

Likelihood estimation - probabilistic approach

Unlike the classical multinomial logit model setting, the proposed model might not have a closed-form analytic solution for the choice probabilities. While estimating those through simulations might be computationally costly. In order to streamline the computational process we implement a neural network (NN) model fitting the mode-choice probabilities $\mathrm{P_m}$ between the six transportation modes m as the function of their log-utilities and the model parameters. Notice that σ can be treated as the scaling factor for the log-utilities in order to simplify the model. In order to fit the model we simulate $\mathrm{P_m}$ for various values of of U_m, σ sampled from the random (normal) distribution and corTFS, corFS (provided corFS>corTFS) sampled uniformly and use it to learn a NN. The model architecture consists of three hidden layers with 8,12,8 neurons respectively, having the with 'relu' activation for hidden and sigmoid for the output layer trained on 'binary cross-entropy' objective function. We further use this pre-trained NN model for estimating $\mathrm{P_m}$ and further computing the likelihood of the observed C2SMART and TLC data give the model for each set of β, σ and correlations parameters instead of using simulations or explicit analytic formulas.

For mode choice probabilities $\mathrm{P_m(o,d,w,\sigma,\beta)}$ for each set of origin(o), destination(d) and wages(w), the log-likelihood for four modes given the observed C2SMART ridership $\mathrm{R_m(o,d,w)}$ is calculated as

$$L(\sigma,\beta) = \sum_{o,d,w} \sum_t R_t(o,d,w) \ln P_t\,(o,d,w,\sigma,\beta) \qquad (4)$$

For each σ,β with their chances weighted by $e^{L(\sigma,\beta)} prior(\sigma,\beta)$ and observed TLC ridership Rt(o,d,w) for t∈{taxi,FHV,SFHV} and estimated Pt(o,d,w,σ,β,corTFS,corFS) calculate

$$L_{FHV}(cor) = \sum_{o,d,w} \sum_t R_t(o,d,w) \ln \frac{P_t(o,d,w,\sigma,\beta)}{P_{TFHV}(o,d,w,\sigma,\beta)} \quad (5)$$

Where $P_{TFHV} = \sum_{t \in taxi,FHV,SFHV} P_t$. Essentially, for each (β,σ,corTFS,corFS), we weight the chances by $e^{L(\sigma,\beta)+L_{FHV}(corTFS,corFS)} prior(\sigma,\beta) prior(corTFS,corFS)$.

Based on the above framework, we obtained the best parameter sets of β= 0.71, σ= 0.38 and corTFS = 0.31, corFS = 0.58 based on likelihood values. The β,σ parameter values are sampled from log-normal prior distributions with $\ln \beta \sim N\left(\ln \mu_{beta}, \sigma^2_{beta}\right)$, such that P(0.33< β <1) = 68%, i.e. P(−ln 3<lnβ <0) = 68% which can be achieved when $\ln \mu_{beta} = -(\ln 3)/2, \sigma_{beta} = (\ln 3)/2$. Similarly, for σ, if we take $\sigma_{sigma} = |\ln(\ln 2)|$ and simulate multiple $\ln \sigma \sim N(\ln(\ln 2), (\ln(\ln 2))^2)$ then for the resulting ϵ the probability of P(0.5< ϵ <2) is again going to be 68% (for a log-normal prior distribution of $\ln \sigma \sim N\left(\ln \mu_{sigma}, \sigma^2_{sigma}\right)$).The correlation parameters corTFS and corFS are sampled from uniform distribution [0,1] such that corFS>corTFS. Then the sampling simply takes the 10%,30%,50%,70%,90% percentiles of each distribution with equal weights.

Once the parameters are sampled and the model fit likelihoods are assessed, it allows simulating the mode-choices for a variety of sampled parameters with the results weighted by the joint likelihood $e^{L(\sigma,\beta)+L_{FHV}(corTFS,corFS)}$ (as the prior sampling ensures even probability intervals). For express-assessment one can simulate the results just for the max-likelihood parameters, however comprehensive parameter sampling provides assessment with respect to the model uncertainty.

Based on the estimated parameter likelihoods, we simulate the final mode choices between origins and destinations for each individual commuter or group of commuters of a given wage group under two different scenarios of interest: (A) intervention scenario (having shared FHV unavailable or after imposing Manhattan Congestion Surcharge) and (B) the baseline scenario with all the transportation modes available with their original utilities. Individual correction factors ϵ are maintained the same between scenarios (A) and (B). For each individual simulation and the set of model parameters the mode-shift can be directly assessed and aggregated into percentage mode-shift over the entire city or origin, destination and/or wage group of interest. Being assessed for multiple sampled parameters it providing probability distributions with respect to parameter likelihood weighting.

The percentage mode-shift can be further translated into the impacts of interest with respect to the differences in travel time, cost and mileage driven between the transport modes.

Subsection 3.3 Model Comparison

We first evaluate the above simulation model against the classic MNL and logMNL (a version with multiplicative random factors for further consistency with the simulation framework above) according to

their capability of fitting the reported choices of four major modes (walking, driving, public transit and taxi) during the pre-FHV era.

All of the discussed approaches estimate mode-choice probabilities P_t for each origin-destination-wage pair based on the defined utility involving income of a commuter, travel time and costs. The Multinomial Logit (MNL) framework gives probabilities based on equation (2). Whereas for individual choice simulation model, we developed an approach to estimate choice probabilities and resulting likelihoods for each parameter sets through a NN model. The simulations corresponding to each parameter set are weighted by the likelihoods having their logarithms estimated by equations (4) and (5). Table 1 reports the likelihood-weighted averages for the mode-choices provided by each model.

	Taxi	Public Transit	Walking	Driving
Ground truth	634,535	10,619,997	9,416,078	6,663,358
MNL (multiplicative)	409,560	9,920,182	8,038,709	8,957,035
MNL (additive)	1,013,408	11,620,805	6,216,792	8,482,963
Choice simulation model	692,391	8,781,046	8,533,870	8,001,092

Table 1: Comparison of aggregated number of modal trips of MNL vs Choice simulation model for 4 modes

We observe that Individual choice simulation model provides estimates much closer to the ground-truth compared to either MNL or logMNL. Specifically, for the taxi ridership estimates (which is the most important for the considered use cases concerning taxi and FHV trips primarily), MNL underestimates the ground truth by over 1.5 times, while $\log\mathrm{MNL}$ overestimates by approximately 1.6 times. While the choice simulation model shows just a 9% deviation. It also gives much closer estimates for walking and driving, while under-performing on the public transit. Furthermore, choice simulation model provides a more adequate estimate for the travel time rationality parameter (the max-likelihood parameter of $\beta = 0.71$ corresponds to a quite realistic 29% undervaluing time, while optimal β for MNL and logMNL is above 1 corresponding to time overestimation, which contradicts common intuition of people generally valuating direct money benefits more than indirect benefits of the same estimated value.

Finally, as discussed, the simulation model framework provides better intuition and flexibility when simulating individual trips and evaluating alternative choices for the mode-shift part of the analysis. Based on this initial evaluation we are going to stick to the simulation model going forward.

Section 4: Uncertainty Analysis

Accounting for uncertainties is critically important for impact assessment in order to assess statistical significance of the reported city-wide quantities as well as their difference per wage group or areas across the city. We address uncertainties from two sources: uncertainty in the data and uncertainty in the model.

Uncertainty in the data is accounted for by incorporating the travel time and fares random distributions into the model and running the simulations multiple times. Then we calculate the mean and variance of trips for each of the four major modes from the simulations.

	Taxi	Public Transit	Walking	Driving
Ground truth	634,535	10,619,997	9,416,078	6,663,358
Simulation averages	692,391	8,781,046	8,533,870	8,001,092
Simulation std	2,241	4,449	5,480	5,808

Table 2: Data based uncertainty results for four modes

The variation in the trips from the data-based uncertainty simulations were observed to be too low to have any significant impact on the mode-shift quantities of interest (Table 2). Thus, we focus on the next aspect of uncertainty - model based uncertainty.

	Taxi	Public Transit	Walking	Driving
Ground truth	634,535	10,619,997	9,416,078	6,663,358
Simulation averages	692,391	8,781,046	8,533,870	8,001,092
Simulation weighted std	105,438	509,274	496,029	459,127

Table 3: Model based uncertainty results for four modes with probabilistic approach

Model based uncertainty is analyzed using the approach mentioned in the probabilistic approach of getting simulation likelihoods. For the log-normal prior distributions of beta, sigma along with uniform distributions of correlation parameters, we simulate utilities Ut(o,d,w) and weight by the likelihood values estimated by the formulas described in previous equations. We then estimate the weighted average and standard deviations of the simulations. We sample (10×10) values of beta, sigma and (10×10) corTFS, corFS pairs (corFS>corTFS) pairs drawn from the above distributions weighted equally. The 4-mode uncertainty results using this approach is given in Table 3.

Section 5: Impact Assessment

To evaluate applicability of the proposed framework to assessing impacts of transportation interventions and policies, the paper considers two use cases - introducing shared FHV after 2014 in NYC and imposing Manhattan Congestion Surcharge in early 2019.

Subsection 5.1 Impacts of ridesharing in NYC

As shared FHV became an integral part of NYC transportation, understanding their actual impact is challenged by the lack of an appropriate control area where shared FHV were not available. Historic pre-2014 mobility cannot serve as an adequate baseline as a rapidly evolving transportation system likely got affected by multiple trends, not only the spread of shared FHV. E.g. increased adoption of an FHV service as such (not necessarily shared) could have had a larger impact.

However, the proposed mode-choice model allows simulating a hypothetical scenario with the same transportation demand if shared FHV were not available. As described before we first train the model on the historic mobility represented by C2SMART simulation test bed and then further estimate FHV-related parameters based on the actual taxi, FHV and shared FHV ridership reported by TLC. Important to mention that the model is used to simulate the relative distribution of the ridership per mode for each origin-destination and passenger wage group, while in order to estimate the actual scale of the

impact we are going to rely on the actual amount of shared FHV reported by TLC (as those are the trips that would not have happened without ridesharing, while the alternative modes that would have been used are to be determined for those). This way dependence of the model on historic simulation test bed data is limited to estimating the likelihood of the parameters.

We analyzed the model-shift (if shared FHV trips were to be facilitated by the second-choice mode in each individual scenario) simulated by the model with different parameters weighted by the model fit likelihood in order to determine the anticipated effect of shared FHV on the NYC transportation system. The mode-shift (i.e. percentage of the observed shared FHV trips that would have been facilitated by public transportation, walking, taxi, FHV, and private vehicles) is reported on the Fig. 5. The model-based uncertainties seem relatively small, highlighting robustness of the pattern.

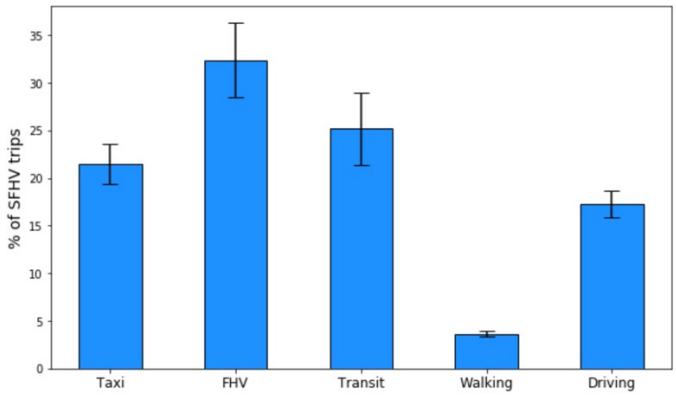

Figure 5: Percent of shared FHV trips accommodated by each alternative mode if shared FHV were not available

As one would expect majority of the shared FHV trips would have been facilitated by FHV and taxi as the closest alternative. Together with driving this adds up to nearly 70%. However around 30% of the trips have actually replaced transit and walking. So while majority of the shared FHV rides potentially (in case ridesharing actually occurred) cut the traffic by combining the trips that would otherwise involve individual driving, around 30% of those trips replace non-driving mobility, this way increasing the traffic.

On the aggregate citywide scale we observe a net travel time decrease of 1.77% (95% confidence interval - 1.71%-1.83%) and the net mileage increase of 1.14% (95% confidence interval - 1.06%-1.22%) even if we assume that each shared FHV trip have actually combined two trips (unfortunately we do not have ground-truth data on that, so this likely represents an optimistic scenario in terms of the traffic impact as some shared FHV might still serve individual passengers, while sharing more than two trips at once seems to be a rather rare case). In absolute numbers this corresponds to more than 62,000 hours saved for the NYC commuters at the price of 100,000 extra miles driven citywide over the year. So on average every hour saved comes at a price of a traffic increase by 1.6 miles. The mode shift impact also

result in a net decrease of around 260,000 trips for public transit which translates to close to $720,000 decrease in revenue for MTA. In terms of environment impact, the extra miles driven translate to close to 5,000 extra gallons of fuel emitting around 40 tons of carbon-dioxide emissions. In terms of economic impacts, the mode shift accounts for the citywide time-cost reduction of $1.1M.

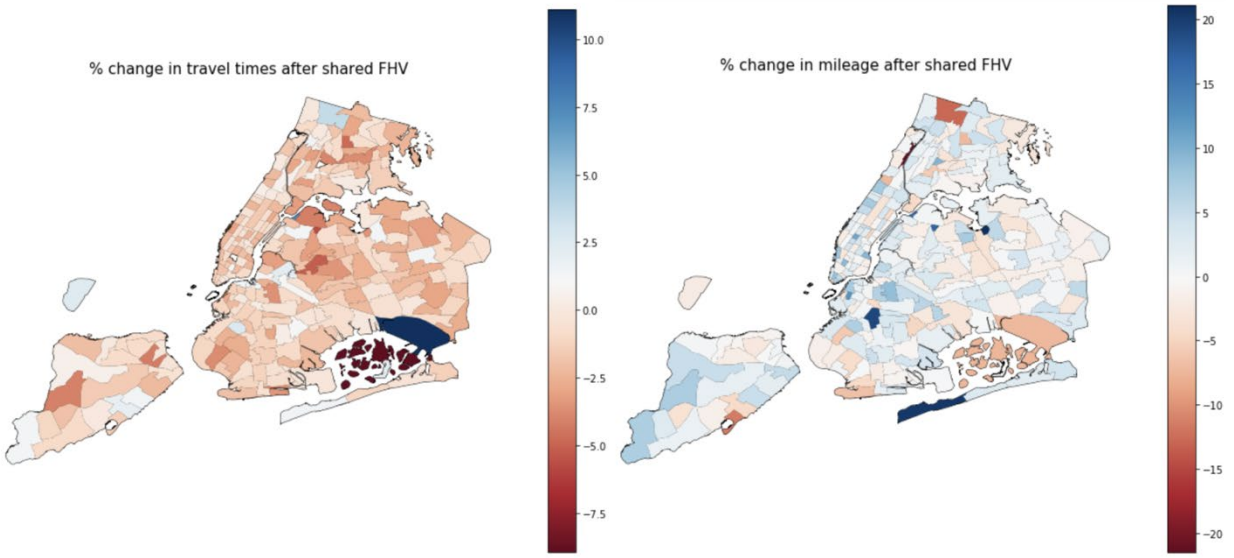

Figure 6: Percent change in travel times and mileage across taxi zones

While shared FHV cause an overall travel time decrease and traffic increase across the city, those impacts are greatly uneven across the city. On the level of individual taxi zones, the largest travel time decrease of up to 8% occurred in inner areas of Brooklyn, Queens and Staten Island which seem to benefit the most (Fig. 6) as the new relatively affordable commute option has likely bridged the local gaps in transportation accessibility. While some areas such as the airports actually saw an opposite effect of up to 8% increase in travel time, which can be related to using the shared FHV as a replacement for more expensive taxi and FHV service heavily used in such locations (having generally lengthy and expensive commutes for which people may compromise travel time for a significant cost savings).

Providing individual simulations with respect to the commuter wealth, the model allows to analyze the equitability of the impacts across urban populations. We observe the most significant changes for the low-income groups in % difference in mileage (Fig. 7), while the highest changes in travel times are observed for higher income groups (>100k annual income). For the high-income groups, the majority of shared FHV trips come from transit and driving. So there is an increase in mileage from transit to shared FHV trips and at the same time decrease from switch from driving to shared FHV. In case of low-income groups (<60k annual income), the mileage increase comes from shared FHV trips are being accommodated from walking and transit modes. In short it looks like the shared FHV service is the most

efficient for the wealthier in terms of the trade-off between improved travel time and the traffic footprint, while when used by low-income passengers it causes much heavier traffic footprint with smaller travel time improvement. Additionally, we observe that the mode-shift differences across income groups are significant with respect to the model-based uncertainties.

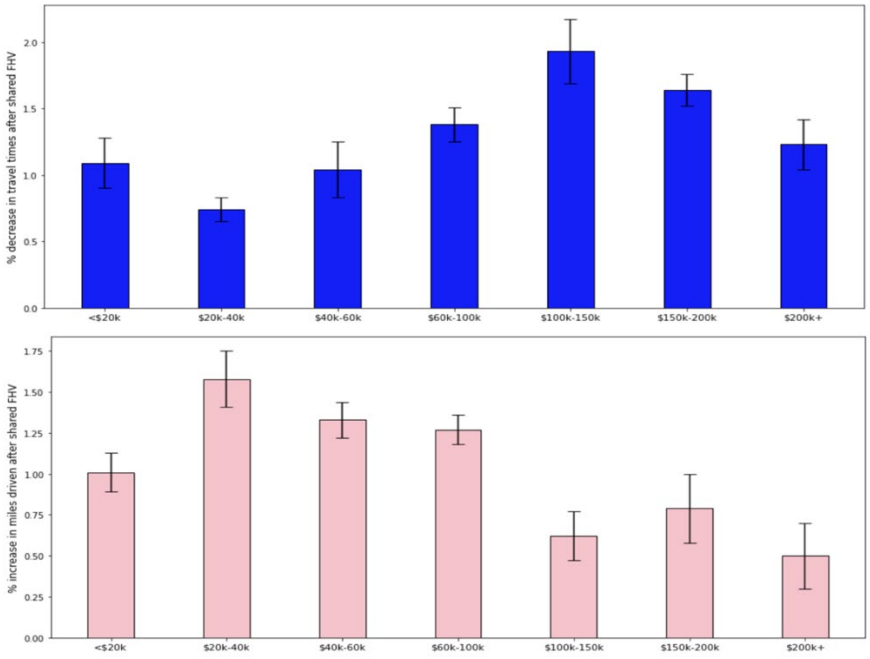

Figure 7: Percent change in travel times and mileage across commuter income groups

Subsection 5.2 Manhattan congestion pricing impact

To assess the impact of the Manhattan congestion pricing, we performed a set of simulations where we added fixed costs to taxis ($2.50), FHV ($2.75) and shared FHV ($0.75/passenger) for all trips originating in Manhattan. For shared FHV, we took an average of 2 passengers per ride at a time, so the total cost added was $1.50.

On a city-wide scale, we observe an increase in 1.09% in travel times and a 0.87% decrease in mileage, which can be attributed to lower usage of taxis and FHV and mode-shift to alternative modes.

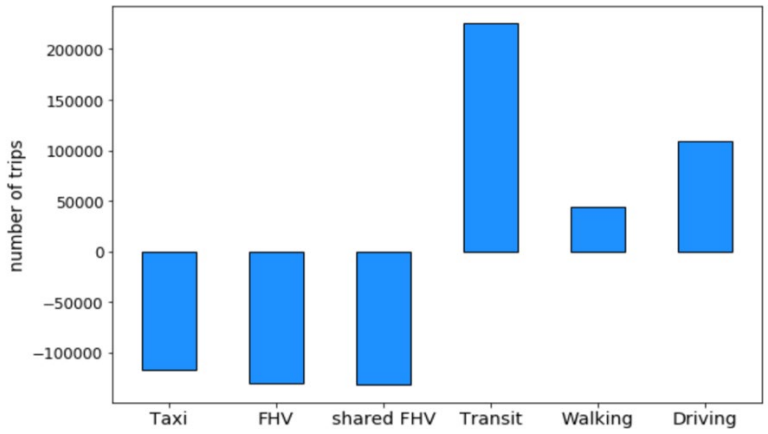

Figure 8: Number of added/reduced across travel modes after Manhattan congestion surcharge

On seeing the number of reduced trips across modes, we observe an almost equal drop across taxis and FHVs, although the highest reduction is seen for shared FHV. Almost 60% of the reduced trips are accommodated by transit mode, which translate into $610,000 projected increase in revenue for the MTA. Driving and walking accommodate 28% and 12% of the reduced trips respectively (Fig. 8). The decrease in number of trips for taxis and FHVs account for around 63,000 less miles driven which comes at a net increased travel time of 42,000 hours. This translate into the citywide economic impact of $570,000 time-cost value increase after the Manhattan congestion charge is added.

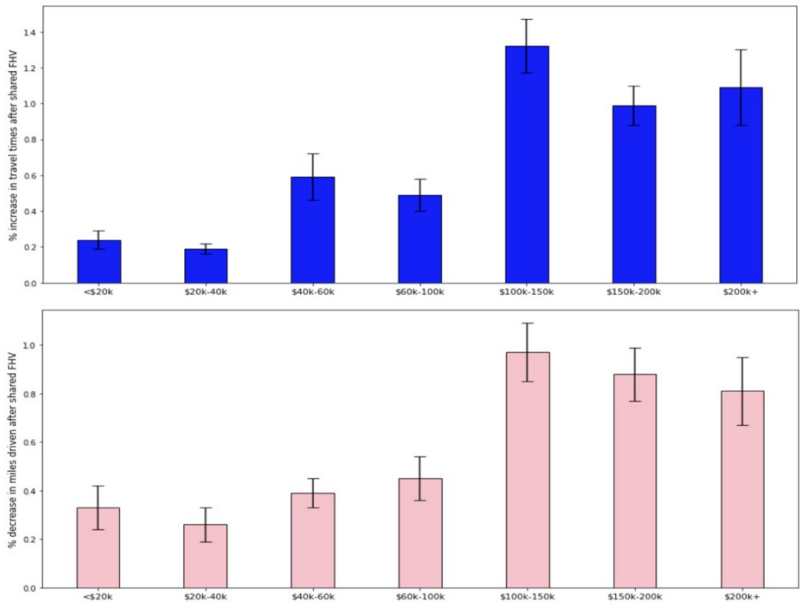

Figure 9: Percent change in travel times and mileage across commuter income groups after Manhattan congestion surcharge

Seeing from an equitability perspective, we observe the most dramatic changes for the high-income groups in % difference in travel times and mileage (Fig. 9), meaning that commute choices of the richest are affected the most. Compared to the low-income population, we see an increase of about 1 percentage points in travel times for high-income groups. The same is observed for total mileage driven, where the decrease is about 0.8 percentage lower for low income populations than high income groups. This makes sense as taxi and FHV ridership is seen across high income population. The highest mileage cut comes from top mode switch from FHVs and taxi to transit. This change is seen the most for the $100k-$150k income group whereas for >$150k income groups, the mileage cut decreases as top mode choice is private car instead of transit after the congestion surcharge. The mileage cut is significantly less for lower income groups as the number of trips of taxis and FHVs are low to begin with. With congestion surcharge, the top mode choice becomes transit/walking but the net number of trip changes are low compared to the higher income groups. The highest change among low income groups is observed for the $60k-$100k group where top mode choice switches to transit from FHV/shared FHV after the congestion charge is introduced. The same trend is seen for the travel times where the rich observe highest time increase owing to their switch from taxis/FHVs to transit mode. In terms of spatial impact, the biggest impact is seen in the high-income neighborhoods of Manhattan, specifically Lower East side, Upper East side and Upper West side parts of the borough. As compared to upper Manhattan neighborhoods like East Harlem and Washington Heights, the impacts in both travel times and mileage are relatively higher in Midtown and Lower Manhattan areas.

The relatively low changes in total travel times and mileage for the whole city can be explained with the low total proportion of taxi trips present in the data. Together, taxis and FHVs make up to around ~7% of the net mobility in the C2SMART simulations. Thus any monetary changes in fares in taxi+FHVs for one borough (Manhattan) translate into a low change in net times and mileages.

Consensus across C2SMART test bed, LEHD and Mobile GPS data

With the intervention scenario of introduction of shared FHV, we can expect an increase in citywide net travel mileage and decrease in travel times. But as the C2SMART simulation test bed data might not be a perfect representation of true mobility within the city, it is important to test our model across different mobility data sets to see how much the results might differ and if the model gives a reasonable estimate across different representations of mobility. Thus we decided to also test it on two other data mobility data sets for NYC – LEHD mobility and mobile GPS data provided by Arcadis.

The LEHD data has mobility information from across 47,000 O-D pairs compared to ~21,000 pairs from C2SMART. We ran the simulation model with the best likelihood parameters on the LEHD pairs. We observed a net citywide travel time decrease of 1.91% and a net mileage increase of 1.29%. The mobile

GPS data contains mobility information of around 27,000 unique O-D pairs. On running the individual simulation model on this data, we observed the top mode choice distribution as ~50% walking followed by ~30% transit, ~10% driving and rest taxis and FHVs under the normal scenario where all 6 modes are available choices. Under the intervention scenario of removing shared FHV as a choice, we notice the net citywide time increase of 1.2% and mileage decrease of 0.95%.

So the resulting impacts from the simulation model mildly depend on the data source of mobility demand (and compared to other quantifiable sources of the assessment uncertainty, the data source remains the most significant one), but generally remain consistent and close to the range of percent changes originally obtained from the C2SMART simulation test bed data.

Section 6: Summary

Subsection 6.1 Findings

The project constructed the simulation modeling and probabilistic inference framework suitable for assessment of city-scale impacts of transportation innovations and policies on the transportation system along with the associated environmental and economic implications with respect to uncertainty of such impacts. The framework applicability is illustrated on two use cases: introduction of shared FHV in NYC and Manhattan Congestion Surcharge. Between the two approaches compared - classical multinomial logit mode-choice modeling and individual simulation modeling - the latter showed better performance providing more realistic outputs and enabling individual-level mode-shift assessment consistently accounting for individual preferences. Also, the framework is capable of learning from diverse and possibly inconsistent datasets (such as historic transportation surveys and actual taxi and FHV ridership) providing partial information on urban mobility, stepwise gaining information from either source.

Broadly, our results indicate that shared mobility helped decreasing travel times for all categories of passengers. However, it does so by slightly increasing the traffic – decreases from trip sharing seem to be offset by growing number of riders due to increased affordability of the service. It works more efficiently for high-income categories of passengers providing higher travel time decrease with lower mileage increase. A Manhattan congestion surcharge, on the other hand, noticeably decreases the traffic, however it does so at the price of increasing travel time and in particular for high income travelers, who are perhaps the most frequent users of taxis and FHVs, to which the surcharge is targeted. The uncertainty analysis confirms statistical significance of the impacts as well as their heterogeneity across populations.

While we hope that this study can be a proof of concept for other cities considering shared mobility, congestion pricing or other similar interventions, it should be noted that New York City's transportation

system is unique in many ways, and makes a switch to public transportation more practical than in many other cities.

Subsection 6.2 Limitations

While the impact assessments in the paper provide proof-of-concept use cases for the proposed framework, further work may be needed to develop a comprehensive and accurate picture of the mode choices and mode shift. The outdated survey-based ground truth refined by C2SMART simulation test bed by itself might not be fully representative to the actual urban mobility. The current landscape of urban mobility might differ significantly from the RHTS and similar available transportation surveys conducted in the pre-FHV era. And although the historic data is only used as part of the parameter estimation for the model, while the scale of the impacts is based on the up-to-date TLC data, the mode-choice proportions and reliability of the impact assessment might still get affected. However, comparing the results for three different data sources of the commute demand distribution further confirm robustness of the findings.

Another limitation of the study is the simplicity of the utility function as presently considered. Accounting only for travel time and cost it may not reflect all the critical factors of how a person makes a transportation choice. The present utility function would work very well in an ideal world where everyone worked out the economics of their commute daily, but the reality is that transportation choices are influenced by habits, comfort preferences, and other human factors as well as environmental conditions. We focused our study on commuters because it allowed us to infer demographic and transportation demand information, but the morning commute takes up a small part of New York City's complex transportation system. The collection of more comprehensive ground truth data and accounting for more aspects of individual choices could further improve the reliability of the impact assessment. Finally, the validity of the mode-choice and impact assessments is conditional on the validity of the model, although an uncertainty assessment related to inaccuracies in the data as well as the model fit allows us to assess the degree of confidence in such an assessment, the specific model behind mode choices need to be assumed.

With that, the main contribution of the project is a proof-of-concept demonstration that a robust data-driven probabilistic modeling framework incorporating incomplete and inconsistent available mobility data, is capable of assessing the holistic picture of the urban commute and impact of transportation interventions with a reasonable degree of certainty.

References

[1] Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven H. Strogatz, and Carlo Ratti. Quantifying the benefits of vehicle pooling with shareability networks. Proceedings of the National Academy of Sciences, 111(37):13290–13294, 2014.

[2] Carol Atkinson-Palombo, Lorenzo Varone, and Norman W. Garrick. Understanding the surprising and oversized use of ride sourcing services in poor neighborhoods in New York City. Transportation Research Record: Journal of the Transportation Research Board, 2673(11):185–194, 2019.

[3] Remi Tachet, Oleguer Sagarra, Paolo Santi, Giovanni Resta, Michael Szell, Steven H. Strogatz, and Carlo Ratti. Scaling law of urban ride sharing. Scientific Reports, 7:42868, 2017.

[4] Michael Kodransky, and Gabriel Lewenstein. Connecting low-income people to opportunity with shared mobility. Institute for Transportation & Development Policy, 2014.

[5] Frank S. Koppleman, and Chandra Bhat. A self instructing course in mode choice modeling: Multinomial and nested Logit models.U.S. Department of Transportation Federal Transit Administration, 2006.

[6] Alex Anas. Discrete choice theory, information theory and the multinomial logit and gravitymodels. Transportation Research Part B: Methodological, 17(1):13–23, 1983.

[7] Farzad Alemi, Giovanni Circella, Patricia Mokhtarian, and Susan Handy. What drives the use of ride hailing in California? Ordered probit models of the usage frequency of Uber and Lyft. Transportation Research Part C: Emerging Technologies, 102:233–248, 2019.

[8] Chieh-Hua Wen, and Frank S Koppelman. The generalized nested logit model. Transportation Research Part B: Methodological, 35(7):627–641, 2001.

[9] Joachim Scheiner, and Christian Holz-Rau. Travel mode choice: affected by objective or subjective determinants? Transportation, 34:487–511, 2007.

[10] Ahmed Hamdy Ghareib. Evaluation of Logit and Probit models in mode-choice situation. Journal of Transportation Engineering, 122(4):282–290, 1996.

[11] Ke Wang, Xin Ye, Ram M. Pendyala, and Yajie Zou. On the development of a semi-nonparametric generalized multinomial logit model for travel-related choices. PloS ONE,12(10):e0186689, 2017.

[12] Pamela Murray-Tuite, Kris Wernstedt, and Weihao Yin. Behavioral shifts after a fatal rapid transit accident: A multinomial logit model. Transportation Research Part F: Traffic Psychology and Behaviour, 24:218–230, 2014.

[13] Shlomo Bekhor, and Yoram Shiftan. Specification and estimation of mode choice model capturing similarity between mixed auto and transit alternatives. Journal of Choice Modelling,3(2):29–49, 2010.

[14] Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies, 68:285–299, 2016.

[15] Marta C. Gonz´alez, C´esar A. Hidalgo, and Albert-L´aszl´o Barab´asi. Understanding individual human mobility patterns. Nature, 453(7196):779–782, 2008.

[16] Yu Liu, Xi Liu, Song Gao, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, and Li Shi. Socialsensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers, 105(3):512–530, 2015.

[17] Yang Yue, Tian Lan, Anthony GO Yeh, and Qing-Quan Li. Zooming into individuals to understand the collective: A review of trajectory-based travel behavior studies. Travel behavior and Society, 1(2):69–78, 2014.

[18] Girardin F, Calabrese F, Dal Fiore F, Ratti C, Blat J. Digital footprinting: Uncovering tourists with user-generated content. IEEE Pervasive computing. 2008;7(4).

[19] Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. arXiv preprint arXiv:08061256. 2008;.

[20] Amini A, Kung K, Kang C, Sobolevsky S, Ratti C. The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Science. 2014;3(1):6.

[21] Kung KS, Greco K, Sobolevsky S, Ratti C. Exploring universal patterns in human home-work commuting from mobile phone data. PloS one. 2014;9(6):e96180.

[22] Sobolevsky S, Sitko I, Des Combes RT, Hawelka B, Arias JM, Ratti C. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In: Big Data (BigDataCongress), 2014 IEEE International Congress on. IEEE; 2014. p. 136–143.

[23] Sobolevsky S, Sitko I, des Combes RT, Hawelka B, Arias JM, Ratti C. Cities through the prism of people's spending behavior. PloS one. 2016;11(2):e0146291.

[24] Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science.2014;41(3):260–271.

[25] Paldino S, Bojic I, Sobolevsky S, Ratti C, Gonź alez MC. Urban magnetism through the lens of geo-tagged photography. EPJ Data Science. 2015;4(1):5.

[26] Kontokosta CE, Johnson N. Urban phenology: Toward a real-time census of the city using Wi-Fi data. Computers, Environment and Urban Systems. 2016;64:144–153.

[27] Marie Douriez, Harish Doraiswamy, Juliana Freire, and Cĺ audio T. Silva. Anonymizing nyc taxi data: Does it matter? Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), 140–148, 2016.

[28] Mingxiao Li, Song Gao, Feng Lu, and Hengcai Zhang. Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data. Computers, Environment and Urban Systems, 77:101346, 2019.

[29] Mariano G. Beir´o, Andr´e Panisson, Michele Tizzoni, and Ciro Cattuto. Predicting human mobility through the assimilation of social media traces into mobility models. EPJ Data Science, 5(1):30, 2016.

[30] Zhiren Huang, Ximan Ling, Pu Wang, Fan Zhang, Yingping Mao, Tao Lin, and Fei-Yue Wang. Modeling real-time human mobility based on mobile phone and transportation data fusion. Transportation Research Part C: Emerging Technologies, 96:251–269, 2018.

[31] Goce Trajcevski. Uncertainty in spatial trajectories. Computing with Spatial Trajectories,63–107, 2011.

[32] Stefano Manzo, Otto Anker Nielsen, and Carlo Giacomo Prato. How uncertainty in input and parameters influences transport model: output A four-stage model case-study. TransportPolicy, 38:64–72, 2015.

[33] Soora Rasouli, and Harry Timmermans. Uncertainty in travel demand forecasting models: literature review and research agenda. Transportation Letters, 4(1):55–73, 2012.

[34] Andrew Boyce. Risk analysis for privately funded transport schemes. Proceedings of the European Transport Conference, 1999.

[35] He, Y., Zhou, J., Ma, Z., Chow, J. Y. J., Ozbay, K. (2020). ”Evaluation of city-scale built environment policies in New York City with an emerging-mobility-accessible synthetic population”. [Revision under review]

[36] Frank S. Koppleman, and Chandra Bhat. A self instructing course in mode choice modeling: Multinomial and nested Logit models. U.S. Department of Transportation Federal Transit Administration, 2006.

Appendix

Subsection A.1 (Travel Demand)

Data from two publicly available products - the Regional Household Travel Survey (RHTS) and the NYC Taxi and Limousine Commission trip records (TLC) - was used to determine the transportation demand between O-D pairs and the wage distribution of commuters. Initial exploration of the RHTS/TLC data supported our choice to focus on the commute hours (i.e., 7am - 10am and 5pm - 8pm).

Subsection A.1.1 (Regional Household Travel Survey / C2SMART)

The data was used to reflect reported choices of transportation modes by commuters serving as partial ground truth for fitting the model. Census tract level estimate data was pulled in order to generate probabilities of a commuter within each taxi zone of having wages within each Census income bracket. Commute information from the RHTS was reported by respondents which form of transportation they used “most days” for commuting to work, as to estimate the percentage of people in each taxi zone that regularly choose each form of transportation for trips to work. Collectively, the RHTS data allows us to estimate the probability of any given resident of each origin zone choosing each distinct mode of transportation, given the commuter’s income.

Subsection A.1.2 (Taxi Trips)

New York City’s Taxi and Limousine Commission provides access to their database of taxi trip data with ride-level granularity. These data give a sense of the high-volume traffic areas in the city, as well as the distribution of trips by time of day. They are crucial to estimate current modal distribution and, when used in conjunction with demographic RHTS data, to predict mode shift under various scenarios and within varying demographics. We found that the actual TLC data was most correlated with the LEHD Origin-Destination Employment Statistics (LODES) demand for the commute hours (see Appendix), and this correlation allows us to assume that the extracted trips are largely originating from the rider’s home taxi zone, enabling us to infer the commute trips by taxi from TLC data between taxi zone pairs (O-D pairs).

Subsection A.1.3 (For-Hire-Vehicles Trips)

The data was also provided by TLC consists of individual trip data for different FHV services (ex. Uber, Lyft, etc.). For analysis, we have separated the FHV trips to FHV and shared FHV and aggregated both data at the level of pick-up and drop-off zone, date and hours between commute hours. The aggregated data contains attributes of date, pick-up location id, drop-off location id, average trip duration (sec), trip counts, and surcharge flag (FHV and shared FHV). A typical month of data includes 15 to 20 million rides, with around 20\% of shared FHV, and 80\% non-shared FHV. We found the zones with high trip amounts are mostly concentrated at lower/middle Manhattan and downtown Brooklyn areas for both pick-up and drop-off locations. Such finding suggests that a large portion of our model simulations will be reflecting the trips in these areas, thus we need to be more carefully consider the regional demographic information of these areas as well as their functionalities (e.g. shopping, parks, etc.), to avoid any false assumptions when profiling people choices.

Subsection A.2 (Travel Budget)

The key elements of our model required us to estimate the time and cost associated with trips between each taxi zone pair for each of the six transportation modes (taxi, FHV, shared FHV, public transportation, walking, private vehicle). Additionally, to evaluate the utility of each transportation mode, we obtained the rider's wages from RHTS. We have aggregated all data sources to the TLC taxi zone level and the final version of the data which is used in the model includes pickup and drop-off locations, commute duration, price, and the wage distribution for that origin-destination pair. We considered mean fare amount, trip duration, and their standard deviations to inform data uncertainty.

Subsection A.2.1 (RHTS based C2SMART simulation test bed data and TLC Records)

To determine the time and cost of taxi trips between each O-D pair, we aggregated trip level (partial) data provided as open data from RHTS and TLC. If trip duration was not available, we supplemented it with duration estimates from the HERE/Google Maps API. The FHV trips presented an additional challenge as the TLC indicates trip duration and which trips are shared but does not include price information for those trips. To determine the cost per rider of FHV trips we aggregated data from the Uber API, pulled live during the commute hours on several workdays, taking the average for each O-D pair. The cost of private vehicle trips was calculated by estimating the cost of gas, vehicle wear and tear, and parking. We assumed a constant $2.75 price for public transportation trips and multiplied the total price according to the number of transfers. Walking trips were considered free of monetary cost.

Subsection A.2.2 (HERE/Google Maps)

HERE/Google technologies are the company that provides mapping and location data. In order to assess travel time, cost and overall utility of each transportation mode considered in the model given the O-D pair, we use HERE/Google REST APIs to gather information such as maps, routing, geocoding, places, positioning, traffic, transit, and weather information. The public transit data can be acquired via specific Public Transit API. Use HTTP GET methods, route information such as the trip time duration, the number of transfers, and the mode for each transfer will be get given departure and arrive location, departure time, and specific mode. To try and limit the impact of any special circumstances that would impact these estimates, we retrieved the data on several occasions and took an average. For those pairs with no route information available for either of the modes, we consider their corresponding time and distance values to be infinity. This situation can happen where there does not exist a way of commuting from one zone to another (e.g. islands).