QUANTITATIVE EFFECTS OF CARSHARING SERVICES ON COLLISIONS IN NEW YORK CITY

A Thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Master of Public Policy
in Public Policy

By

Matthew Emeterio, B.A.

Washington, DC
April 5, 2016
Copyright 2016 by Matthew Emeterio
All Rights Reserved
QUANTITATIVE EFFECTS OF CARSHARING SERVICES ON COLLISIONS IN NEW YORK CITY

Matthew Emeterio, B.A.

Thesis Advisor: Erica Johnson, Ph.D.

ABSTRACT

Using regression analysis on publicly available datasets, this research examines the effects of car-sharing services on collisions in three locations in New York City: the Financial District, Brooklyn, and JFK Airport, while controlling for weather effects and different types of cab services that operate in New York City. Prior studies have focused on the Central Business District, the core of central to lower Manhattan, which has a markedly different profile in terms of collisions and usage rates. In addition to finding both positive and negative statistically significant outcomes in different areas, this research further identifies a number of key areas for data to be collected or made available from both private and public stakeholders for future study in the field.
Special thanks to Sarah, for all the reminders and encouragement,

Matthew Emeterio
# Table of Contents

Introduction .................................................................................................................. 1

Background .................................................................................................................. 2

Literature Review ......................................................................................................... 5

Data and Conceptual Framework ................................................................................. 10

Empirical Strategy ....................................................................................................... 13
  Data and Methods .................................................................................................... 14

Empirical Results ......................................................................................................... 20
  Financial District ..................................................................................................... 20
  JFK ............................................................................................................................ 21
  Brooklyn .................................................................................................................. 22
  Conclusion ............................................................................................................... 23

Works Cited .................................................................................................................. 25
LIST OF FIGURES

Figure 1. FHV Ecosystem, 2016................................................................. 7
Figure 2. Conceptual Framework ............................................................... 11
Figure 3. Collisions ............................................................................. 16
Figure 4. Uber Pickups ....................................................................... 17
Figure 5. Green Cab Pickups ................................................................. 17
Figure 6. Yellow Pickups.................................................................... 18
INTRODUCTION

Taxi services have, in conjunction with mass transit, historically been the canonical way in which New Yorkers navigate their city. In 2006, 241 million passengers rode in taxi services (Schaller 2006) - by 2014 this number had fallen to 236 million (Stiles 2014). The advent and rise of services, largely beginning with Uber’s introduction to the New York City market in May 2011, allowing private citizens to use their own cars and a phone app to act as on-demand ridesharing services has contributed to that decline. These services - Uber, Lyft, Sidecar, and others - provided a litany of benefits over traditional taxi-hailing mechanisms. Of the three mentioned, Uber is overwhelmingly the largest operator in every metric. As a comparison, in October 2015, Uber raised venture capital at a $70 billion valuation while Lyft, the second largest operator in the field, sought a funding round in November 2015 at a $4 billion valuation. The data bears this point out as well - the dataset from fivethirtyeight that aggregates multiple “for hire vehicle” services’ trips per day (Flowers 2015), Lyft provided 267,701 trips, while Uber provided 2,053,535 trips in the same time period.

The discussion over ridesharing services’ contribution to traffic congestion has been both very public and very politicized. In early 2015, Bill DeBlasio, mayor of New York City, proposed a cap on the number of Uber drivers operating in New York City at any given time over traffic congestion concerns. Uber argued that their drivers largely replace trips that would have been undertaken by car, meaning they are net neutral in traffic congestion. The prima facie response to additional cars being on the road is that they must obviously increase congestion - very roughly, Mayor DeBlasio’s argument - but the case may simply be that Uber riders are not taking their own cars, and there is some evidence that Uber (Fischer-Baum 2015) is taking rides
away from taxis. In either case, this paper will provide analysis on the effect of Uber on an important component of traffic congestion - collisions - and ensure that the conversation around regulating ridesharing services is driven by data.

Specifically, this paper will attempt to identify an important measure of traffic congestion: collisions. By looking at a dataset of all Uber pickups for the large majority of 2014 in New York City provided by fivethirtyeight, another dataset provided by the New York Taxi and Limousine Commission on both yellow and green taxi usage, New York City’s Vision: Zero dataset of traffic collisions, and a series of variables that could be expected to impact collisions like weather, public events, and street closures, I will measure the effect that Uber has on collisions in a series of intersections in New York City, from the very densely urban “central business district” in Lower Manhattan to the less dense Brooklyn and Queens, and around the major airports.

BACKGROUND

The ridesharing industry, as a piece of the larger “on-demand economy,” is perhaps the piece most likely to enact long-term changes to the way people do business. While companies like Airbnb (short-term housing rentals), DogVacay (pet-sitting), and TaskRabbit (household help) have gained footholds in their respective markets, none have captured public - and investor - interest like Uber.

Borne out of the founder’s poor experience trying to hail a taxi, Uber operates by using a phone app to connect riders with “driver-partners,” independent contractors using private vehicles to provide rides. The rider requests a ride on their phone, notifying drivers in the area, and a driver accepts the fare, and moves to pick up the rider. By GPS, the rider can watch the
driver’s progress to pick up, or be alerted when the driver is nearby. When the ride is over, payment is handled by a credit card or other payment account already attached to the rider’s Uber profile, and no cash changes hands. Of this payment, Uber retains 20% of the total fare and pays out the rest to the driver.

Because Uber acts as a platform to connect riders and drivers, Uber is incentivized to match demand from riders to supply by drivers, and it does this through a process called surge pricing. In scenarios with high demand for rides - e.g., a concert ending - Uber algorithmically enacts surge pricing, which places a multiplier on the cost of the ride. The multiplier, based on geographic zones and recalculated every five minutes (Chen 2015), scales based on demand, so using Uber for a ride on New Year’s Eve in Times Square could cost multiple times what it could cost outside of a surge. Surge pricing has multiple functions: first, it prices some users out, reducing demand at the rider level. Second, it draws drivers into the “market” by directly increasing their per-hour earnings. The surge pricing process allows supply and demand to be matched more efficiently, and coupled with making the transaction process for ridesharing as simple as possible, Uber and other ridesharing services tapped into a market that was, for decades, inconvenient at best. The various taxi commissions responded with partnerships and tools like Hailo that emulated Uber’s ride-hailing, but the options presented never became standardized and have had limited application.

The academic study of traffic congestion has been largely limited to the realms of urban planning and traffic engineering, and policy proposals and analysis therein have been tangential to those fields. The policy proposals that have surfaced are often designed to generate revenue rather than decrease demand for driving - toll roads that scale pricing based on traffic, and
similar “congestion pricing” (“What Is Congestion Pricing?” 2014) methods applied to cities, but both of these are only encountered after drivers are already on the road, and are simply treated as a cost of owning a car. Furthermore, these tools can be deeply regressive - a sort of “sin tax” for driving. A study by the RAND Corporation in 2008 of traffic in Los Angeles noted that “even 2 percent to 3 percent fewer cars on the roads could reduce congestion by 10 percent to 15 percent.” (“Comprehensive” 2008) The full study continues, saying “automotive travel is also underpriced, in the sense that drivers are not required to bear the social costs associated with their travel decisions.” (Sorensen 2008) Uber’s application of surge pricing may coincidentally be addressing this - if surge pricing is keeping an appropriate amount of drivers on the road for the exact amount of rides needed and pricing it appropriately, then traffic congestion could be reduced by Uber’s presence in a market. There is a natural amount of slack in the surge algorithm, and it may speak more to the value of public and alternate forms of transportation, but the logic of appropriate pricing reducing the total amount of cars on the road, and thereby reducing congestion, is sound.

Finally, a research report by the Maryland Department of Transportation State Highway Administration notes in the summary of its literature review that “it can be determined that traffic volume, as a surrogate variable of congestion, plays a significant role in accident frequency, rate, and severity analyses. Some significant relationships were identified including the relationship that a higher traffic volume usually results in higher accident frequency.” (Chang 2003) This paper explores and hopes to build on the findings of the Maryland report, by isolating the effect of Uber pickups - and therefore, Ubers on the road - on traffic accidents in various locations in New York City. So while there is a sort of tautological link between collisions and
congestion, and because the mere proposal of a policy prescription from Mayor DeBlasio ignited a dramatic response from the company - in the Uber phone app, riders were shown a version of what “DeBlasio’s Uber” looked like, with 25 minute wait times and were entreated to email their local representatives - the study of Uber’s contribution to collisions, and by association, congestion is core to policy proposals for not just Uber and New York City, but other cities, the ridesharing industry, as well as taxis, public transportation, and the broader pursuit of policy in urban planning.

LITERATURE REVIEW

Because of the novelty of the industry, the depth of literature in the “on-demand economy” compared to healthcare, education, or other established industries is relatively shallow. The only major pieces of academic literature associated with the Uber specifically is “An Analysis of the Labor Market for Uber’s Driver-Partners in the United States” (Hall 2015) by Hall and Kreuger, looking at the effective labor force of the Uber driver pool, and “Peeking Beneath the Hood of Uber” (Chen 2015) by Chen, Mislove, and Wilson, which tries to “reverse engineer” Uber’s surge pricing methodology.

Less traditional academic research has been done on the effects of Uber on traffic at fivethirtyeight, but the articles by Fischer-Baum, Bialik, Silver, Flowers, and Mehta have provided interesting insight into the activities of Uber in New York. The article “Uber Is Taking Millions of Manhattan Rides Away From Taxis” argues that “the ride-share service probably isn’t increasing congestion,” (Fischer-Baum 2015) coming to the conclusion that the rides in the city core are largely Uber replacing taxis, while the Brooklyn and Queens are seeing net increases in total pickups. The major caveat for this article, as it specifically relates to traffic
congestion is laid out as such: “it’s also possible that many more empty cabs are cruising Manhattan streets than last year, although it would not be financially viable for drivers - who primarily rent their cabs - to keep this up for very long.” (Fischer-Baum 2015) If, as the article states, Uber drivers are accounting for nearly four million new pickups in the core of Manhattan from 2014-2015 while taxi drivers are cruising while empty, by definition there should be an increase in congestion, with more cars operating in the same physical space.

The other fivethirtyeight pieces look at where New York City’s Green Cabs - cabs that are explicitly excluded from Manhattan’s core – operate (Fischer-Baum 2015), the intersection between Uber’s growth and public transit (Silver 2015), and an analysis of Uber’s operation in the outer boroughs compared to taxis (Bialik 2015). These pieces are less salient to the discussion and analysis of congestion and associated traffic issues, but provide important context to the activity of Uber in the city. The final fivethirtyeight article, “The Debate on Uber’s Impact In New York City Is Far From Over,” focuses on the political machinations around the effects of Uber in New York City, which is at the heart of this paper, as explained in the Introduction and Background sections.

In 2015, the New York City Office of the Mayor commissioned McKinsey & Company to produce a report (“For-Hire Vehicle Transportation” 2016) on for-hire vehicles operating in the city, and the report was published in January 2016. The McKinsey report differentiated for-hire vehicle services - cabs, black cars, livery, and for-hire vehicle services - and differentiated them by access type: traditional, e-hail/e-dispatch, and street hail.
The McKinsey report identified several “capacity reducing factors,” and analyzed their effects on congestion in the Central Business District in Manhattan, south of 60th. The report found that reductions in speeds, an analogue for congestion, had been primarily induced by “increased freight movement, construction activity, and population growth.” Furthermore, E-dispatch services were identified as contributing to congestion, but “did not drive the recent increase in congestion in the CBD,” however, McKinsey found that E-dispatch services “could drive modest growth in congestion in the future.” The report further found that yellow cab utilization, that is the ratio of number of miles travelled with a passenger to the number of miles travelled without a passenger, has remained at “approximately 54%,” despite a decline overall usage, which are largely being substituted by e-dispatch services. On congestion, the report also found that ride-sharing, carpooling services rendered by e-dispatch - Lyft Line, Uber Pool, and others - were both becoming more popular and may have “inadvertent, detrimental impacts on congestion,” depending on whether riders were coming from car or public transit options: “the congestion mitigation of an 11-13% switch in yellow taxi and e-dispatch trips over to ride-
sharing would be completely offset if less than 1% of public transit riders also switched to ride-sharing.”

The McKinsey report also produced a number of findings that are policy-relevant in New York’s case. First, for-hire vehicles are not expected to significantly change air quality in New York, largely as a function of stricter emissions standards. Next, the rise of e-dispatch services will reduce the number of for-hire vehicles that are wheelchair accessible, as private cars first supplant taxi rides, and as the supply of both green and yellow cabs drop. Finally, every taxi ride that is replaced by a for-hire vehicle reduces the pool of available funds for a number of transit-related projects, from accessibility to mass transit; however, services like Uber and Lyft do pay into traditional sales tax funds, and a portion of this is directed specifically to MTA expenditures.

There are two caveats to consider in the first section of the literature: first, while fivethirtyeight makes their data for articles open, neither those articles, nor the McKinsey report are peer-reviewed. Furthermore, while the McKinsey report approaches the effects of car-sharing services from the angle of congestion, an important point to consider is the focus on lower Manhattan. As our conceptual framework will expand on, different areas of New York have substantially different traffic profiles - the CBD often being at “carrying capacity” for traffic, while the less-densely urban boroughs have more variance in the flow of traffic rates.

Since the mid-1990s, modern assessments of the built environment and traffic engineering have used statistical analysis to identify risk factors and predict outcomes from the large datasets produced by governments. The Transportation Research Board, part of the United States National Research Council, has done significant basic research into traffic engineering, and much of the work done in the last decade is derivative of or otherwise based on the baseline
work of the TRB, particularly “Quantifying Congestion” (Lomax 1997). While much of the TRB’s work revolves around highway planning, the journal *Accident; Analysis and Prevention* has a range of articles on both the root causes of collisions and the variety of roads where they happen, and it seems as though some of the larger body of literature has conflicting results. In “Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England,” (Wang 2009) the authors note their “robust technique [. . .] developed to map M25 accidents onto its segments,” and that “since existing studies have often used a proxy to measure the level of congestion, this study has employed a precise congestion measurement.” This is essentially the core of my work as well, but in my case, I am mapping Uber pickups to intersections in New York City - segmented analysis - and observing the collision rate, which is a proxy to congestion.

An interesting note from Wang, Quddus, and Ison is that “the results suggest that traffic congestion has little or no impact on the frequency of road accidents on the M25 motorway. All other relevant factors have provided results consistent with existing studies.” Done in 2009, this directly contradicts the findings from a paper done for the Maryland Department of Transportation’s State Highway Administration, in conjunction with University of Maryland, College Park’s Civil Engineering department by Dr. Gang-Len Chang in 2003. The results of Dr. Chang’s report disagree with the Wang, Quddus, and Ison results twice:

The estimation results, based on the available sample data, reveal that accident rates on local arterials tend to decrease with an increase in traffic volume. In contrast, accident rates on freeway segments during peak hours indicate a positive correlation with traffic volume per lane. Additionally, freeway accident rates
during off-peak periods appear to be random in nature, and not necessarily correlated to any specific factors.

The disparity between these outcomes is notable because the Chang paper disagrees with the Wang, Quddus, and Ison paper in both directions - counterintuitively, that on local arterials, accidents decrease with an increase in traffic volume, and that freeway segments have accident rates increase with increases in traffic volume. Both papers looked at their data using statistical analysis and looking at similar areas of study, came up with opposite results. This may be the result of a number of functions: there is some core difference between drivers in Maryland and England, the data collected was mechanically different, or pure statistical anomaly. I suspect that similar results may appear in my work as well, there is a compelling argument that, as a function of the amount of traffic, there are enough collisions in Manhattan’s central business district that the addition of Uber pickups will not statistically significantly modify that in any way. To that end, I will be emulating the WQI paper in segmenting New York City into discrete areas of study, and looking at intersections in areas less congested in less traffic-dense areas - Brooklyn, Queens, and JFK airport for evidence of statistical significance.

DATA AND CONCEPTUAL FRAMEWORK

The proposed conceptual framework is comprised of several functional pieces: collisions, weather, taxi pickups, and Uber pickups. The data itself will be trimmed by latitude and longitude to areas of interest, per regression, and I will generate three different analyses: one for the central business district in lower Manhattan, one for Brooklyn, and one for JFK Airport. The boundaries are a rough outline of the areas of interest,
Collision data is drawn from the New York Vision: Zero dataset, and will be the dependent variable in the regression. It contains a number of variables related to cyclist and pedestrian collisions, so the dataset will be trimmed to only include collisions that involve vehicles, ruling out cyclist/pedestrian collisions.

Weather data is drawn from WeatherUnderground.com by a custom script. It is an hourly pull from KNYC, a weather station in Central Park. It includes a number of weather-related variables, but for the purposes of the regression, the data is collapsed into a dichotomous variable based on whether or not there is precipitation in the hour of interest. Using data from a single location in New York City has the potential for introducing some bias - it may be raining in Central Park but not at LaGuardia - but it represents a centralized location in the city that in
absence of more granular weather data, provides a rough estimate of the overall weather condition in New York City.

Taxi pickup data is drawn from the NYC Taxi and Limousine Commission datasets on Yellow and Green cab activity in New York City. For clarification, Green cabs are ones only permitted to operate in the outer boroughs and north of East 96th and West 110th in Manhattan, while Yellow cabs are permitted to operate in lower Manhattan and the airports. This data includes a large number of factors related to trip length, fare totals, and pickup and dropoff locations, but for the purposes of the regression, the data will be trimmed to only include time, date, and pickup latitude and longitudes.

Finally, the Uber pickup data is the final component of the regression. This data, initially from the NYC Taxi and Limousine Commission, was requested for release through New York’s Freedom of Information Law by the journalists at fivethirtyeight. The data includes pickup time and date, latitude and longitude, and the TLC “base company” code attached to the pickup. For purposes of the regression, the data will be limited to time, date, and location.

Their hypothetical relations to each other consist exclusively of positive correlations. With automobile collisions being our dependent variable, all of the variables of interest – Uber pickups – and controls variables have a positive correlation to collisions. The point of interest with automobile collisions as a variable is the feedback loop with road closures – closures hypothetically generating collisions, which could further generate more closures.

The final independent/control variable I am using in the model is inclement weather. While inclement weather can lead directly to collisions because of reduced visibility, more
difficult mechanical control, and other reasons, it also contributes to road closures – e.g., snow routes – and to “traffic outsourcing” by way of Uber pickups and taxi hailing.

The piece of data that is difficult to acquire, or even quantify, is “traffic congestion.” In this model, we use collisions as a proxy for congestion, but they are not the same, and can be dependent on each other. In the model, collisions fulfill the role of proxy reasonably well, fitting into the conceptual framework in the same way that congestions would, but there are other datasets that would be similarly useful as a proxy for congestion. For example, mechanically in terms of effect on congestion, average traffic speed works in the same way in the model as collisions, and would be an interesting dataset to include, but the data does not have the breadth or depth in granularity to be useful, when it does exist. Other data that would be useful to incorporate into the model would be mass transit onboarding numbers. The correlations between mass transit usage rates and public events, road closures, and inclement weather would be an important piece of information as it would likely take substantial explanatory power away from the Uber and taxi usage variables, but again, the data that exists is not granular enough to justify involving in the model.

**EMPIRICAL STRATEGY**

The empirical strategy I am using is a standard OLS regression, with the dependent variable of traffic collisions and independent variables of Uber pickups, taxi pickups, public events, road closures, and inclement weather. Weather is the least granular variable as it is measured hourly, and the unit of analysis is collisions in an area per hour. The format of the regression is as follows:

\[ Y_{\text{collisions}} = \beta_{\text{Uber Pickup}} + \beta_{\text{Weather}} + \beta_{\text{Green Pickup}} + \beta_{\text{Yellow Pickup}} + \mu \]
My goal, and a place for further research, is to identify a statistical tool to compare the regression results across geographic areas. This is driven out of my concern that places like the Financial District in downtown Manhatten will likely see no statistically significant results in any direction: the intersection of, for example, Broadway and Wall Street has likely maxed out on “carrying capacity” at any given time, so adding Uber pickups will not have any statistical effect on the level of traffic collisions seen in that particular part of the city. More residential areas in the other boroughs, however, may be more susceptible to collisions as a function of increased Uber pickups. The goal then, would be to prove that $Y_{\text{collisions}_{\text{Financial District}}}$ is statistically significantly different from $Y_{\text{collisions}_{\text{Brooklyn}}}$ which may require more sophisticated statistical techniques.

**Data and Methods**

Our analysis is based on five datasets. The first is the collision dataset (“NYPD Motor Vehicle Collisions” 2015) as provided in the NYC Open Data portal. It provides information on date, time, latitude, longitude, and assorted incident data like vehicles involved and borough. The second is the dataset of Uber pickups by latitude and longitude and time as provided by fivethirtyeight via a FOIA request from New York City (Flowers 2015). The third and fourth are information on yellow and green cabs with pickup latitude and longitude, time, date, and a number of other data points associated with cab rides (“Trip Record Data” 2015). The final dataset is a custom script scraping data from Weather Underground’s reporting from the Central Park weather station, which provides precipitation and condition information by hour, allowing for integration of weather conditions. A caveat about the weather data: Central Park was chosen
for generalizable weather data with respect to the rest of the city. In short, if there is precipitation in Central Park, there is likely to be precipitation in Brooklyn, the Financial District, and elsewhere in the city within the hour.

The first phase of the data manipulation was trimming data by location. To do this, I first identified the latitude and longitude coordinates of the locations I was interested in: Brooklyn\(^1\), the Financial District\(^2\), and JFK airport\(^3\). From there, we trimmed the data by latitude and longitude for each location. As the Financial District and Brooklyn are more loosely defined geographically, in the case of JFK airport, I included a distance around the airport, to capture data local to the airport itself. The simplest trimming mechanism was to simply identify a rectangle and trim to data points with a latitude and longitude that fell within that rectangle. More sophisticated trimming mechanisms would have capture more location-specific data with diminishing returns.

The next step of data manipulation was generating dummy variables indicating that a pickup had occurred, a collision had taken place, or that there had been precipitation in a given location in a given hour. Once those were generated, they had to be joined on a unique identifier. In this case, I generated `hour_of_may`, a variable that uniquely tracked data across individual hours in May, and the disparate datasets were then merged to produce a single one to do analyses.

---

1 Latitude: 40.57N through 40.7N, Longitude: 74.0412W through 73.8612W  
2 Latitude: 40.70416N through 40.7194N, Longitude: 73.998W through 74.0196W  
3 Latitude: 40.6404N through 40.675N, Longitude: 73.7532W through 73.8108W
The datasets are relatively unique in that descriptive statistics are ineffective in assessing the full state of the data; for example, the standard deviation of pickup location of Ubers in New York is not illuminative of the state of the world. However, the data is susceptible to mapping, and included are maps of Uber pickups, yellow and green cab pickups, and collisions. Each point represents a pickup or collision, per map.

Figure 3. Collisions
Figure 4. Uber Pickups

Figure 5. Green Cab Pickups
The maps highlight a number of interesting elements based on a purely visual estimate. First, there are a number of green cab pickups south of W110th and E96th; this is primarily interesting in that Boro Taxis - green cabs - “can be dispatched to pick you up in northern Manhattan, the Bronx, Queens, Brooklyn, and Staten Island and at the airports, but may not pick up any trips - pre-arranged or street hail - in the Manhattan exclusionary zone.” (“Boro Taxis”)

As expected, collisions are much more densely packed in the Financial District, becoming more disperse in the other boroughs. This comports with the hypothesis in that the Financial District has unique elements to congestion that make traffic uniquely bad, while in the more residential boroughs, collisions are less frequent, with greater room for variance. Additionally, both yellow and green cab pickups follow the arterial roads tightly, whereas Uber pickups, especially in the outer boroughs, are much more dispersed into side streets. This is a
function of Uber’s ride hailing app compared to traditional street hails of green and yellow taxis. Uber seems to also better serve distant communities like Coney Island than yellow cabs especially. Finally, a point to take into consideration for analysis: collisions in the departure and arrival zones for terminals at JFK are not be reported, while pickups of all types are. This road, the Van Wyck Expressway, is maintained by the Port Authority, and collisions are presumably not logged as collisions for New York’s overall numbers, affecting the analysis on the area surrounding JFK Airport.

In terms of limitations, the overarching problem is incomplete, non-archived, and unavailable data, causing omitted variable bias. A number of datasets would have been useful to incorporate into the analyses, including average traffic speeds, road closures, public events, construction, and street resurfacings from the public side, among others.

Public events can generate road closures – this is often a component of hosting an event in a public space. The reason for including public events into the model for predicting automobile collisions is largely the same as for including road closures: by forcing drivers into something outside of their routine, we increase the likelihood of mistakes being made. The canonical example being a driver trying to look at their phone’s GPS to navigate and drive at the same time, so whether it’s a road closure from construction or a road closure from a parade, I would hypothesize an increase in collisions.

Both road closures and public events feed into the larger world of “traffic outsourcing” – hailing a taxi or calling an Uber. In short, rather than driving during a public event or road closure, people may simply hire someone else to help them complete their transit. The rationale for including them in a model that predicts collisions is that both processes disrupt traffic flow,
but in a different way from public events and road closures do. To make a pickup, both Ubers and Taxis have to either stop mid-street creating congestion or pull to the side, creating an opportunity for cars to go around them, perhaps a worse option for safe driving than simply coming to a complete stop to do their pickup.

Uber dropoff data, even locally anonymized, would have been helpful as well, as both types of taxis provide dropoff data, and it seems likely that dropoffs are similarly “dangerous maneuvers” to pickups in terms of analyzing contributions to collisions. Given that only cab dropoff data was available, it seems that analyzing that data without a comparable set from Uber would disproportionately place predictive “burden” on the cabs.

**EMPIRICAL RESULTS**

For the regression itself, I ran a basic OLS estimates regression with collisions as the dependent variable and uber_pickup, knyc_precipitation, green_pickup, and yellow_pickup as independent variables. I ran this regression against the three different datasets, first on the Financial District geocoded data, then on the JFK geocoded data, and finally on the data geocoded around Brooklyn.

**Financial District**

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) collision</th>
</tr>
</thead>
<tbody>
<tr>
<td>uber_pickup</td>
<td>-0.0188***</td>
</tr>
<tr>
<td></td>
<td>(0.000172)</td>
</tr>
<tr>
<td>knyc_precipitation</td>
<td>-0.0200***</td>
</tr>
<tr>
<td></td>
<td>(0.000175)</td>
</tr>
<tr>
<td>green_pickup</td>
<td>-0.0182</td>
</tr>
<tr>
<td></td>
<td>(0.0253)</td>
</tr>
</tbody>
</table>
In the Financial District, for both **uber_pickup** and **knyc_precipitation**, there are negative, highly statistically significant (p<0.01) coefficients. The model predicts a decrease of 1.8 accidents for each 100 additional Uber pickups per hour. **yellow_pickup**, the variable for yellow cab pickups, has a highly statistically significant, positive coefficient, and the overall model’s R-squared is 0.017.

### JFK

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>collision</th>
</tr>
</thead>
<tbody>
<tr>
<td>uber_pickup</td>
<td>-0.0107***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000189)</td>
<td></td>
</tr>
<tr>
<td>knyc_precipitation</td>
<td>-0.0113***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000129)</td>
<td></td>
</tr>
<tr>
<td>green_pickup</td>
<td>0.00258</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00352)</td>
<td></td>
</tr>
<tr>
<td>yellow_pickup</td>
<td>-0.000315</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000197)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.0114***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.78e-05)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>14,769,860</td>
<td></td>
</tr>
<tr>
<td>R-squared</td>
<td>0.001</td>
<td></td>
</tr>
</tbody>
</table>

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1
For JFK, `uber_pickup` and `knyc_precipitation` are the only statistically significant coefficients at $p<0.05$, and both negative. For an additional hundred Uber pickups in the reporting area in an hour, the model predicts a decrease of slightly over one collision.

`yellow_pickup` is also worth noting, coming in with a $p$-value of .11, just under a 90% significance level. The R-squared for JFK is 0.001.

**Brooklyn**

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) collision</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>uber_pickup</code></td>
<td>0.0135***</td>
</tr>
<tr>
<td></td>
<td>(0.000365)</td>
</tr>
<tr>
<td><code>knyc_precipitation</code></td>
<td>0.0330***</td>
</tr>
<tr>
<td></td>
<td>(0.000562)</td>
</tr>
<tr>
<td><code>green_pickup</code></td>
<td>-0.0382***</td>
</tr>
<tr>
<td></td>
<td>(0.000285)</td>
</tr>
<tr>
<td><code>yellow_pickup</code></td>
<td>-0.00238**</td>
</tr>
<tr>
<td></td>
<td>(0.000945)</td>
</tr>
<tr>
<td><code>Constant</code></td>
<td>0.276***</td>
</tr>
<tr>
<td></td>
<td>(0.000139)</td>
</tr>
</tbody>
</table>

Observations 14,769,860
R-squared 0.001

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard errors in parentheses</strong></td>
<td></td>
</tr>
<tr>
<td>*** $p&lt;0.01$, ** $p&lt;0.05$, * $p&lt;0.1$</td>
<td></td>
</tr>
</tbody>
</table>

The final model, for Brooklyn, reports variables are highly statistically significant ($p<0.01$), with `uber_pickup` and `knyc_precipitation` being positive and `green_pickup` and `yellow_pickup` being negative. The model for Brooklyn provides an R-squared of 0.0015, and estimates the effect of 1.3 additional collisions for an additional 100 Uber pickups. While this model matches my hypothesis, that the streets are not at carrying capacity and have room for variance in collisions, this is only one of a number of possible explanations for the variance in
collisions, and does not have the statistical weight of a panel study, and should be treated as such.

CONCLUSION

Each instance of running the model on different geographic sets found R-squared results of less than .01, indicating that the model predicts less than one percent of variance in collisions. Furthermore, effect sizes were measured in hundredths, if not thousandths, indicating very small marginal effects of, for example, an additional Uber pickup in an area in an hour. Even 100 additional Uber pickups in an hour in Brooklyn would only predict 1.3 additional collisions, and based on R-squared, the model only predicts 0.15% of collision variance. Much of this is tied to the size of the data - nearly 15 million yellow cab pickups, hundreds of thousands of Uber pickups, and deep issues with variables being omitted because of insufficient data or access result in very small effect sizes and very little in the way of explanatory power for the variance in the collisions variable.

It is difficult to assess policy implications of any variety based on the results above - a wide number of variables contribute to traffic collisions, from intersection design and number of potholes to driver training requirements and distracted driving, and my models looked at four explanatory variables total. In using those four, this paper lays groundwork for a full assessment of availability of data about street transportations of all types from both the public and private sectors. In identifying avenues for future research, with additional data - not just observations, but explanatory variables - we would be able to model collisions more successfully with statistical tools. Additionally, while this would require cooperation with Uber and other ride-sharing services, developing a panel study with collision and explanatory variable data from
multiple cities so as to account for city-unique effects would allow for research into causality, an overarching goal for stakeholders in both the public and private sectors.
WORKS CITED


"For-Hire Vehicle Transportation." For-Hire Vehicle Transportation Study (n.d.): n. pag.


