2016 IEEE International Congress on Big Data

Green cabs vs. Uber in New York City

Lasse Korsholm Poulsen[1], Daan Dekkers[1], Nicolaas Wagenaar[1], Wesley Snijders[1], Ben Lewinsky[1]
Raghava Rao Mukkamala[1] and Ravi Vatrapu[1,2]
[1]*Copenhagen Business School, Denmark and* [2]*Westerdals Oslo School of Arts, Comm & Tech, Norway*
{*rrm.itm, rv.itm*}*@cbs.dk*

***Abstract*—This paper reports on the process and outcomes of big data analytics of ride records for Green cabs and Uber in the outer boroughs of New York City (NYC), USA. Uber is a new entrant to the taxi market in NYC and is rapidly eating away market share from the NYC Taxi & Limousine Commission's (NYCTLC) Yellow and Green cabs. The problem investigated revolves around where exactly Green cabs are losing market share to Uber outside Manhattan and what, if any, measures can be taken to preserve market share? Two datasets were included in the analysis including all rides of Green cabs and Uber respectively from April-September 2014 in New York excluding Manhattan and NYC's two airports. Tableau was used as the visual analytics tool, and PostgreSQL in combination with PostGIS was used as the data processing engine.Our findings show that the performance of Green cabs in isolated zip codes differ significantly, and that Uber is growing faster than Green cabs in general and especially in the areas close to Manhattan.We discuss meaningful facts from the analysis, outline actionable insights, list valuable outcomes and mention some of the study limitations.**

***Keywords*-Big social data, Social set analysis, Social business, Visual analytics, geo-spatial, GIS, Taxi, Green cabs, Uber.**

I. Introduction

The NYC Taxi & Limousine Commission (NYCTLC) is a governmental agency created in 1971, and is responsible for the licensing and regulating of New York City's yellow taxicabs, for-hire vehicles, para-transit, commuter vans and other luxury limousine services. The NYCTLC licenses and regulates approximately 50,000 vehicles and counts 100,000 drivers. The paper presented here will focus on the Green cabs, that were introduced by the Five-Boro Taxi Plan, a NYCTLC initiative that aims to meet the demand surplus for taxi rides in the outskirts of New York City.

In August 2013, the NYCTLC introduced a fleet of Green cabs to the city of New York. These Green cabs were introduced with the goal of providing the residents of Brooklyn, Queens, the Bronx, and Upper Manhattan more access to metered taxis. Considering that the Yellow cabs prefer to operate in the areas of NYC that are most dense in pick-ups (Manhattan and the airports), the availability of Yellow cabs tends to be low in the outer boroughs of NYC. Hence, Green cabs are not allowed to pick up street hails from the largest part of Manhattan (below 110th St. on the West Side, and below 96th St. on the East Side), or either of JFK or LaGuardia airports. However, they can pick up hails anywhere outside of these areas. Perhaps most importantly, Green cabs can be ordered (via phone or online) for a pick-up anywhere in the city, unlike Yellow cabs which can only be hailed from the street.

In May 2011, the rideshare service Uber expanded into NYC, severely influencing the taxi market. With the amount of total Uber pick-ups increasing from 2 million in April-June 2014 to 8 million in April-June 2015 [1], Uber can be called a success from a market share perspective. Research by Bialik et.al. [1] has shown that Uber has been replacing Yellow cab rides in the centre of the city (Manhattan), and supplementing Green cab rides in the outer boroughs of NYC. It is in these outer boroughs where Uber has realised most of its growth. Thus, Green cabs have been facing fierce competition from Uber since its introduction, with Uber offering lower fare rates and easier accessibility through the mobile application [2]. Situated within this real-world context, this research paper aims to analyse in which specific outer boroughs Green cabs are underperforming Uber cabs. More specifically, a visual analysis of Green cab rides and Uber cab rides over 6 months in 2014 outside of Manhattan and the two airports will be conducted to reveal demographic and geographic usage patterns.

A. Problem Formulation and Research questions

The introduction of the Green taxis has been successful when compared to the decreasing performance of the Yellow taxis [1]. One of the NYCTLC's biggest competitors is Uber, who services both central Manhattan and the outer boroughs of NYC in one integrated app. Initial reports suggests that Green taxis were gaining success in particular regions of NYC [3]; it would be valuable to the NYCTLC if the percentage of rides delivered in certain zip codes were indexed to the same percentage delivered by Uber. For example, if 15% of all of Uber's outer borough pickups occur in Williamsburg and only 5% of Green cab pick-ups occur in the same zip codes, it suggests that the NYCTLC's marketing efforts should be increased in that region. This leads to our first research question: *Using trip data records, how does NYCTLC's share of rides per zip code compare to Uber's in the outer neighbourhoods of New York?*

After identifying which zip codes are under-indexed when compared to Uber, we consider the socio-demographic properties of the specific areas. Thus, our second research question is: *For the New York Taxi Commission's under-*

978-1-5090-2622-7/16 $31.00 © 2016 IEEE
DOI 10.1109/BigDataCongress.2016.35

represented zip codes, what socio-demographic attributes are most prevalent in the New York Taxi Commission's under indexed zip codes?

Recognising that socio-demographic properties may impact a customer's behaviour, we decided to explore the potential impact of time on a customer's choice between a Green cab and an Uber. Thus, our third research question is: *Does customer preference change with respect to time of day (night/day or weekend/weekday)? Does this vary between zip codes?*

There are many other competitors in the taxi marketplace for NYC. However, we believe it to be a market that shows characteristics of diverging on a duopoly and that the information supplied from both companies give a thorough overview of the state of the market. Therefore, for the purposes of this paper, we only consider data from the NYCTLC's Green cabs and Uber.

II. Related Work

The taxi industry is highly regulated and a great deal of literature exists detailing the implications of such stringent rules and suggested changes [4], [5]. In more recent years, a large amount of Freedom of Information Law (FOIL) requests made the NYCTLC decide to make trip data readily available to everyone to download from their website. The availability of the data has resulted in many research projects being conducted in recent years. We intend to build on the existing work and try to explore new research questions as stated above with a focus on providing insights for the modern competitive environment. In the next few paragraphs, we present a selected review of related work on the industry sector, similar research questions and data types.

Kim [6] explored the use of car sharing services in the lower income regions of NYC and found that car sharing in low-income neighbourhoods did not differ from typical car sharing locations. As such, affordability is important when trying to understand why NYCTLC or Uber may be underrepresented in certain boroughs of NYC. It is critical to consider the possible impact of the existing public transport system and other car sharing services when drawing conclusions from the dataset.

Yazici, Kamga and Singhal [7] explored ways in which the service for taxi customers at JFK Airport can be improved. Savage and Vo [8] assessed that drop-offs in part of outer Manhattan by far outweighed the number of pickups suggesting that the pickup service being delivered the boroughs of NYC was most likely to be greatly less than optimal. Whilst the introduction of the Green cabs would have assisted in solving this problem it is important to consider that a great deal of the Green cabs brand equity can be diluted if yellow taxis who make drop-offs in the suburbs deliver a lower quality service than what the Green cabs strive to do. Ferreira et. al. [9] worked with the same dataset as above and grouped trips based on their zip codes in order to display ride information in a more precise manner. Similarly, this study intends to use to PostGIS to store the bounding polygons of all NYC zip codes as geometric data types in the database, which can then be used to group rides by zip code.

We utilise Zhu, Gorman & Horel [10]'s technique for combining geospatial and demographic analysis. They conducted a geospatial analysis in order to examine the relationship between alcohol outlet density and the occurrence of violence. In order to do so, they make use of three data sets. First, for the active alcohol outlets the authors obtained a dataset from the Texas Alcoholic Beverage Commission. Each record included the name, geographic location and type of permit or license of the outlet. Second, the data set regarding crime records counts the different types of crime (murder, rape, robbery and aggregated assault) and the location of the crime on a monthly basis. Third and last, the authors make use of socio-demographic data sets of 12 different neighbourhoods. They found that there is a clear association between alcohol outlet density and violence, and the authors suggest that alcohol availability and access are fundamental to the prevention of alcohol-related problems within communities

Funk et.al [11] used a choropleth map to display a potential relationship between animal Ebola counts and human Ebola outbreaks across different geographical regions in order to allow for better disease control. With the choropleth map the authors have been able to identify the African regions in which wild animals are most often diagnosed with Ebola. Moreover, they identified the places where human Ebola outbreaks occurred. One can note from the map that the areas in which Ebola is most frequently diagnosed among wild animals also counts the most human Ebola outbreaks. We employ the technique of a cloropeth map to understand taxi ride characteristics in terms of demographics and temporal dependencies, if any.

III. Methodology

Before the introduction of the Five-Borough Taxi plan, Yellow taxi cabs were the only vehicles allowed to pick up passengers through hailing. However, due to the limited number of Yellow cabs (13237 vehicles in 2013) and the high passenger density of central Manhattan, Yellow cab drivers were rarely providing hail taxi services in the outer neighbourhoods. It is estimated that, before the introduction of the Five-Borough Taxi plan, roughly 97% of the Yellow taxi rides begun their rides in central Manhattan, while approximately 80% of the New York Population lives outside of Manhattan [12] as shown in Fig. 1.

The Green cabs introduced by the Five-Borough Taxi plan are not allowed to pick up customers through hailing in Inner-Manhattan and the different airport areas, making the outlying neighbourhoods of New York their primary pick-up areas. The Five-Borough Taxi plan allowed for 18.000

street hail permits for Green cabs, which were released in three batches of 6,000 over a period of 3 years. Please see an overview of the pick-up areas of the different types of taxis below.

A market analysis conducted by the TLC in 2015 shows that the Five-Borough Taxi Plan and its Green cabs are tapping into the earlier identified unmet demand [12]. Up until June 2015, TLC has issued 8050 Green cab licenses. The report additionally mentions that the unmet passenger demand remains strong, even though an increasing number of Green cabs are being introduced. Lastly, trip data shows that the current fleet of Green cabs is too small to meet existing demand, arguing for an expansion of the current fleet of 8050 Green cabs.

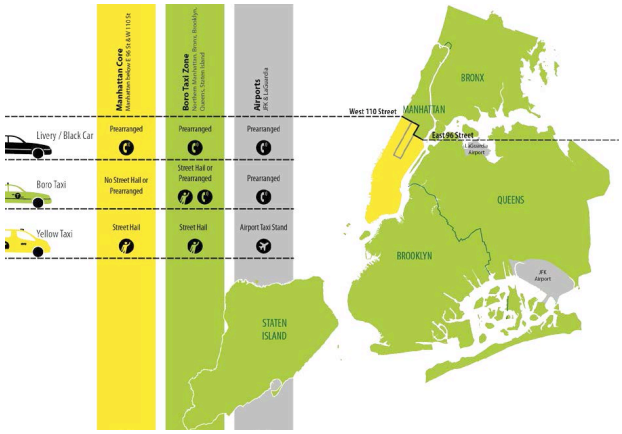

Figure 1. Livery/Black Car, Boro Taxi, and Yellow Taxi Service Areas [12]

In 2011, For-Hire Vehicle Dispatch Applications entered the New York City taxi market. The main example of such an application is Uber. Uber has developed a mobile application that allows users to electronically hail taxis connected to the Uber platform. In 2015, Uber can be considered a strong competitor of TLC's taxi services [1]. Bialik et.al. [1] found that Uber's services are not expanding the total fleet of taxis in New York, but that Uber's taxi's are replacing TLC's yellow and Green cabs. When looking at the five main outer neighbourhoods of New York, we found that Uber experienced an increase in pick ups between April-June 2014 and April-June 2015, whereas Yellow and Green cabs pick-ups remained at a fairly similar level.

The overall process of data analysis is shown in Fig. 2.

A. Data Acquisition and Dataset Description

Two different datasets were combined for the analysis. The first dataset consists of trip record data that was uploaded by the NYCTLC on their website. The second dataset was one consisting of trip record data of Uber pick-ups in NYC, obtained by the website fivethirtyeight.com (part of ESPN internet ventures) as a result of a Freedom of Information Law request on July 20, 2015. Both datasets

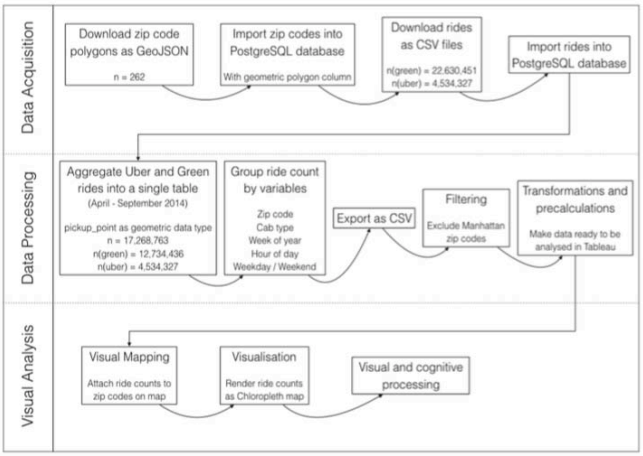

Figure 2. Overview of Data Analysis

were provided as CSV files in which each row represents a taxi ride and each column an attribute specific to that ride. We will refer to 'rows' as 'rides' in this paper.

1) NYCTLC Dataset: The NYCTLC states that the data was collected and provided to it by technology providers authorised under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the NYCTLC, and NYCTLC makes no representations as to the accuracy of these data. As will appear from the description of the variables of this dataset, the 'technology providers' responsible for the data collection are Creative Mobile Technologies and Verifone Inc. The NYCTLC does not elaborate on the purpose of collecting and uploading their trip data records. However, this is likely to be part of the NYC OpenData project, which is a collaboration between the Mayor's Office'a Analytics (MODA), the Department of Information Technology and Telecommunications (DOITT), and NYC Digital, to improve data-based decision making and promote public use of NYC data [13].

The TLC provides data for Green and Yellow taxis from 2009 till June 2015 with frequent additions of more recent months. For this paper, only the trip record data for Green taxis of the period April to September 2014 was used which adds up to approximately **8 million rides**. The main reason for selection of this particular time period is to allign to the second dataset (Uber) described next.

2) Uber Dataset: The second dataset consisted of trip record data of Uber pick-ups in NYC, obtained by the website fivethirtyeight.com (part of ESPN internet ventures) as result of a Freedom of Information Law request on July 20, 2015. The dataset amounts to trip data of **18.8 million** Uber rides, spread out over two periods (April-September 2014, and January-June 2015).

Two different formats were used to indicate the pickup location in the attributes. For the trip records of the period April-September 2014, latitudes and longitudes were used,

while for the period January-June 2015, locationID was used. This is why we have only analysed the period of April-September 2014. Originally, we intended to analyse the entire period of April 2014 to June 2015, which is why we downloaded and imported all data for Green cab trips for that period. However, due to the limitations of the Uber dataset, we limited queries on the Green cab dataset to April-September 2014.

B. Data Pre-Processing

Given the structured nature of the data, we decided that a relational database would be the best storage and processing device for this project. Several such database engines exist (e.g. MySQL, PostgreSQL, Microsoft SQL Server) and each has strengths and weaknesses compared to the others. A key requirement of the processing of our data was the ability to perform queries on geospatial data. We found that PostgreSQL has an extension called PostGIS, which allows the database to perform a plethora of geospatial queries. It also adds to ability to store geometric data types such as polygons and points. For this reason we have used the combination of PostgreSQL and PostGIS as the processing engine for the datasets.

We have set up the database using Amazon's Relational Database Service [14]. Amazon RDS provides support for PostgreSQL + PostGIS out of the box, which makes setting up a database in the cloud a matter of filling out a form. With the database set up we accessed it throughout the process using PostgreSQL's *psql* command line interface for unix- based systems, as well as a visual user interface via pgAdmin. The pre-processing of the data consists three distinct steps:

1) The importing of NYC zip codes into the database
2) The importing of the raw CSV ride data into the database
3) The aggregation of green rides and Uber rides into one big normalised table

The scripts used in data pre-processing step are located in GitHub Respository [1].

C. Data Analytics: Tools and Methods

The commericial tool Tableau (version 9.2) was used to perform the visual analytics of this paper. The 'calculated field' option proved difficult to use within Tableau and we had to perform a large number of supporting calculations through SQL in the database and with Excel before importing the aggregated results into Tableau. Whilst the in-built data layers in Tableau (which included demographics) were initially thought to be comprehensive; in practice they were not as extensive as first perceived and were unable to filter or edit as needed.

[1] GitHub Respository

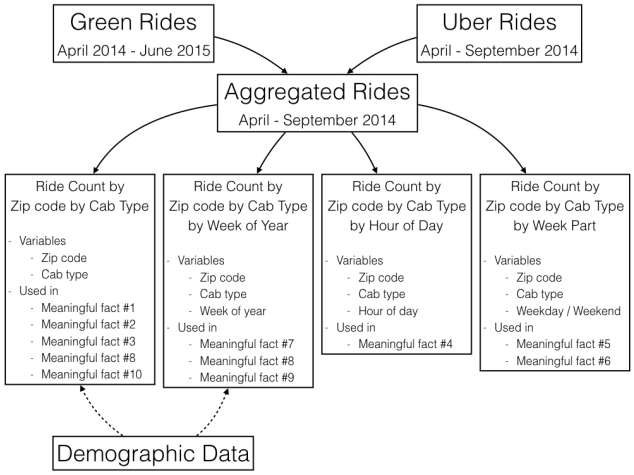

Figure 3. Data Analytics Modeling

A choropleth map displays data according to their geographical location [15] and can be used to create a clear overview that allows for comparison between different geographical areas. There are two types of choropleth maps: classified and unclassified. Classified choropleth maps divide geographical areas in a way that does not allow for a data relationship. It sets class intervals and assigns, for example, a certain colour to these intervals. Unclassified choropleth maps, however, divide geographical areas with a colour spectrum without setting intervals, which allows users to visually establish relationships between different parts of the map. Because we are conducting a comparative study between NYCTLC and Uber, we decided to create an unclassified choropleth map, to precisely show the difference in market share between Uber and NYCTLC per zip code.

In [16] compared classed choropleth maps to unclassed choropleth maps that is, maps that make groups data into coloured classes and maps that apply a spectrum. They found that the unclassed choropleth maps are more accurate with regard to quantization error. Moreover, they state that unclassed choropleth maps tend to display superior visual quality. Critics state, however, that the unclassed choropleth map could display too much information, and that the visual comparison of areas might be problematic when differences are relatively small [16]. When choosing a type of choropleth map, one should therefore carefully assess the amount of information one wants to display, and in which format trends are most easily identified.

IV. Findings

As mentioned in the methodology, we have the total numbers of rides for both companies sorted by zip codes. We organise our findings into the traid of meaningful facts, actionable insights and valuable outcomes.

A. Meaningful Facts

1) Meaningful Fact #1: Green cabs are just as popular as Uber on the weekend. The distribution of rides according to weekends versus weekdays comparison is very similar in regards to Green cabs and Uber as shown in Fig. 4. Also, the distribution is close to equal in both cases with approximately 40% of the rides occurring during the weekends. It should be noted, though, that the distribution is not really equal in terms of days as the weekends constitute 2.5 day and the weekdays 4.5 days. This means that even though the visualization deceives the interpreter to think of the distribution as a close to 50/50, one should realise that there are more rides taken place during one weekend day than one week day.

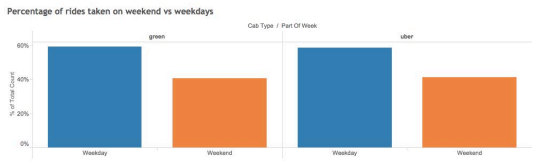

Figure 4. Distribution of rides over weekday vs weekend

2) Meaningful Fact #2: Weekdays versus weekend rides per hours. To make up for this difference in days with 4.5 weekday days and 2.5 weekend days, we took the total of number of rides occurring during weekdays and divided them by the total number of hours in 4.5 weekdays. Similarly we took the total number of rides occurring in weekends and divided that by the total number of hours in those 2.5 weekend days. The bar charts in Fig. 5 show the difference between the average rides per hour in weekends and weekdays for Green cabs and Uber respectively in total numbers, on the left, and in percentage increase, on the right. The difference is clearer in the right bar chart, as it shows a 3% higher increase of Uber rides per hour in the weekends i.e. compared on average hours, Uber increases during weekends by 48% while Green cabs increase by 45%.

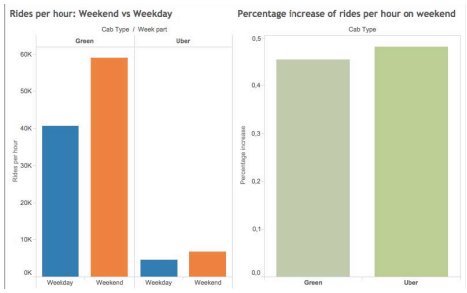

Figure 5. Comparison of hourly rides over weekday and weekend

3) Meaningful Fact #3: There is no clear correlation between the negative and/or positive growth of Uber and Green cabs. With the explosive growth of Uber, one could imagine that when looking at both negative and positive growth, Green cabs would see a negative growth where Uber is experiencing a positive growth, i.e. a 'takeover' growth by Uber. However, as seen in the growth visualization below this is not the case in all areas.

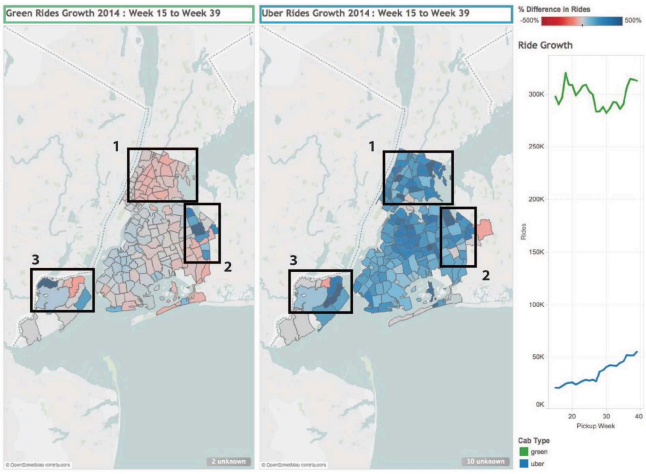

Figure 6. Growth in rides in week 15-39 for Uber and Green Cabs

First of all, the growth of the Green cab rides over the period of time investigated is very unstable and it peaked within the first weeks of the data set time. In contrast Uber's growth is more steady over time and is also higher in total. Uber grows from 18008 to 55089 rides a week yielding a 206% growth. Green grows from 297976 to 312877 rides a week yielding a 5% total growth. As for the relationship between the negative and positive growth, three areas have been highlighted to provide examples as shown in Fig. 6. Quadrant 1 shows a clear sign of Uber taken over market share from Green cabs. Looking at this quadrant of both Uber and Green cabs isolated would lead one to assume a clear correlation between Uber gaining more market share and Green cabs losing it. But quadrant 2 data reveals another pattern, which is that both Uber and Green cabs are experiencing growth in the same 7 zip code areas. This pattern would lead one to assume that both companies are experiencing overall growth. The last quadrant, number 3, shows a completely pattern-less up visualization of the growth comparison. In the bottom right corner zip code area both companies are experiencing growth. In the top right corner Uber is taking over market share from Green cabs, and in the top left corner Green cabs are taking over market share from Uber.

4) Meaningful Fact #4: : Interesting but irrelevant: Green cabs are winning in the poorest areas, which they could eventually lose to Uber. As previously analysed Green cabs are winning in the area just north of Manhattan. This area is one of the relatively poorest of the zip code areas included in our dataset as shown in Fig. 8.

5) Meaningful Fact #5: - Makes sense: Uber is gaining market share - Green Cabs are losing Looking at an overview

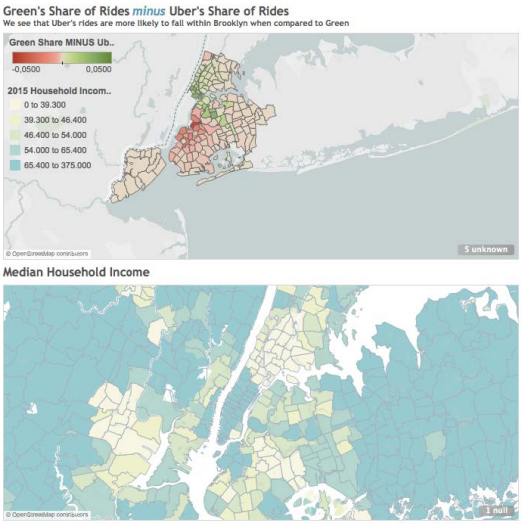

Figure 7. Growth of green cabs in the poorest areas

of the total percentage of rides week by week, it is clear that Uber is, generally speaking, gaining in on the Green Cabs' position as market leader as shown in Fig. 8.

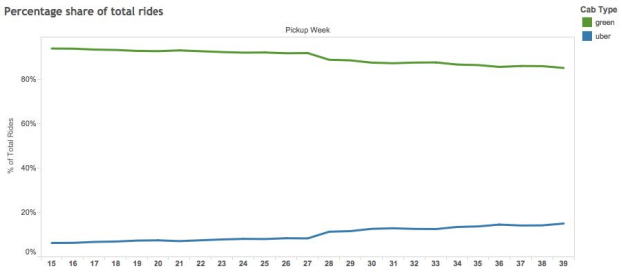

Figure 8. Percentage share of total rides

B. Actionable Insights

1) Continue Doing More: There is one main thing that the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) is already doing for the Green cabs and what they should continue doing more. As described earlier they have created a format, which is used to guide Verifone and CMT in their data collection and storing. For public use, this is stored on github in .csv format on a half-year basis or whenever someone is filing a FOIL request. This should be continued but they could consider boosting this effort and setting up data storage and data processing processes which would make the data available more frequently.

Both vendor companies are investing in technological setups, which can retrieve data, take multiple forms of payments and deliver entertainment for the passengers during their rides. One of the newest efforts to follow the technological development is made by a company called Arro through collaboration with CMT. This new app, Arro, allows users to order a cab by the push of a button, and pay automatically merely by leaving the cab when the ride is done (as described in the screenshot from Google Play store on the right). This is definitely a step in the right direction in terms of keeping up with the technological development, and they should continue to put effort into this area.

2) Continue Doing Less: As seen in the meaningful facts, the zip code areas furthest away from Manhattan are the only zip code areas where Uber actually has more rides than Green cabs when looking at total numbers. During the time span analysed, one can note that the service of Green cabs in these outer areas is diminishing. These far out areas are argued to be not profitable to serve for a full time Green cab driver, as the density and frequency of pick ups are relatively low compared to the rest of the areas included in the analysis. Because Uber's business model allows individuals to combine taxi driving with other income generating activities, Uber drivers can serve these outer areas more profitable, as they do not rely on driving full days. Therefore they should continue to serve these areas less.

3) Do Differently: Based on findings, it could be argued that LPEP should add an attribute to their dataset, which contains the zip code or other predefined areas of that specific ride. Practically, this is merely a matter of running the geo-location of the pick-up and the drop-off through a database containing all the chosen areas, which then contain all the geo locations belonging to the different areas, and add two area IDs to that line item.

4) Start Doing New: There are at least two new things which can be done benefitting the Green cabs. First, it would be beneficial to set up a real time data processing system, that would make it possible to analyse data week by week, day by day and even hour by hour. With the Arro app this should not require a big investment for the Green cabs using this type of vendor i.e. CMT, but as previously mentioned in our dataset CMT represented 78% of the data collected. It constitutes a challenge that all the Green cabs are not using the same vendor with regards to data collection and therefore also data analysis. A way to solve this issue would be to allow Verifone drivers to use the Arro app as well or to convince drivers to shift to CMT. Second, NYCTLC should analyse performance based on isolated and different areas. As seen in the meaningful facts section above, the difference between performances in different zip codes is significant and ignoring these differences would yield suboptimal returns.

5) Do Not Do Ever: In the 2015 analysis of the Five-Boro Taxi plan conducted by the NYCTLC, the company stated that over the course of the coming years it will make the price of Green cab medallions more expensive. Our analysis has shown that over 2014, the demand for cab rides in the outer boroughs of New York is rising, a trend that continued in 2015 [1]. Raising the price of Green medallions would most likely result in less medallions being purchased by potential cab drivers, resulting in lower growth of total

fleet size. Doing so would negatively impact the ability of the NYCTLC to keep up with growing demand, potentially damaging their competitive position.

C. Valuable Outcomes

1) Short-Term: As previously analysed, the entire market for taxi rides outside Manhattan is growing. However, this new market space is being lost to Uber, as they are growing faster than the Green cabs. Based on the actionable insights, it is assumed that the short-term value would be highest if the NYCTLC starts analysing the data already available at the moment as done in this paper. From these analyses, they will be able to focus their marketing efforts on where they are needed the most, and consequently attempt to re-capture the share of the customers that have been lost to Uber. Hence, it would be beneficial for NYCTLC to catch up with the growth rate, by understanding better where they are losing market share to Uber. It is advised that the NYCTLC immediately focuses on the following aspects: a) Areas (one or more zip codes) where Uber is growing faster than Green cabs b) Weekends as Uber's number of rides per hours during weekends increases by 3% more c) Doing nightlife campaigns to increase the number of rides taking place from 3 am till morning.

2) Mid-Term: On a mid-term period, NYCTLC could develop a dashboard, which shows already processed and analysed data in the ways previously described. This would imply showing rides per hours in different zip codes, their number of rides in weekdays and weekends compared, and their performance in different zip codes compare to the average performance of the market. As previously argued NYCTLC should furthermore continue to push the diffusion of the ARRO app, as doing so will strongly develop the company's e-hailing capabilities. Digitising the NYCTLC's services will improve the accessibility of Green cabs and empower the NYCTLC to better monitor consumer demand and execute strategic decisions accordingly. For example, the NYCTLC could analyse the Green cab density per region and assess whether this density is able to meet the current demand. If not, the NYCTLC could try to identify a region with a Green cab surplus and instruct drivers to relocate.

3) Long-Term: Our findings have shown that Uber's services is attracting an increasing amount of customers and that the number of Green cab rides is growing more slowly than Uber's. Thus, the NYCTLC has to revise their long-term strategy for Green cabs. As explained in the previous two sections, the NYCTLC is advised to further develop their data collection and analytics practices, allowing them to leverage this data in a more real-time manner. Additionally, they should continue to develop their mobile application in order to sustain a similar service level to Uber.

However, mimicking Uber's business model is not sufficient. In the long-term, the NYCTLC is advised to leverage those factors that give them an edge over Uber. These factors are twofold. First, the NYCTLC could utilise its affiliation to the NYC government. Our paper gave examples on how the combination of data on geo-location and socio-demographics can give insight on the boroughs and consumer demographics where Green cabs has to focus its targeted marketing campaigns. Assuming that the NYC government possesses even more detailed data on these categories and would grant access to these datasets, this could be a beneficial opportunity for the NYCTLC. Second, the NYCTLC should leverage its ability to better manager its drivers. Since Uber's drivers maintain a relatively high degree of freedom, the NYCTLC could create an edge over Uber by deploy a workforce that can be managed according to real-time monitored demand. This, together with the increased intelligence of the previous point, could potentially enable the NYCTLC achieve an inimitable business model, solidifying its position in the market.

V. Discussion

The data analysis performed in this research has shown that Green cabs have clearly maintained the market leader position in terms of total rides in the areas outside of Manhattan and the airport. Overall, a growth in rides has been identified of 5% for Green cabs. However, Uber has been gaining market share rapidly in a very large amount of the zip codes in NYC. Comparing this growth, Uber has clearly done better over the six months that have been analysed in this research.

1) For the New York Taxi Commission's under-represented zip codes, what socio-demographic attributes are most prevalent in the New York Taxi Commission's under indexed zip codes?: The zip codes where Green cabs have lost the most market share compared to Uber are mainly located to the north of Manhattan. The growth in market share for Uber in these areas has surfaced relatively late and slowly. Analysing the socio-demographic data of these areas indicates that these are inhabited by a population with a relatively low income. Thus, it is assumed that, in terms of innovation diffusion, this consumer segment can be identified as the "Late Majority" [17], in addition to being attracted to Uber's cheaper pricing.

2) Does customer preference change with respect to time of day (night/day or weekend/weekday)?: It can be concluded that customers slightly prefer Uber over Green cabs in the weekends. No customer preference has been identified during weekdays. Regarding night rides Uber's rides starts inclining at 3 am while Green cab rides declines from midnight to 5 am, indicating that Uber is slightly more popular as a taxi service during night time proportionally.

3) Limitations: The NYCTLC dataset included variables relating to fair amount, tips and drop-off locations, whilst the Uber set did not. This limitation resulted in a dramatic reduction of scope when it came to comparing the datasets as the only variables that were present in both datasets were

pickup coordinates and time. Further, differences existing the ride data for Uber between 2014 and 2015. For unexplained reasons, the 2015 pickup data (obtained from Github) was already classified based on Neighbourhood Tabulation Areas (NTAs), and was already aggregated from its original geographic coordinates form. This meant that we were unable to match the 2015 data to a zip code using PostGIS. To account for this; the idea of using PostGIS to match the 2014 coordinatesto NTA was explored. However, Tableau supports zip codes by default as opposed to NTAs. Whilst it was possible to draw our own boundaries in Tableau using polygon coordinates to support NTAs; due to its complexity, it was decided instead to exclude the 2015 Uber data and compare the 2014 datasets. Both sets were additionally missing information regarding driver and car medallion number. A security breach in 2014 led to the NYCTLC not releasing this information in an anonymous form as it could technically be traced back to the driver and their home address and personal income [18]. This meant that no analysis could be made on an individual driver level between the two companies and the efficiency of their fleets.

VI. Conclusion

In this paper, we conducted a geo-spatial analysis of Green cab and Uber rides in the outer neighbourhoods of New York City, in order to assess the competitive position of the NYCTLC. We have found that demand for Green cab's is still growing, but that the number of Uber rides in the same area is growing more rapidly. Additionally, the analysis showed that especially in relatively poor neighbourhoods, Green cabs are performing better than Uber's. When looking at differences between weekdays and weekends, we found no differences between Green cab's and Uber. Our research recommends the NYCTLC to create a dashboard that analyses and visualises data real-time, because we believe that this will raise their competitiveness in comparison with Uber.

Acknowledgments

We thank the members of the Computational Social Science Laboratory (http://cssl.cbs.dk) for their valuable feedback. The authors were partially supported by the project *Big Social Data Analytics: Branding Algorithms, Predictive Models, and Dashboards* funded by *Industriens Fond* (The Danish Industry Foundation). Any opinions, findings, interpretations, conclusions or recommendations expressed in this paper are those of its authors and do not represent the views of the Industriens Fond (The Danish Industry Foundation).

References

[1] C. Bialik, A. Flowers, R. Fischer-baum, and D. Mehta, "Uber is serving new yorks outer boroughs more than taxis are. but most of its rides, like those of taxis, still start in manhattan." http://fivethirtyeight.com/features/uber-is-serving-new-yorks-outer-boroughs-more-than-taxis-are/, 08 2015. 1, 3, 6

[2] S. Silverstein, "These animated charts tell you everything about uber prices in 21 cities," http://www.businessinsider.com/uber-vs-taxi-pricing-by-city-2014-10?IR=T, 10 2014. 1

[3] A. Tangel, "Green taxis gaining ground in new york city," http://www.wsj.com/articles/green-taxis-gaining-ground-in-new-york-city-1403145481, 06 2014. 1

[4] B. P. Loo, B. S. Leung, S. Wong, and H. Yang, "Taxi license premiums in hong kong: Can their fluctuations be explained by taxi as a mode of public transport?" *International Journal of Sustainable Transportation*, vol. 1, no. 4, pp. 249–266, 2007. 2

[5] J. M. Cooper and W. Faber, "Taxi demand modeling to ensure appropriate taxi supply. do both open access and model based restrictions fail?" *Traffic and Transportation Studies 2010*, p. 60, 2010. 2

[6] T. J. Kim, "Metadata for geo-spatial data sharing: A comparative analysis," *The Annals of Regional Science*, vol. 33, no. 2, pp. 171–181, 1999. 2

[7] M. A. Yazici, C. Kamga, and A. Singhal, "A big data driven model for taxi drivers' airport pick-up decisions in new york city," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 37–44. 2

[8] T. H. Savage and H. T. Vo, "Yellow cabs as red corpuscles," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 22–28. 2

[9] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2149–2158, 2013. 2

[10] L. Zhu, D. M. Gorman, and S. Horel, "Alcohol outlet density and violence: a geospatial analysis," *Alcohol and alcoholism*, vol. 39, no. 4, pp. 369–375, 2004. 2

[11] S. Funk and P. Piot, "Mapping ebola in wild animals for better disease control," *Elife*, vol. 3, p. e04565, 2014. 2

[12] NYC Taxi and Limousine Commission, "2015 hail market analysis," http://www.nyc.gov/html/tlc/downloads/pdf/hail_market_analysis_2015.pdf, 2015. 2, 3

[13] ——, "Tlc trip record data," http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml, 2016. 3

[14] Amazon Web Services, "Amazon relational database service (amazon rds)," https://aws.amazon.com/rds/, 2016. 4

[15] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo." *Commun. Acm*, vol. 53, no. 6, pp. 59–67, 2010. 4

[16] N. Gale and W. C. Halperin, "A case for better graphics: The unclassed choropleth map," *The American Statistician*, vol. 36, no. 4, pp. 330–336, 1982. 4

[17] V. Mahajan, E. Muller, and R. K. Srivastava, "Determination of adopter categories by using innovation diffusion models," *Journal of Marketing Research*, pp. 37–50, 1990. 7

[18] A. Hern, "New york taxi details can be extracted from anonymised data, researchers say," http://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn, 06 2014. 8