ESTIMATING THE EFFECT OF UBER & LYFT ON PARKING VIOLATION IN NYC

Junjie Cai
Institution: Center for Urban Science + Progress, New York University
Address: 370 Jay Street, Brooklyn, NY 11201
Email: jc9033@nyu.edu

Junru Lu
Institution: Center for Urban Science + Progress, New York University
Address: 370 Jay Street, Brooklyn, NY 11201
Email: lj1230@nyu.edu

Yuxuan Wang
Institution: Center for Urban Science + Progress, New York University
Address: 370 Jay Street, Brooklyn, NY 11201
Email: yw1665@nyu.edu

Pranay Anchan
Institution: Center for Urban Science + Progress, New York University
Address: 370 Jay Street, Brooklyn, NY 11201
Email: pya209@nyu.edu

Shijia Gu
Institution: Center for Urban Science + Progress, New York University
Address: 370 Jay Street, Brooklyn, NY 11201
Email: sg5718@nyu.edu

Zhan Guo (Corresponding Author)
Position: Associate Professor of Urban Planning and Transportation Policy
Institution: Wagner Graduate School of Public Service, New York University
Address: The Puck Building, 295 Lafayette Street, Room 3010, New York, NY 10012
Email: zg11@nyu.edu

Word Count:
words + 3 table(s) $\times$ 250 = 750 words

Submission Date: August 2, 2019

ABSTRACT

This paper aims at exploring one potential ride-hailing services impact, taking the two biggest ridesharing companies Uber & Lyft as examples: whether daily Uber & Lyft trips affect New York City (NYC)'s parking demand, or specifically parking violations. NYC daily Uber & Lyft trips, parking tickets data and some additional data were collected and aggregated by taxi zones. Three technical models, Bayesian Network, Fixed Effects and Difference in Difference (DID) were applied on these data. The results of these models showed a negative correlation and causal effect between the number of Uber & Lyft trips and parking tickets, suggesting that Uber & Lyft may help reduce parking violations and likely parking demand in NYC. Given the controversial issues around ride-hailing services, this paper sheds light on the impact of Uber & Lyft and offer policy insight to the NYC Taxi and Limousine Commission (TLC) regulation.

Keywords: Ride-hailing services, Parking Violations, Causal Inference, Bayesian Network, Fixed Effect Model, Difference in Differences, New York City

INTRODUCTION

Ride-hailing services, like Uber & Lyft, are taking over marketing shares from Taxis (*1*). They together have earned a revenue of $13.46 billion in 2018 (*2*), while the revenue in U.S ride-hailing segment is roughly $44.8 billion (*3*). Nevertheless, the impact of Uber & Lyft is not limited to the For-Hire vehicles (FHV) market alone. As a critical part of the transportation system, they are reshaping our urban settings; although these effects are being questioned. On the one hand, Uber & Lyft have created thousands of new jobs with an hourly salary between $8.55 and $11.77 on average (*4*). On the other hand, they are suspected of increasing major city congestion significantly (*5*), and even damage the public transit system of United States (*6*).

However, due to the limited availability of Uber & Lyft trip data (Uber & Lyft only have public trip data available in NYC) and related scientific research, most of the Uber & Lyft effects have not been uncovered and proven. To impose effective regulations on ride-hailing services, it is critical for city administrators to understand them comprehensively. So, the goal of this paper is to investigate one potential ride-hailing services impact – parking demand.

Theoretically, ride-hailing services may affect parking demand through several channels. They may attract customers from private vehicles, traditional taxi cabs, transit systems, or people who would forge their travel if these service were not available. The first scenario may reduce parking demand as passengers do not need to park their cars. The second one may have little effect on parking, while the 3rd and 4th scenarios may potentially result in more parking demand as they are extra new vehicle travel even the parking need from rider-hailing service are relatively minimum. The net outcome depends on the relative proportion of the four scenarios, so parking demand might increase, decrease, or remain the same.

Note that the actual parking demand is hard to observe so a proxy measure is adopted by this paper – parking violations. Parking violations are only applicable to street parking and for a small portion of the parked cars, but it provides a proxy to understand parking demand. This paper examines whether Uber & Lyft increase/reduce parking violations in New York City (NYC). It leverages the power of large-scale machine learning and traditional statistics to model and empirically prove the causal effect of Uber & Lyft rides on parking violations in NYC and shed light on the impact of ride-hailing services on parking demand.

LITERATURE REVIEW

In a 2013 paper, Henao used a self-collected dataset containing information of 311 surveys to ridesharing passengers. He asked a few questions in the survey, including the driving frequency and the purpose of taking a ridesharing trip. By analyzing the dataset, he found out there are lots of frequent drivers, who take TNC rides to avoid parking (*7*). This could be a potential factor in the decrease in parking violations in the city. However, due to the size and nature of his dataset, his analysis cannot statistically prove this effect.

Looking at patterns of parking violations in NYC, 'tickets for stopping or parking in illegal zones, blocking traffic, tend to be most in common in areas with a greater commercial concentration of traffic.' The violations usually occur as a 'result of a driver needing to stop temporarily in a busy area where parking is scarce; they try to park in a loading zone or double park for a few minutes to avoid having to search for a parking spot.' (*8*)

As mentioned above, the current research mainly focuses on the study of ride-hailing services or parking violations alone, but not to the extent of the relationship between them. Even if some scholars is trying to fill the gap, they are frequently limited by the small size of the data.

This paper proposes a strategy combining large-scale machine learning methods and traditional statistics with several open datasets to effectively infer the causal effect on the parking violations brought by ride-hailing services in NYC.

Basically, causal inference is the process of drawing conclusions about a causal connection based on the conditions of the occurrence of an effect. In Pearl's 2009 paper, he discusses the advances in causal inference and "the approaches to be undertaken when moving from traditional statistical analysis to causal analysis of multivariate data." He surveys the "development of mathematical tools for inferring answers to three types of causal queries: queries about causal effects, probabilities of counterfactuals and direct and indirect effects (mediation)." (*9*) He discusses how "combining the algebraic component of the structural language of causality along with the graphical component provides a powerful and comprehensive methodology for empirical research" (*9*). In this paper, specifically three models - Bayesian Network, Fixed Effects and Difference-in-Differences (DID) are adopted.

In a 1996 study, it was shown how the "use of causal independence in a Bayesian network could greatly simplify probability assessment as well as probabilistic inference." (*10*) Heckerman and Breese look at an important weakness in a real-world Bayesian network representation, i.e. an effect could have potentially many causes; probability specification and inference could be impractical or impossible. They find some "potential sources for gains in inference efficiency, although these gains are not very apparent for more general Bayesian networks." They used a few models on several artificial and real-world Bayesian networks to better understand general Bayesian networks. In each model, they used the noisy-MAX model (*11*) to encode all parent-child relationships. They "transform Bayesian networks into an annotated undirected tree where each clique (node) corresponds to a set of nodes in the original network". Their results indicate the "use of decomposition can decrease inference complexity substantially when nodes have many states and many parents."

Fixed effects models provide a way to estimate causal effects in analyses where units are measured repeatedly over time. It can eliminate the effects of confounding variables without measuring them or even knowing exactly what they are, if they remain constant over time. One significant drawback with this model is that 'a great deal of information can be lost by focusing only on variation within individuals, thereby ignoring the variation across individuals.' Since this model removes the effects of all time-invariant causes, the standard fixed effects model is unable to estimate the effects of time-invariant measured causes (*12*).

Difference-in-differences (DID) is another mainstream statistical technique widely applied in econometrics and social sciences for causal inference. DID attempts to use observational data to mimic the design of experimental research, studying the differential effect of a treatment. (*13*) Alberto Abadie from Harvard University is one of the representatives who applied the DID model to the field of sociology. He used this idea in 2010 to effectively estimate the negative impact of California's Proposition 99 on local tobacco sales (*14*).

DATA

This paper uses three datasets to quantify the causal effect of estimating ride-hailing services versus parking tickets. All of them are aggregated with taxi zones, which is an officially defined geographical unit that roughly based on NYC Department of City Planning's Neighborhood Tabulation Areas (NTAs) and are meant to approximate neighborhoods. See the general attributes of taxi zones at `https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc`.

Parking Ticket Data

The parking violations data were obtained from NYC Open Data, and this project used data from 2014 to 2018. NYC issues about 11 million parking tickets each year. Four columns, Issue Date, Violation Code, Street Name, and Borough, and 80 types of parking tickets, which related to private vehicle, are used. This parking tickets data is completely open and being updated daily, which can be referred at this website: `https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2014/jt7v-77mi`.

Since the dataset only provides borough and street name for each parking ticket, Google API was used to geocode the given borough and street name for each observation to an approximate location with longitude and latitude and then mapped on to according taxi zones. Eventually, the dataset was grouped by date and taxi zone to show the number of parking tickets issued in each taxi zone in New York for each day from 2014 to 2018.

Uber & Lyft Ride Data

The Uber & Lyft trip data was extracted from the NYC For-Hire Vehicle (FHV) dataset that is available on the Taxi & Limousine Commission (TLC) website: `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. Dispatching base number, which is provided for each observation, is different for different FHV companies. So it was used to identify which trips were conducted by Uber & Lyft.

The collected dataset for Uber & Lyft trips includes fuzzy location by taxi zone (there are 263 taxi zones in New York City, and the mean partition granularity is 3 square kilometer) and time of pick up and drop off for each trip. Spark was used to group the dataset by trip date and pick-up and drop-off taxi zones. Pick-up trips were finally used as the number of Uber & Lyft trips. Each row of the prepared dataset indicates the number of trips conducted by Uber & Lyft originated within one taxi zone for each day in the four-year window. Parking ticket data and Uber & Lyft trips data were merged according to the taxi zone and the date to form a final prepared dataset. **Table 1** shows the prepared data.

TABLE 1 A sample of the prepared data

Taxi zone id	Date	Pickup counts	Ticket counts
1	2015-01-01	1	10
3	2015-01-01	9	0
4	2015-01-01	411	43
6	2015-01-01	2	22

Other Data

A general idea of causal inference research is to control thrid-party factors that uncorrelated with the independent variable but correlated with the dependent one. Thus, five additional spatial datasets that may affect the parking tickets from other aspects (geographical, demographic and socio-economic) were included. They were used in the DID as clustering variables and in the Bayesian Network as potential causation for parking violation (**Table 2**):

- **ACS data** was extracted from the American Community Survey 2015 5-year estimates.

18 statistical attributes, including population density, poverty rate, etc., were used to add the socioeconomic information for all models.

- **NYC crime data** is accessible on NYC Open Data. Three variables: number of felonies, number of violations, and number of misdemeanors, were mapped to taxi zone level and used in the model.
- **SAT result data** is hosted on NYC Open Data. The average score of SAT reading, math, and writing sections was included to measure the education level of each zone.
- **Transportation accessibility data** is available on NYC Open Data. This dataset includes subway entrance and bus stop coordinates. The number of subway entrances and the number of bus stops represent the transportation accessibility.
- **Parking capacity data** was collected from NYC Open Data, including meter parking and parking lot. They indicate the parking capacity of taxi zones.

TABLE 2 Composition of 5 additional datasets

Dataset	Attributes
ACS	'DensityPop': Population Density
	'TotalPop': Total Population
	'IncomePerCap': Income per capita ($)
	'Poverty': under poverty level rate (%)
	'Professional': % employed in management, business, science, and arts
	'Service': % employed in service jobs
	'Office': % employed in sales and office jobs
	'Construction': % employed in natural resources, construction, and maintenance
	'Production': % employed in production, transportation, and material movement
	'Employed': employed rate (16+) (%)
	'Unemployment': Unemployment rate (%)
	'Drive': % commuting alone in a car, van, or truck
	'Carpool': % carpooling in a car, van, or truck
	'Transit': % commuting on public transportation
	'Walk': % walking to work
	'OtherTransp': % commuting via other means
	'WorkAtHome': % working at home
	'MeanCommute': commute time (minutes)
Crime	'FELONY': Number of felony crimes in the taxi zone
	'VIOLATION': Number of violation crimes in the taxi zone
	'MISDEMEANOR': Number of misdemeanor crimes in the taxi zone
SAT	'sat': Average score of SAT reading, math, and writing
Transportation	'subway': Number of subway entrances
	'bus': Number of bus stops
Parking	'meter': Number of merter parking
	'parkinglot': Area of parking lot

METHODOLOGY

Three models - Bayesian Network, Fixed effects, and Difference in Differences (DID) were developed. The combination of the three methods was expected to comprehensively explain the causal effect of Uber & Lyft on parking violations. All of our data preparation and coding work can be reached at: `https://github.com/uberlyftparkingviolation/NYU-CUSP-Capstone-2019`.

Bayesian Network

The key idea of Bayesian Network is that one can distinguish correlation from causation if independent causes can be observed (*9*). This project integrated 14 observational features of taxi zones: 'Pickup', 'Ticket', 'PopDensity', 'IncomePerCap', 'Poverty', 'Professional', 'Employed', 'Crime', 'Subway', 'Bus', 'Meter', 'ParkingLot', 'Sat', and 'Carpool'. Bayesian Networks can be learned from the datasets to demonstrate whether Uber & Lyft directly/indirectly cause parking violations in NYC. There are three main steps: structure learning, parameter learning, and link strengths measuring.

Structure learning estimates a directed acyclic graph (DAG) that captures the dependencies between the variables given a set of data records. This paper combines two algorithms to learn the DAG: score-based structure learning and constraint-based structure learning.

Score-based Structure Learning approach construes model selection as an optimization task. It has two building blocks: first apply a 'scoring function' $s_D : M \rightarrow \mathbb{R}$ that maps models to a numerical score, based on how well they fit to a given data set D; second perform 'search strategy' to traverse the search space of possible models M and select a model with optimal score. For the first building block scoring function 'Bayesian Dirichlet', scores such as BDeu, K2, and BIC (Bayesian Information Criterion) were used to measure the fit between model and data. For the second building block 'Search strategy', a heuristic search 'HillClimbSearch' was used, which implements a greedy local search that starts from the DAGÂăstart (disconnected DAG) and proceeds by iteratively performing single-edge manipulations that maximally increase the score.

Constraint-based Structure Learning attempts to correctly capture the directionality of causal relationships. It has two building blocks: identifying independencies in the data set using hypothesis tests and constructing a DAG according to identified independencies. For the first step, independencies in the data can be identified using χ^2 conditional independence tests. TheÂăp-value is the probability of observing the computed χ^2 statistic or a higher χ^2 value, given the null hypothesis that X and Y are independent for given Zs. This can be used to make independence judgments, at a given level of significance: p-values less than the chosen significance level α reject the null hypothesis, and it can be concluded that X and Y are conditionally dependent given Zs. P-values greater than or equal to α fail to reject the null hypothesis, from which it can be concluded that X and Y are conditionally independent.

Parameter learning is a reinforcing stage of Bayesian Network method. The task of parameter learning is to estimate the values of the conditional probability distributions (CPDs). Given a Bayesian network structure and a training dataset, one can learn the parameters (conditional distribution of each node for each distinct combination of parent values) by maximum likelihood. There are two main algorithms: Maximum Likelihood Estimation (MLE) and Bayesian Parameter Estimation. MLE simply uses the relative frequencies with which the variable states have occurred. The Bayesian Parameter Estimator starts with already existing prior CPDs, that express our beliefs about the variables before the data was observed. Those "priors" are then updated, using the state counts from the observed data. Parameter learning would be performed in the future to further

improve the causal inference ability of our network.

Pearson correlation coefficients are calculated to measure the strength of connection along a specific edge. Link strengths among socioeconomic features are represented by spatial correlations since they are differential though zones but rarely vary through years. Link strength between Uber & Lyft and parking violations is represented by the combination of spatial correlation and temporal correlation. For the time series correlation, we are focusing on the periodic variance rather than the constant part. Thus Hodrick-Prescott (HP) Filter is implemented to split the overall trend and cyclical pattern. The overall trends represent the fixed parts of tickets and Uber & Lyft trips that must come up no matter what happens. And the cyclical patterns represent the daily variations. Finally the temporal correlation was calculated with the filtered cyclical variations.

Fixed Effects Model

Fixed effects model (FEM) is widely used to control for unobserved variables in causal inference (*15*). FEM fits this study well, which aims to explore the causality using multidimensional datasets shown in **Table 1**. The datasets were standardized before modeling in order to improve the optimization training process and measure variable importance. To develop a FEM, dummy variables of each taxi zone were added to a standard OLS model to evaluate the fixed effects of the number of Uber & Lyft trips across different taxi zones. Thus the effect of Uber & Lyft trips can be seperately measured after filtering the unobserved fixed effects. Considering the model for taxi zones $i = 1, \ldots, N$ which is observed at certain time periods $t = 1, \ldots, T$ (**Equation 1**):

$$y_{it} = \alpha + \beta \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} + \gamma \sum_{i=1}^{N} z_i + u_{it} \tag{1}$$

where y_{it} and x_{it} are, respectively, the number of parking tickets (dependent variable) and the number of Uber & Lyft trips (independent variables) in the i-th taxi zone at time t; z_i are unobserved, time-invariant dummy variables in different taxi zones; β and γ represent coefficients of the independent variables and dummy variables; α and u_{it} refers to the intercept and the error term. The coefficient, β, of the independent variables would indicate the causal effect exerted by Uber & Lyft trips on parking violations.

DID: Difference in Differences

Within the framework of natural science, two identical NYCs, one having TNC companies operating in it and the other not, were needed to strictly prove the causal effect of Uber & Lyft on parking violations. However, such ideal 'treatment groups' and 'control groups' were impossible to be found or developed in social science research. Consequently, DID model was applied in this project to solve this problem. **Figure 1** shows the general idea of DID and how it works. The 'treatment object' and 'control object', P and S, have status P_1 and S_1 on time period 1, and status P_2 and S_2 on time period 2. Based on the DID's most critical assumption, parallel trend assumption, that any existing factor would have same effect on P and S, shown by the parallel S_1 to S_2 line and P_1 to Q dotted line DID, the difference between Q and S_2 on the second time period equals the difference between P_1 and S_1 on the first time period, and the difference between P_2 and Q represents the effect of the outside factor or treatment.

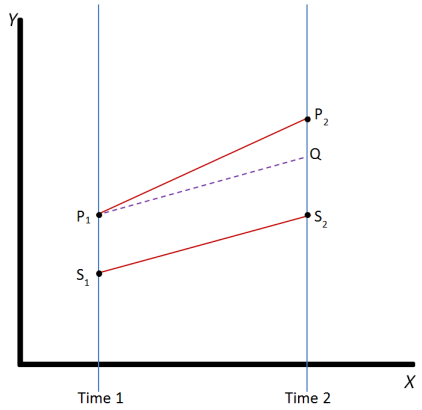


FIGURE 1 Illustration of DID model, by Danni Ruthvan

Average causal effect can be expressed by **Equation 2**:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta Outside}{(P_2-S_2)-(P_1-S_1)} \tag{2}$$

where $\Delta Outside$ refers to the difference between the average amount of Uber & Lyft trips in 2015-2018 between the 'treatment taxi zone' P, and the 'control taxi zone' S, for every i. P_1, P_2, S_1, and S_2 refer to the average amount of parking tickets in 2014 and 2015-2018 between P and S. n refers to the number of 'treatment vs. control' pairs.

To meet the parallel assumption, cross-clustering was applied to find multi-dimensionally homogeneous taxi zones to be used as 'treatment group' and 'control group.' The taxis zones were clustered based on two dimensions separately, extracting those zones always belonging to the same clusters: the number of parking tickets for each taxi zone in 2014, which, shown by data, had not been affected by Uber & Lyft; and the above 5 additional datasets, which summarizing different aspects of taxi zones from geographic, demographic and socio-economic views that will also hardly affected by Uber & Lyft over time. In each cluster, the remaining zones were pairwise paired, which generated over 900 pairs of homogenous taxi zones.

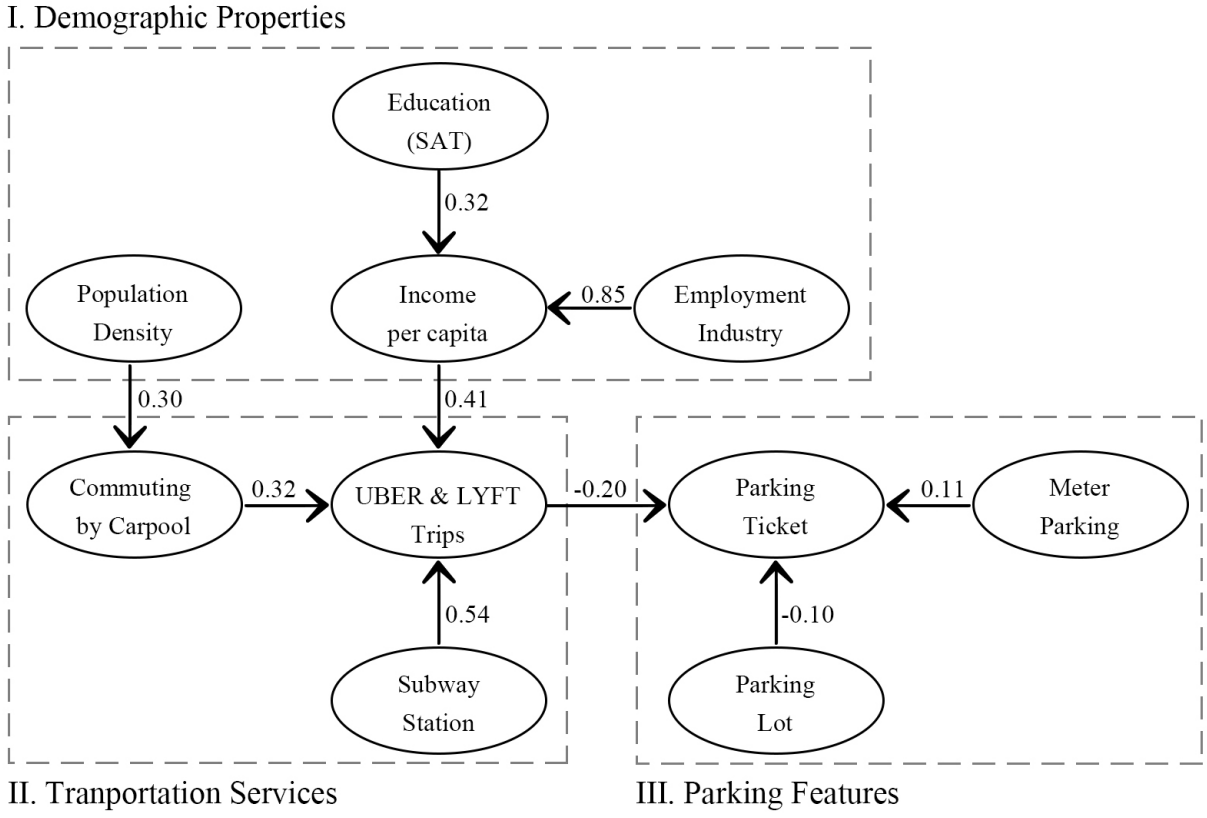
RESULTS

Results of these three models are listed below, which proves the negative casual effect on parking tickets brought by Uber & Lyft trips in NYC.

Bayesian Network

The causal network structure generated by Bayesian Network indicates that Uber & Lyft have a direct impact on parking tickets. According to **Figure 2**, there are three variables that directly influence the parking ticket, and Uber & Lyft is the strongest one of them. After applying HP filter (**Figure 3**), the Pearson's correlation coefficient is -0.20 between the cyclical variations of Uber & Lyft trips and parking tickets. And p-value is $6.52 * e^{-14}$, much lower than the significant level of 0.05. In conclusion, Uber & Lyft statistically significantly reduce the parking violations.

However, there are limitations and potential biases with Bayesian Network. Initially, Bayesian Network method bases on many hypotheses, including causal Markov, causal faithfulness, causal sufficiency, and acyclicity, but it is difficult to meet all of them in real world. Meanwhile, unobserved intermediate factors may be missing between nodes.That is, Bayesian Network could never be absolutely certain whether the influence is direct.



Notes:

Causal Network struture is learnt by Bayesian Network algorithm.
Association strengths among edges are represented by Pearson's Correlation Coefficients.

FIGURE 2 Bayesian Network

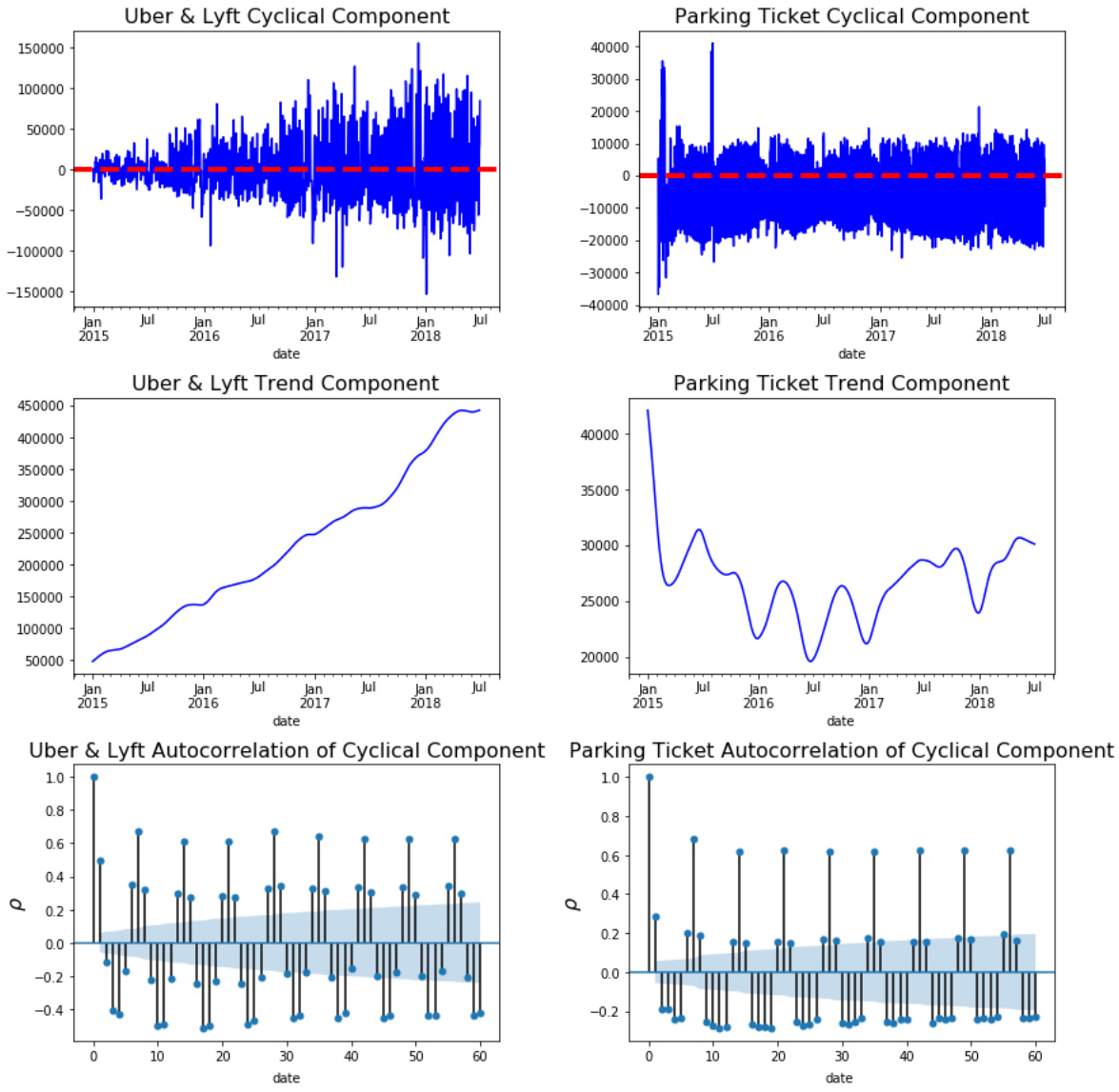


FIGURE 3 Fixed Trend and Cyclical Variations Filter

Fixed Effects Model

The fixed effects model (FEM) controls for the unobserved factors. According to **Figure 4**, R-squared of the Uber & Lyft regression is low since the number of parking tickets cannot be explained by Uber & Lyft trips solely. But after adding dummy variables, the model achieves an R-squared of 0.70, well explaining the variation of number of parking tickets. The coefficient of Uber & Lyft trips (pickup) is -0.0007, indicating Uber & Lyft trips slightly reduce NYC parking violations. However, this effect is not statistically significant, because the p-value of is 0.677, much larger than the significance level of 0.05. And the confidence interval across zero (-0.0039, 0.0026) shows the effect of Uber & Lyf is volatile through taxi zones.

Fixed Effect Model Result Summary

Dep. Variable:	tickets	R-squared:	5.197e-07
Estimator:	Pooled Least Squares	R-squared (Between):	-0.0002
No. Observations:	333768	R-squared (Within):	5.197e-07
Date:	Sun, Jul 29 2019	R-squared (Overall):	-8.903e-05
Time:	14:55:35	Log-likelihood	-2.71e+05
Cov. Estimator:	Unadjusted		
		F-statistic:	0.1733
Entities:	264	P-value	0.6772
Avg Obs:	1264.3	Distribution:	F(1,333503)
Min Obs:	8.0000		
Max Obs:	1277.0	F-statistic (robust):	0.1733
		P-value	0.6772
Time periods:	1277	Distribution:	F(1,333503)
Avg Obs:	261.37		
Min Obs:	255.00		
Max Obs:	263.00		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-1.037e-15	0.0009	-1.098e-12	1.0000	-0.0018	0.0018
Uber & Lyft	-0.0007	0.0017	-0.4163	0.6772	-0.0039	0.0026

F-test for Poolability: 2982.9
P-value: 0.0000
Distribution: F(263,333503)

OLS Regression Results

Dep. Variable:	tickets	R-squared:	0.703
Model:	OLS	Adj. R-squared:	0.703
Method:	Least Squares	F-statistic:	2989.
Date:	Sun, 21 Jul 2019	Prob (F-statistic):	0.00
Time:	15:36:44	Log-Likelihood:	-2.7104e+05
No. Observations:	333768	AIC:	5.426e+05
Df Residuals:	333503	BIC:	5.454e+05
Df Model:	264		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.7892	0.015	-51.613	0.000	-0.819	-0.759
C(zone)[T.1.0]	0.2108	0.022	9.765	0.000	0.168	0.253
...	...	...	...	...	...	...
C(zone)[T.263.0]	0.0008	0.022	0.038	0.970	-0.042	0.043
Uber & Lyft	-0.0007	0.002	-0.416	0.677	-0.004	0.003

Omnibus:	79835.840	Durbin-Watson:	1.183
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5088941.734
Skew:	-0.080	Prob(JB):	0.00
Kurtosis:	22.129	Cond. No.	265.

FIGURE 4 Fixed Effects Model Result

DID: Difference in Differences

A pair of homogeneous taxi zones in **Figure 5** and **Figure 6** show how the parallel trend assumption was met. The map in **Figure 5** indicates that Flatiron and Midtown Center are similar. Most of their attributes, displayed in the table, are highly close. Their parking violation distribution in 2014 is also similar, shown in blue in the first two line charts in **Figure 6**. However, the distribution of the number of Uber & Lyft trips and the parking tickets after 2014 of these two areas are different, shown by the orange and green lines on the line charts in Figure **Figure 6**. The Uber & Lyft trips in Midtown Center are almost twice the trips in Flatiron. Therefore, the difference between Uber & Lyft trips between the two homogeneous taxi zones is the only external factor that causes the significant change of the parking ticket.

	Flatiron	Midtown Center
Shape_Leng	0.0307592	0.0358039
Shape_Area	5.53132e-05	7.19131e-05
Income Per Capita (2015)	53729.1	53744.1
Under poverty level rate (%)	19.6629	18.8357
Employed in sales and office jobs (%)	22.4886	20.8286
...	...	...
Commuting alone in a car, van, or truck (%)	14.7486	15.2024
Working at home (%)	5.02857	5.48095
Average commute time (minutes)	36.0571	35.9366
Average SAT score (2019)	428.287	428.287
Men in Population (%)	50.1118	48.9601
Women in Population (%)	49.8882	51.0399
White in Population (%)	41.4686	39.3095

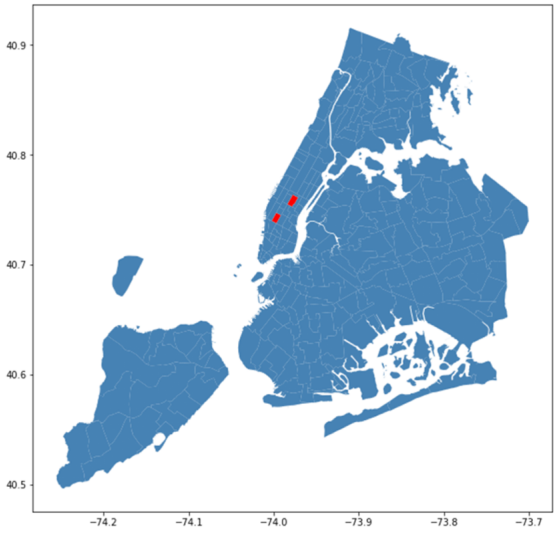


FIGURE 5 Features and Locations of a pair of highly homogeneous taxi zones

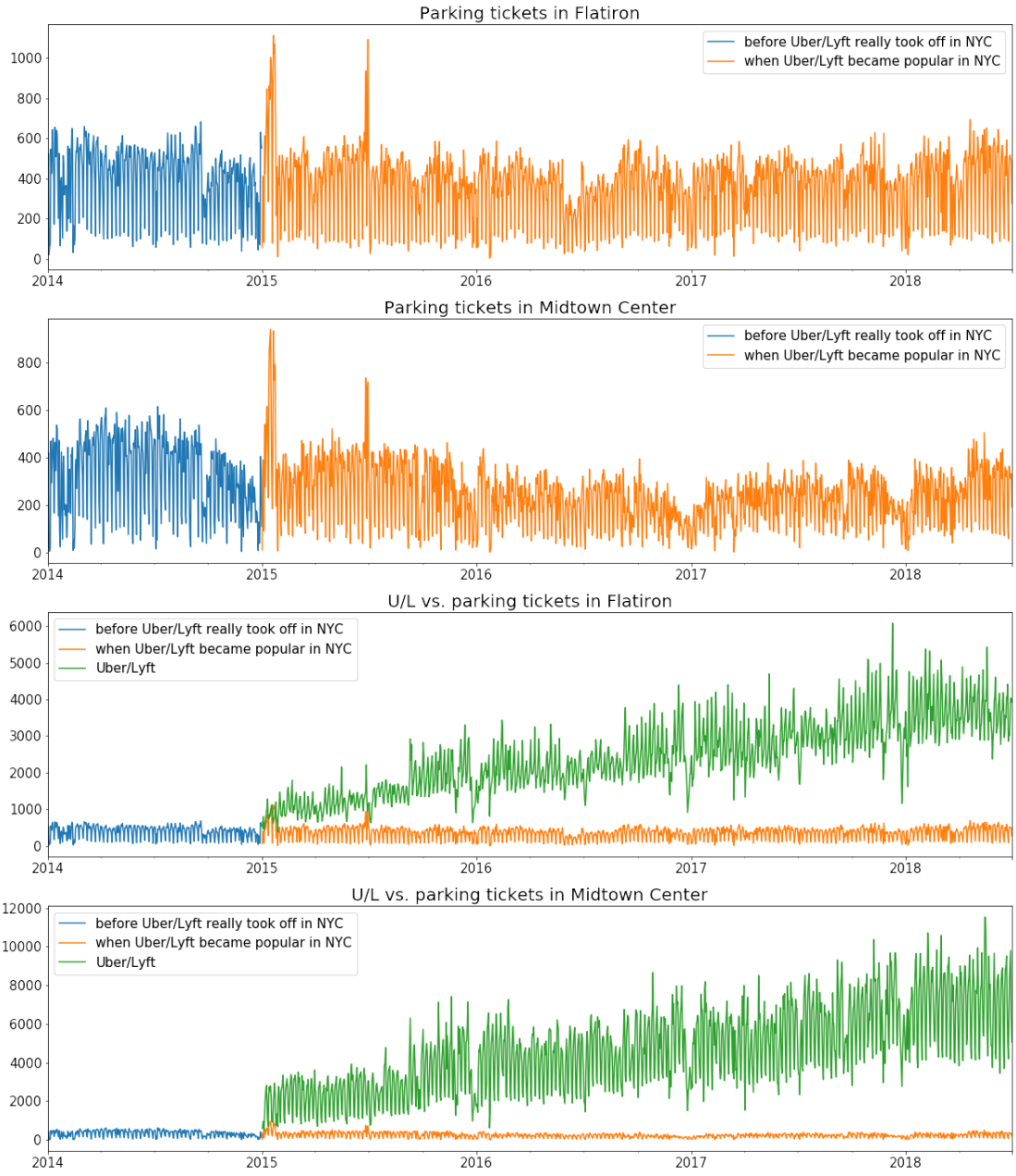


FIGURE 6 Tickets and TNC Trips of a pair of highly homogeneous taxi zones

Based on all qualified pairs, the DID model gave an average causal effect intensity of 224.2, which indicates that 1 parking ticket will be decreased with every 224.2 Uber or Lyft pickups increasing. So, the DID model suggests that Uber & Lyft rides can reduce parking violations in NYC and the effect is much more significantly than the result given by the previously explained FEM.

Pearson correlation analysis on the difference between Uber & Lyft trips and the corresponding difference between parking tickets, and hypothesis testing for non-correlation were used to validate the correctness of DID's conclusion. The results for all 900 qualified pairs among the whole NYC and grouped pairs within each borough are listed in the following **Table 3**:

	NYC	Manhattan	Brooklyn	Queens	Bronx	Staten Island
Coeff	0.021	-0.219	0.063	0.093	0.166	-0.034
P-value	0.534	0.028	0.612	0.420	0.225	0.908

TABLE 3 Pearson correlation analysis and hypothesis testing

The series of testing shows that although the DID concludes that increase of Uber & Lyft leads to decrease of parking tickets in NYC, the correlation analysis, on the other hand, indicates number of Uber & Lyft trips is positively related with number of parking ticket in overall NYC and three specific boroughs, Brooklyn, Queens, and Bronx. However, it should be noted that the statistical significance of this positive correlation is very low according to the very large P-values. But at the same time, Manhattan's result has significantly supported the conclusion of DID: negative correlation with p-value less than 0.05. Conversely, if we only use homogeneous taxi pairs within the Manhattan, DID provides a even higher average estimate of casual intensity: 472. Based on the above results, we believe that at least in Manhattan, the conclusion that Uber & Lyft have a negative causal effect on parking tickets is reliable.

CONCLUSIONS

The main goal of this paper was to explore the causal effect between ride-hailing services represented by Uber & Lyft and parking violations: determining whether Uber & Lyft increases or reduces the number of parking violations in NYC, and how strong the effect is. After a series of data collection, preparation, and analyses, the results demonstrated that Uber & Lyft have negative causal effect on NYC parking violations, suggesting Uber & Lyft are able to reduce parking violation. However, the extent of the impact has not been fully confirmed yet, due to the difference between the result of the FEM and DID. Future improvements could be made to further modeling the scale of the effect.

With more than 100 million users in the world (*16*), Uber & Lyft has far-reaching impacts; some yet to be identified and explored. This paper serves as an example to apply large scale machine learning methods on the available data to study and prove unnoticed yet potential effects of ride-share companies on cities. Reducing parking violation can be regarded as a positive impact, and also have following effects such as the reduction of parking enforcement pressure for the city. In this aspect, Uber & Lyft could be beneficial to NYC, and potentially to other major cities in the world. However, this is only one aspect of Uber & Lyft's influence. More research could be done in this area to further explore other aspects of Uber & Lyft influence in the city.

ACKNOWLEDGEMENTS

This research was supported by Center for New York University's Urban Science + Process and Wagner Graduate School of Public Servic.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: Junjie Cai prepared datasets, assisted in organizing the codes, and estimated Fixed Effect Model and Bayesian Network Model; Junru Lu developed Difference in Difference Model, assisted in organizing the codes, and drafted this manuscript; Pranay Anchan prepared data (Geocode FHV data with google API) and proofread the manuscript; Shijia Gu helped in data collection and preparation, and model development. Yuxuan Wang helped in data collection, model exploration, built the codes repository and proofread the manuscript. Zhan Guo developed the research questions, provided feedback on model development and policy implications, and proofread the manuscript. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Richter, W., *Uber and Lyft are gaining even more market share over taxis and rentals*, 2018.
2. Grocer, S., *How Uber and Lyft Compare, in Four Charts*, 2019.
3. statista, *Ride Hailing*, 2019.
4. Reed, E., *How Much Do Uber and Lyft Drivers Make in 2018?.*, 2018.
5. Brinklow, A., *Lyft, Uber Increase Traffic 180 Percent in Major Cities, Says Report*, 2018.
6. Schmitt, A., *Study: Uber and Lyft Caused U.S. Transit Decline. Streetsblog*, 2019.
7. Henao, A. and W. Marshall, A Framework for Understanding the impacts of ridesourcing on transportation. In *Disrupting Mobility*, Springer, 2017, pp. 197–209.
8. Ackerman, S. S. and R. E. Moustafa, *Red Zone, Blue Zone: Discovering Parking Ticket Trends in New York City*, 2011.
9. Pearl, J., Causal inference in statistics: An overview. *Statist. Surv. 3*, 2009, pp. 96–146.
10. Heckerman, D. and J. S. Breese, Causal Independence for Probability Assessment and Inference Using Bayesian Networks. *IEEE Transactions on Systems, Man, and Cybernetics âĂŞ Part A: Systems and Humans*, Vol. 26, 1996, pp. 826–831.
11. Kim, J. H. and J. Pearl, A computational model for causal and diagnostic reasoning in inference engines. *Proceedirzgs /JCAf-83*, 1983, pp. 190–193.
12. Glenn Firebaugh, C. W. and M. Massoglia, *Handbook of Causal Analysis for Social Research*, Springer, chap. 7, p. 113, 2013.
13. Angrist, J. D. and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press, Vol. 4, chap. 5, pp. 227–243, 2008.
14. Abadie, A., A. Diamond, and J. Hainmueller, Synthetic control methods for comparative case studies: Estimating the effect of CaliforniaâĂŹs tobacco control program. *Journal of the American statistical Association*, Vol. 105, No. 490, 2010, pp. 493–505.
15. Imai, K. and I. S. Kim, When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data. *American Journal of Political Science*, Vol. 63, No. 2, 2019, pp. 467–490.
16. Niu, E., *Uber Has Nearly 5 Times More Users Than Lyft*, 2019.