Passenger Saver based on Uber Surge Multiplier Progress Report

PENG ZHENG*, Georgia Institute of Technology
JINLI YAN*, Georgia Institute of Technology
XIPENG CHAI*, Georgia Institute of Technology
HAOMIN LIN*, Georgia Institute of Technology
YANJUN DING*, Georgia Institute of Technology
YUNTIAN ZHANG*, Georgia Institute of Technology

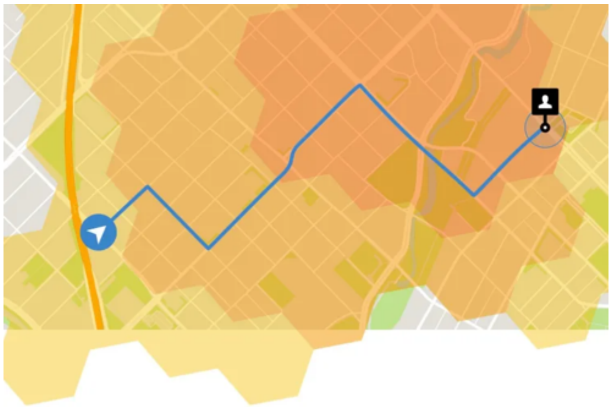


Fig. 1. Schematic Diagram of Distribution of Uber Surge Multiplier.(https://www.uber.com/us/en/drive/partner-app/how-surge-works/)

1 INTRODUCTION

Uber is a technology company that is famous for managing a ride-sharing platform. One of the most important characteristics of Uber is the flexibility of supply[4].Today, what we will focus on is Uber's dynamic pricing called "surge multiplier".Moreover,Uber will charge passengers based on two factors: time and distance. During times of high demand, Uber will use this to

*All team members have contributed similar amount of effort.

Authors' addresses: Peng Zheng, peng_zheng@gatech.edu, Georgia Institute of Technology; Jinli Yan, jinli_yan@gatech.edu, Georgia Institute of Technology; Xipeng Chai, xchai30@gatech.edu, Georgia Institute of Technology; Haomin Lin, humaslin97@gatech.edu, Georgia Institute of Technology; Yanjun Ding, yding366@gatech.edu, Georgia Institute of Technology; Yuntian Zhang, yzhang3469@gatech.edu, Georgia Institute of Technology.

increase prices[3]. The surge multiplier will vary from region and time. **Fig. 1** quoted from the Uber website illustrates how this works. The darker the color is, the higher the price is. To be honest, this method may increase the efficiency of ride-sharing. With the help of surge, it reduces some customers out of the market, for example, many young students may give up calling a Uber when the price is relatively high. Besides, a higher price is able to attract more drivers to drive during the high-demand time or come to this region[3]. The surge multiplier will be updated every 5 minutes. It may sound reasonable, but this information is only visible for drivers. Passengers can be charged a higher price unconscious. Our purpose in this project is letting this information be accessible to passengers as well. For example,in **Fig. 1**, one person is on the edge of a high surge multiplier region. We will use data from the past to give him a guide to move to the nearest region with the lowest surge multiplier. To use more valid data, we decide to use the Uber data from San Francisco due to the largest number of Uber[5].

2 LITERATURE REVIEW

More and More researchers are interested in analyzing the spatial and temporal disparity of Uber demands[6, 7, 18], becuase Uber a representative of "sharing economy". Wang, et. al[18] studied the Uber accessibility by analyzing average wait time in different areas of Atlanta (**Fig. 2(a)**)[6, 7]. They tried to fit the data with models to figure out what factors (like wealth, race, public transportation etc.) influence the spatial disparity of Uber. Similar researches went on in New York City , which focused on analyzing number of pick-up data (**Fig. 2(b)**). Also, comparison between Uber and traditional taxi are made (**Fig. 3**), predicting that demands for Uber may exceeds that of taxi in the following years due to its convenience and relatively low price[6, 15]. Other reviewed papers are shown in the reference lists[9, 13].

Now that more and more users join the Uber ride, it's important for riders to know how the price of trip is calculated. As economist pointed out, truthfulness and information disclosure between riders and riding-sharing companies is vital to further development of ride sharing. Some researchers begin to pay attention on the surge mechanism of Uber[1, 3–5, 10, 14]. Chen, et. al tried to open a black box of surging price algorithms by applying a model to analyze Uber trip data based on cars supply, demand, wait times, locations, etc [3]. As they figured out, surge multiplier changes with areas and time (**Fig. 4**). It is refreshed every 5 minutes. Other researchers criticized the unfairness of surge, suggesting that surge price should be disclosed to public [3, 4]. Therefore, we want to make surge exposed to users while help them to choose a better place and occasion to book a Uber ride. In order to realize these functions, papers on map constructions are reviewed[2, 8, 11, 12, 16, 19] . Several studies are based on Google map API which is used to implement data visualization on maps [8, 12]. Brakatsoulas, et. al[2] studied algorithms that exploit the trajectory nature of the tracking data which is helpful to us in designing a path leading to an area with lower surge multiplier.

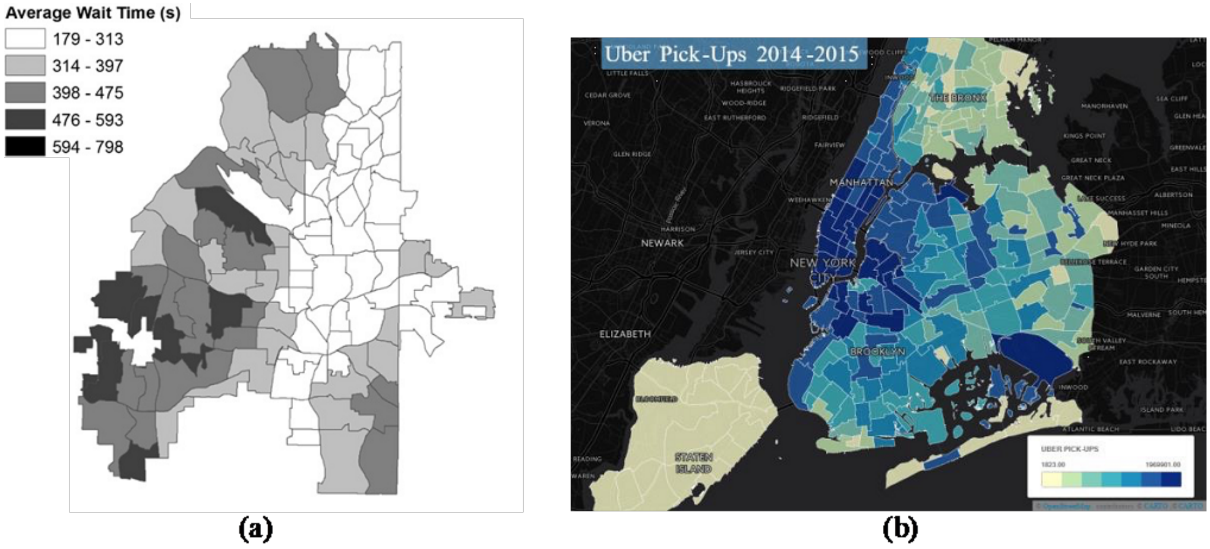


Fig. 2. (a) Estimated wait time of Uber in Atlanta in 2016 (reprinted from Wang, et. al, 2018) (b) Uber demands in New York city in 2014-2015 (reprinted from Correa, et. al, 2017)

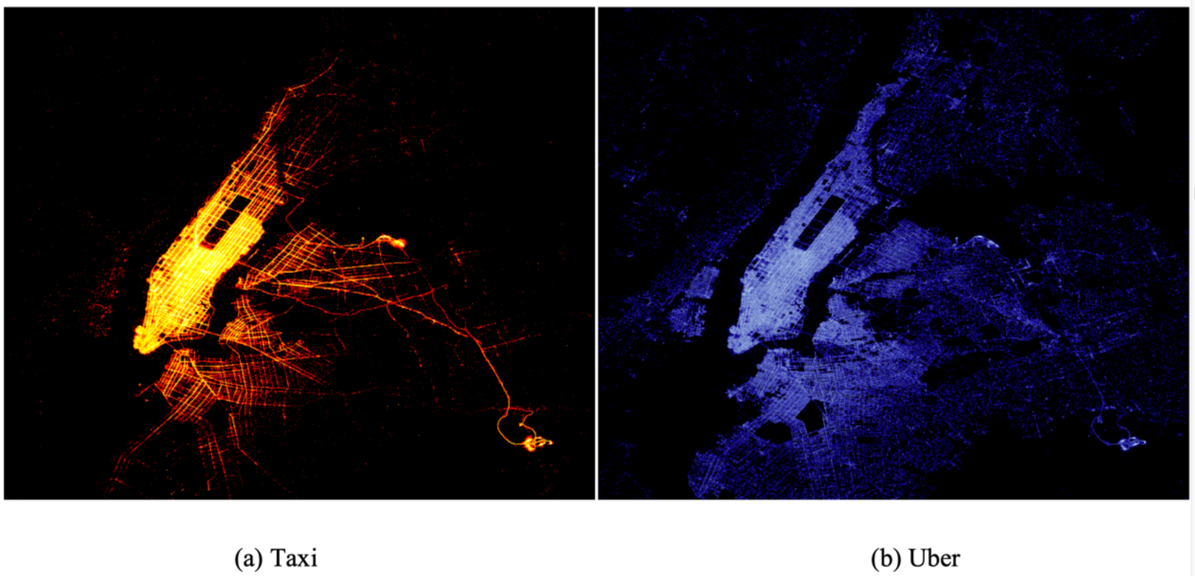
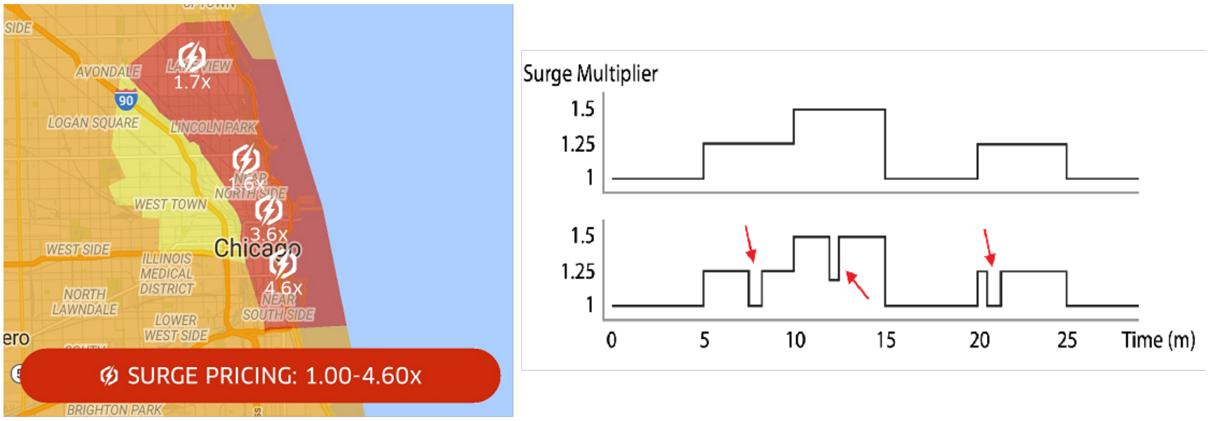


Fig. 3. Heatmaps of taxi and Uber pick-ups in New York City (reprinted from Correa, et. al, 2017)



(a)

(b)

Fig. 4. (a) Surge multiplier changes with different areas (b) Surge multiplier changes with time (reprinted from Chen, et. al, 2015)

3 PROJECT CURRENT PROGRESS

In our previous proposal, we planned to get data from Uber with the help of Uber API. At our first meeting, we tried to ***use API with Python*** to scratch data from uber. Because Uber closed access to API for the public, we must ***change our strategy***.

In order to overcome this obstacle, we kept searching and finally got ***a dataset of Uber and Lyft*** in Boston from Kaggle. We are looking for the relationship between surge and other factors such as weather and time and we find it hard to guide users to a place with smaller surge multiplier since dataset is set in places of Boston. In the second meeting, we started to deal with the dataset. The first step is clean the data, which means removing the data with abnormal value. This dataset includes data of multiple products from uber and Lyft. We used ***Tableau*** to virtualize the data and find that the different prices among these products will affect our calculation of surge. Then we calculated the average price of each product such as Uber x and Uber pool. For each single price, divide the average price according to its product name, we get the value of surge. We tried to use ***Map Reduce*** but found it was too time-consuming. After discussion, we used ***Apache Spark*** to process our data. After a whole alternation of discussion, we came up the algorithm and finished cleaning and calculating the data. For more details, Before we start establishing models for the prediction, we decide to use ***Tableau*** draw some pictures to show the relationships between variables and the "surge" parameter, so that we could have a direct understanding of how the variables are affecting "surge", as well as make a rough estimate for the final outcome.

From **Fig. 5(a)** surges in different weekdays are different. This requires us not to integrate all weekdays into one category when establishing the model, or it could lead to inaccuracy. From **Fig. 5(b)** we find that different product types have approximately the same distribution

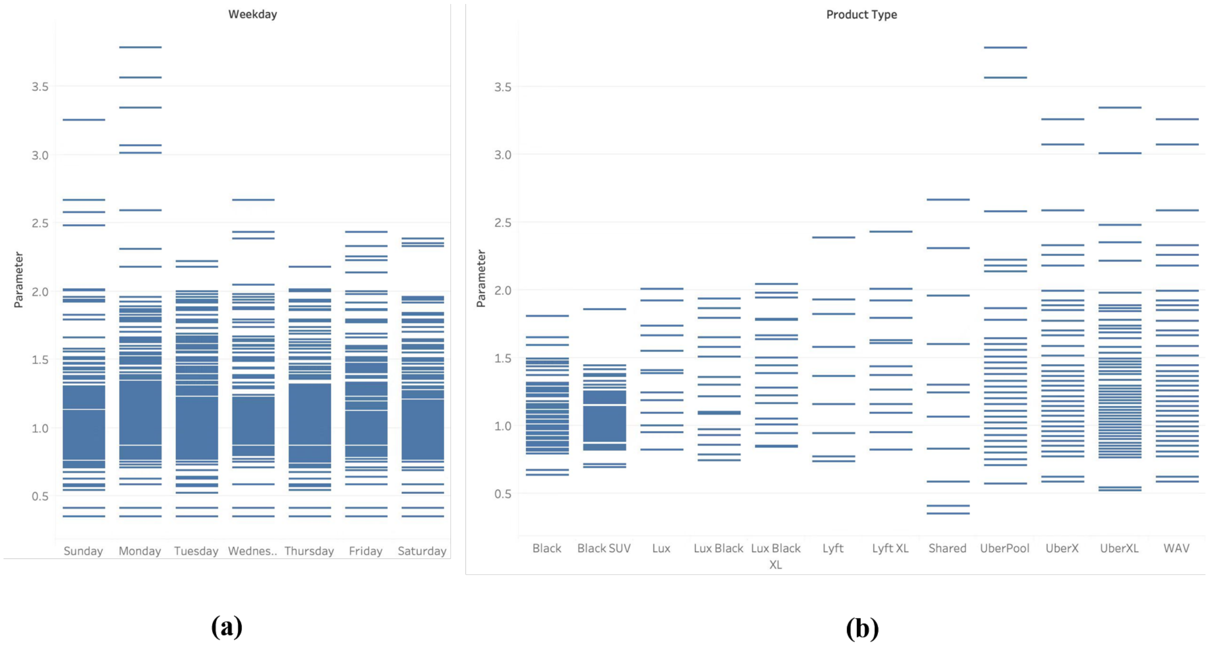


Fig. 5. (a)Surge multipliers in different weekdays; (b)Surge multipliers for different product types.

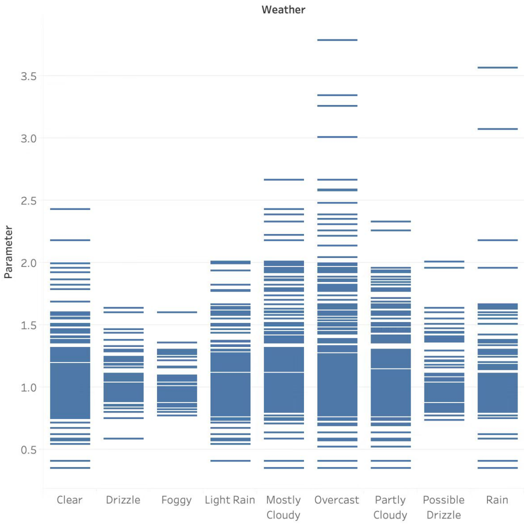
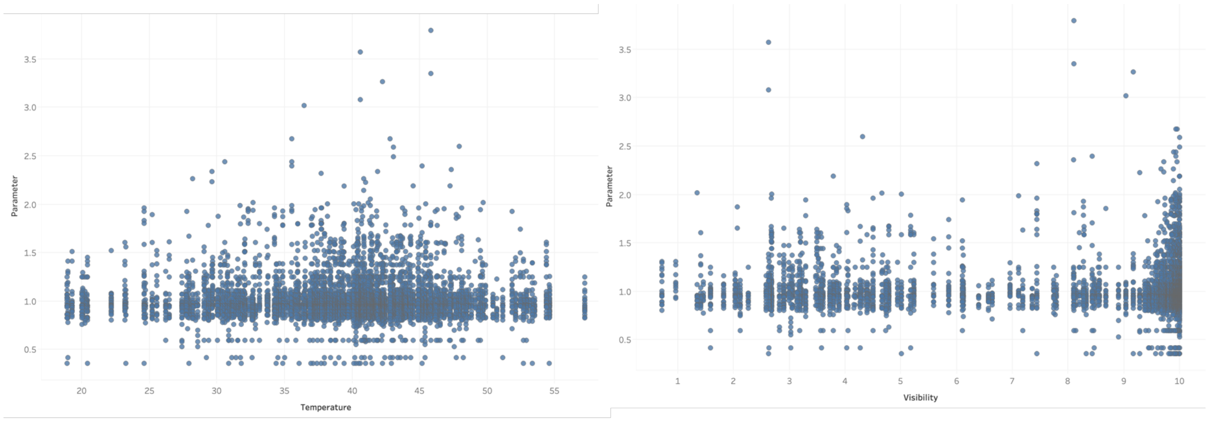


Fig. 6. Surge multipliers in different weather.

of surge parameter despite Uber Pool. We will discard this type of product in future works and use normalization to flatten different product types into one. **Fig. 6** , **Fig. 7(a)**, **Fig. 7(b)** and



(a)

(b)

Fig. 7. (a)Surge multipliers in different temperature; (b)Surge multipliers in different visibility.

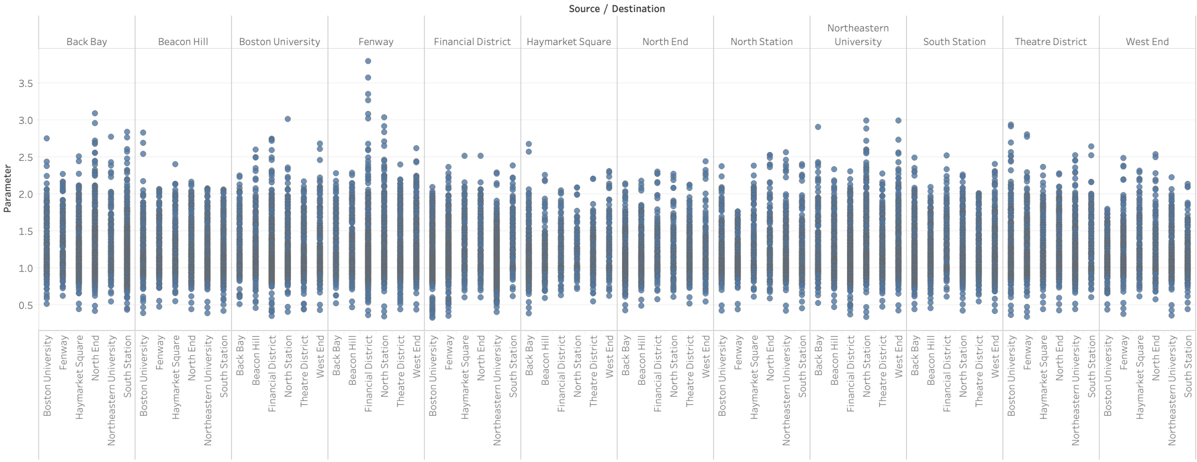


Fig. 8. Surge multipliers for different sources and destinations.

Fig. 8 exhibit respectively how weather, temperature, visibility,departures and destinations affect surge. We find that the distribution of surge parameter changes with the change of weather and visibility, while it is mostly the same for various temperature. Therefore, we will only use weather and visibility as factors of the prediction model.

4 REVISED PROJECT METHODS AND PLANS

Previously in the Proposal, our plan of activities is that we firstly get the data and build our database, then use certain algorithms to analyze the data, and finally put the data into visualization. Importantly, we try to analyze surge multiplier of the whole area (color certain areas[17]) near users' locations so that we can give a suggestion for users. For plans, finishing data analysis will be our "midterm" and the final exam is running our system to get plans to avoid high surge multiplier. ***Previously***, our plan can be divided into four parts: data collection (Peng and Jinli), program design (Yuntian and Yanjun), interface design (Haomin) and program refine (All members)

Currently, however, we realize that data clean analysis is much more time-consuming that we thought and fetching dataset is not as free as we thought. As a result, we extend our time for analysis and focus more on Machine Learning Analysis and Prediction of Surge based on time, departures, destinations, weathers, showing users the border of the surging area and giving suggestion about where to go for users will not be considered due to the limitation of datasets.***Currently***, plan of activities has been rescheduled based on the current progress: Scikit-Learn Analysis(Jinli and Peng), Azure Studio Analysis (Yuntian and Yanjun), Prediction Interface (Haomin and Xipeng).

Concretely speaking, after cleaning the data and calculate some necessary value for our final products like the "surge parameter", we plan to further process our data with ***Machine Learning*** methods. As in our final product, we wish to give users a prediction of Uber or Lyft prices and surge multiplier according to the time, weather, location and other potential factors input by users. Therefore, we need to make suitable analysis on our current datasets.

One possible way is using Machine Learning methods to fit the data. We have two possible plans to implement our ideas. Firstly, we can build a random forest. Current datasets are used to train the decision trees in the forest. We have multiple attributes like locations, weather, temperature, date and time. The values to be predicted is the surge multiplier which is a discrete value since it can be limited to an integer in the range of 0 to 3 so that it's a classification problem. However, when it comes to the prediction of continuous price, it becomes a regression problem. With the help of python package ***Scikit-Learn*** we can train classifiers with our datasets. Then we can make some pseudo data to test the classifiers and let them make prediction for us. With the prediction results, we could further visualize it or use them to build user interfaces. An alternative way to make predictions of prices based on our data is to use the ***Microsoft Azure Studio***. With packaged machine learning algorithm, it's quite easy for us to do regression calculations for our datasets. Once the model is established, we could make predictions as we give it a set of attributes. With these prediction results, it's possible to build a user interface conforming to our goals.

After the data wrangling and analyzing process, the project does some prediction based on the result of the analysis. During the data analysis process, we analyze the relationship between different conditions (time, region, weather, etc.) Our interface allows the user to input their ***destinations and departures***. Then, the project can show the price by either taking Uber or Lyft. This useful when two departure areas are not far but normally has different surge. Users of the project can move to a departure point by checking the predicted price our project provides. A line chat will also be illustrated to show the price fluctuation over different time.

This prediction is useful is the user of this program aren't in a rush and want to save some money by ***adjusting their travel time***. In addition, if the data analysis before shows a big difference of price in a same route under ***different whether situation***, the program will be refined further based on whether information and give more precise price prediction. A bar chart (or other illustration way) can also be made to show the weather-price relationship in order to give the users a more direct and sharp illustration.

ACKNOWLEDGMENTS

All team members have contributed similar amount of effort.

REFERENCES

[1] Niels Agatz, Alan Erera, Martin Savelsbergh, and Xing Wang. 2012. Optimization for dynamic ride-sharing: A review. *European Journal of Operational Research* 223, 2 (2012), 295 – 303. https://doi.org/10.1016/j.ejor.2012.05.028

[2] Sotiris Brakatsoulas, Dieter Pfoser, Randall Salas, and Carola Wenk. 2005. On Map-matching Vehicle Tracking Data. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB '05)*. VLDB Endowment, 853–864. http://dl.acm.org/citation.cfm?id=1083592.1083691

[3] Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking Beneath the Hood of Uber. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. ACM, New York, NY, USA, 495–508. https://doi.org/10.1145/2815675.2815681

[4] M. Chen. 2016. Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform. 455–455. https://doi.org/10.1145/2940716.2940798

[5] Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. 2016. *Using Big Data to Estimate Consumer Surplus: The Case of Uber.* Working Paper 22627. National Bureau of Economic Research. https://doi.org/10.3386/w22627

[6] Diego Correa Barahona, Kun Xie, and Kaan Ozbay. 2017. Exploring the Taxi and Uber Demand in New York City: An Empirical Analysis and Spatial Modeling.

[7] Sabiheh Sadat Faghih, Abolfazl Safikhani, Bahman Moghimi, and Camille Kamga. 2017. Predicting Short-Term Uber Demand Using Spatio-Temporal Modeling: A New York City Case Study. arXiv:stat.AP/1712.02001

[8] C. Fu, Y. Wang, Y. Xu, and Q. Li. 2010. The logistics network system based on the Google Maps API. In *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, Vol. 3. 1486–1489. https://doi.org/10.1109/ICLSIM.2010.5461215

[9] Masabumi Furuhata, Maged Dessouky, Fernando Ordóñez, Marc-Etienne Brunet, Xiaoqing Wang, and Sven Koenig. 2013. Ridesharing: The state-of-the-art and future directions. *Transportation Research Part B: Methodological* 57 (2013), 28 – 46. https://doi.org/10.1016/j.trb.2013.08.012

[10] Nikhil Garg and Hamid Nazerzadeh. 2019. Driver Surge Pricing. arXiv:cs.GT/1905.07544

[11] Josh Greenfeld. 2002. Matching GPS Observations to Locations on a Digital Map. (01 2002).

[12] Maged Kamel Boulos. 2005. Web GIS in practice III: creating a simple interactive map of England's Strategic Health Authorities using Google Maps API, Google Earth KML, and MSN Virtual Earth Map Control. *International journal of health geographics* 4 (10 2005), 22. https://doi.org/10.1186/1476-072X-4-22

[13] Farshad Kooti, Mihajlo Grbovic, Luca Maria Aiello, Nemanja Djuric, Vladan Radosavljevic, and Kristina Lerman. 2017. Analyzing Uber's Ride-sharing Economy. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 574–582. https://doi.org/10.1145/3041021.3054194

[14] Kyungmin (Brad) Lee, Marcus A. Bellamy, Nitin Joglekar, Christo Wilson, and Shan Jiang. 2019. Surge Pricing on A Service Platform under Spatial Spillovers: Evidence from Uber. *Academy of Management Proceedings* 2019, 1 (2019), 16279. https://doi.org/10.5465/AMBPP.2019.16279abstract arXiv:https://doi.org/10.5465/AMBPP.2019.16279abstract

[15] L. K. Poulsen, D. Dekkers, N. Wagenaar, W. Snijders, B. Lewinsky, R. R. Mukkamala, and R. Vatrapu. 2016. Green Cabs vs. Uber in New York City. In *2016 IEEE International Congress on Big Data (BigData Congress)*. 222–229. https://doi.org/10.1109/BigDataCongress.2016.35

[16] Nadine Schuessler and Kay Axhausen. 2008. Processing GPS Raw Data Without Additional Information. (01 2008).

[17] S. Seipel and N. J. Lim. 2017. Color map design for visualization in flood risk assessment. *International Journal of Geographical Information Science* 31, 11 (2017), 2286–2309. https://doi.org/10.1080/13658816.2017.1349318 arXiv:https://doi.org/10.1080/13658816.2017.1349318

[18] Mingshu Wang and Lan Mu. 2018. Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. *Computers, Environment and Urban Systems* 67 (2018), 169 – 175. https://doi.org/10.1016/j.compenvurbsys.2017.09.003

[19] Xiaolu Zhou, Mingshu Wang, and Dongying Li. 2017. From stay to play – A travel planning tool based on crowdsourcing user-generated contents. *Applied Geography* 78 (2017), 1 – 11. https://doi.org/10.1016/j.apgeog.2016.10.002