Speed Anomalies and Safe Departure Times from Uber Movement Data

Nabil Al Nahin Ch
nc1145@wildcats.unh.edu
University of New Hampshire
Durham, New Hampshire

John Krumm
jckrumm@microsoft.com
Microsoft Research
Redmond, Washington

Andrew Kun
andrew.kun@unh.edu
University of New Hampshire
Durham, New Hampshire

ABSTRACT

Analyzing traffic data to find useful statistics for specific routes can help city planners, ride-sharing service providers, and travelers. Uber Movement datasets are especially useful in this type of study since they provide hourly speed data for individual road segments. In this study, we analyzed traffic patterns of New York City using the Uber Movement datasets for 2018 and 2019. We built a model to predict traffic speeds, which let us find anomalies on individual road segments. We found that speed anomalies mostly occur during rush hours and that longer than usual travel times are more frequent compared to shorter than usual travel times. We also found that speed patterns on some routes do not follow the conventional commute speed pattern. We used the speed statistics to compute safe departure times such that a traveler would likely reach their destination by a certain time with some prespecified probability.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**.

KEYWORDS

Uber movement, traffic, anomaly detection, travel time

ACM Reference Format:
Nabil Al Nahin Ch, John Krumm, and Andrew Kun. 2020. Speed Anomalies and Safe Departure Times from Uber Movement Data. In *UrbComp 2020: 9th SIGKDD International Workshop on Urban Computing, August 24, 2020, San Diego, CA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

A clear understanding of a city's traffic patterns is essential for urban planning, and it also allows optimal allocation of resources for ride-sharing service providers [11]. One often underappreciated aspect of these types of studies is that understanding the traffic condition of a route can also help the travelers plan their trip efficiently. Analyzing traffic data can be useful for accurately forecasting demand [6] and detecting anomalies caused by weather conditions [5] or any events. Most of the studies in this area rely on data from traffic surveys [1, 2] which may not be precise as they depend on participants' memories. More importantly, traffic data from surveys can be useful to analyze the overall traffic pattern of an area, but they can not be used to examine what happens on individual roads.

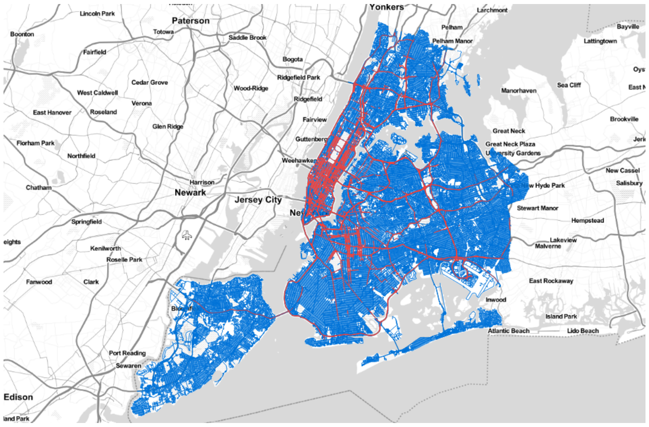


Figure 1: Roads included in Uber Movement New York 2018 dataset (blue) and roads we selected for analysis (red).

Crowd-sourced traffic data like the Uber Movement datasets [3] provides an opportunity to examine traffic patterns for a whole city as well as on individual roads. Uber Movement datasets also provide actual hourly statistics of speeds on roads instead of average speed over a longer period of time. We used this fine-grained speed data for two related tasks. The first is predicting speeds on roads as a function of time. We show how a periodic regression model accurately predicts speed over a typical week. Besides providing guidance on when travel speeds will be low, this predictive model lets us detect speed anomalies by finding when actual speeds deviate significantly from the predicted speeds. We can then give statistics on the timing and severity of anomalous traffic slowdowns.

The second task is using speed statistics to decide safe departure times. Based on a probability distribution of travel time, this analysis lets a traveler decide when to depart in order to reach the destination within a particular amount of time with a particular probability.

We next describe the Uber Movement data we used and our preprocessing. Then we describe the two tasks.

2 DATA ANALYSIS

We used the Uber Movement data for our analysis. This section describes the data and our preprocessing.

UrbComp 2020, August 24, 2020, San Diego, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2.1 Uber Movement Data

Uber recently released the Uber Movement datasets which provide anonymized traffic data for different cities around the world [3]. The expectation is that these datasets will be useful to researchers for analyzing traffic conditions of cities and helping city planners with urban planning. Datasets of different cities provide different statistics of the Uber rides. For our study, we selected the datasets of New York City which provide average and standard deviation of speeds of Uber rides on different road segments for each hour of the year. Road segments are defined as the part of the road between two OpenStreetMap (OSM) nodes. The Uber data comes with ID pointers to the OSM data, from which we can get road features, including their endpoints and polylines. Studies show that the Uber Movement data can be used as a proxy for car-based traffic data [7] and can also be compared with other crowd-sourced traffic datasets [9]. Thus the average speed of Uber rides can be used as a proxy for an average speed of car-based traffic for that particular road segment at that hour. This allows us to examine the pattern of car-based traffic of individual roads by analyzing the Uber Movement dataset.

2.2 Data Pre-processing

Uber movement data for New York City is available for 2018 and 2019. We used data from 2018 for training a prediction model (Section 3.2) and analyzing traffic patterns. Data from 2019 were used for testing the prediction model. There were 116,126 road segments in the 2018 dataset. To avoid ambiguity, we removed road segments that had multiple speed data for the same hours for 2018 or 2019 datasets. For our study, we selected the road segments on which there were Uber rides during at least 90% of the hours of 2018. Thus we used the Uber Movement data from 2018 and 2019 for 10,628 distinct road segments for analyzing traffic patterns, and for training and testing prediction models. Figure 1 shows the roads included in the Uber Movement dataset for New York City in 2018 (blue) and the roads we selected for our study (red).

3 SPEED: PATTERN, PREDICTION AND ANOMALIES

This section describes the temporal patterns we found in speed and our predictions of those speeds. By looking at deviations from the predictions, we were able to find anomalies, which can be important for city planners and drivers.

3.1 Analyzing Speed Patterns

Several studies found that traffic flow and travel time depends on both the day of the week and time of the day [4, 7, 8, 10]. Our initial analysis showed that the speed pattern is different on different road segments, but we wanted to see the overall pattern for the whole city. We examined how the speed changes in New York City depending on the day and time by combining average speed data of all 10,628 selected road segments. For each road segment, the hourly speed was expressed as a fraction of the whole-year mean speed of that road segment. Then the data from the selected road segments was combined to find the hourly mean speed. From that hourly mean speed of all road segments of 2018 and 2019, we found the weekly speed pattern, shown in Figure 2. We can see that on weekdays, the speed starts to go down in the morning and reaches a minimum speed between 8:00 am and 9:00 am, presumably commuters going to work. The lowest speed is observed at the hours between 4:00 pm and 6:00 pm, likely commuters returning home. There is no morning rush hour effect on speed during weekends, but the speed is lowest between 4:00 pm to 6:00 pm as on weekdays.

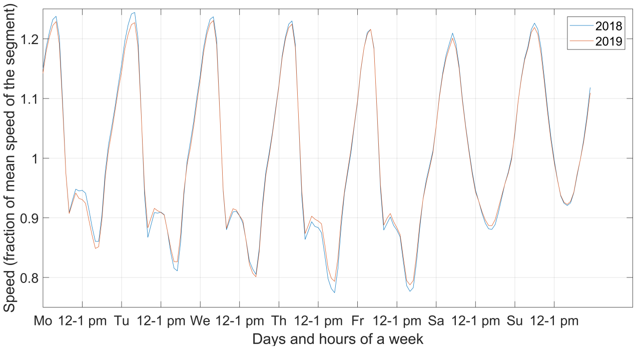


Figure 2: Weekly speed pattern of New York City.

We also found that even though most road segments follow the commute speed pattern shown in Figure 2, some routes follow different speed patterns. We selected two example routes to demonstrate this difference. The first route was in Manhattan, from the Empire State Building to Times Square. The second route was in Queens, from Highland Park to Cunningham Park. The first route was about 1 mile long and consisted of 17 road segments. The second route was about 7.3 miles long and consisted of 45 route segments. For each route, we found the mean speed of each road segment for each hour as a fraction of the 90th percentile speed of that segment and calculated the average to obtain the weekly speed pattern of the route. Using fractions of the 90th percentile means we could sensibly average together these fractions for an overall profile of the whole route. Figure 3 shows the weekly speed pattern for the Manhattan (top) and Queens (bottom) route. The route in Manhattan does not follow the overall speed pattern of New York City shown in Figure 2. There is not a trough during morning and afternoon rush hours, and the difference between the weekdays and weekends speed patterns is also not very apparent. The route in Queens, on the other hand, closely resembles the commute speed pattern with a trough during rush hours and different speed patterns for weekends. The reason behind the first route not following the usual commute speed pattern could be because the route is connecting two major tourist attractions. So there is more traffic during the middle of the day instead of during morning peak hours. A large number of tourists in the area means more pedestrians crossing streets, which can also contribute to the slower speed.

3.2 Prediction Model

Predicting speed and looking for deviations is one way to find anomalies. Here we describe our prediction model that we used on each road segment. From our initial analysis, we have seen that the speed of a road segment depends on the speed of previous hours and the speed also changes periodically. So we considered an auto-regression (AR) model and a harmonic regression (HR)

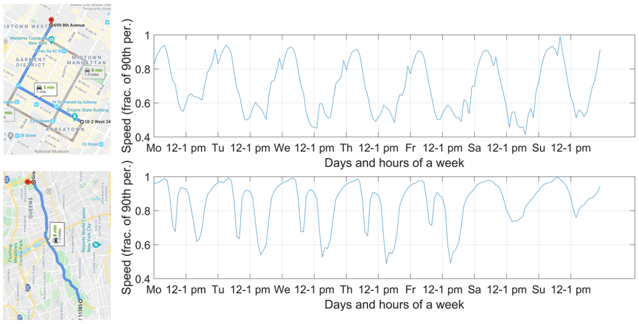


Figure 3: Weekly speed pattern of example routes in Manhattan (top) and Queens (bottom).

model for speed prediction. However, AR models are not suitable for detecting anomalies lasting for a few hours, since the predicted speed depends on the speed at previous hours. Since the purpose of the prediction model was to compare predicted and actual speed to find anomalies, we used HR models for predicting speed.

For each road segment, this is the form of the prediction model.

$$\hat{s}(t) = a_0 + \sum_{n \in C} a_n \cos(\frac{2\pi nt}{T}) + b_n \sin(\frac{2\pi nt}{T})$$

where T is the signal length and a_0 is the mean of the time series. The a_n and b_n come from minimizing the squared error between the measured speeds $s(t)$ and predicted speeds $\hat{s}(t)$ in the training data from 2018. There is a separate set of a_n and b_n for each road segment. The set of harmonic components C comes from the 40 most significant frequency components for each road segment.

We tested the HR model for each road segment using 2018 and 2019 datasets. Figure 4 shows the comparison of the predicted speed and actual speed of a road segment (a) and the mean absolute error (MAE) of all the road segments for both testing datasets (b). We can see the median MAE while testing on 2018 and 2019 datasets were 1.98 mph and 2.32 mph respectively. The MAE was lower when testing on the 2018 dataset since the HR model was trained on the same dataset. We found unusually large MAE for some road segments when testing on the 2019 dataset. After examining some of those road segments we found that they have very different speed patterns for the year 2019 compared to 2018. An example of such a road segment is shown in Figure 5, which shows that the speed pattern of that road segment changed halfway through 2019. This yields a large MAE while testing on the 2019 dataset since the speed pattern in the testing dataset of 2019 is significantly different from the 2018 dataset used for training the prediction model.

3.3 Analyzing Anomalies

To find the speed anomalies of a road segment, we calculated the difference between the actual speed and the predicted speed for each hour of the year using following equation.

$$\Delta s(t) = s(t) - \hat{s}(t)$$

where $\Delta s(t)$ is the difference between the actual speed $s(t)$ and predicted speed $\hat{s}(t)$ of that hour t for a road segment.

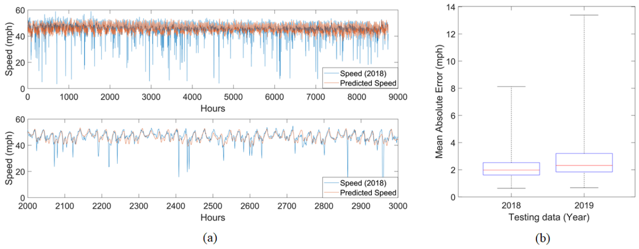


Figure 4: (a) Comparison of predicted speed and actual speed. (b) MAE for testing with 2018 and 2019 data. The whiskers extend to the minimum and maximum data points. The bottom and top edges of the box indicate first and third quartiles, respectively. The central line indicates median value.

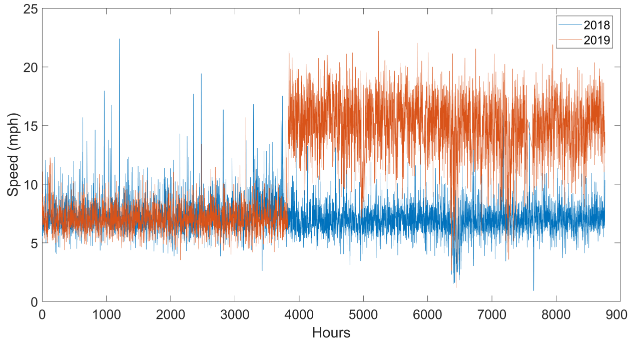


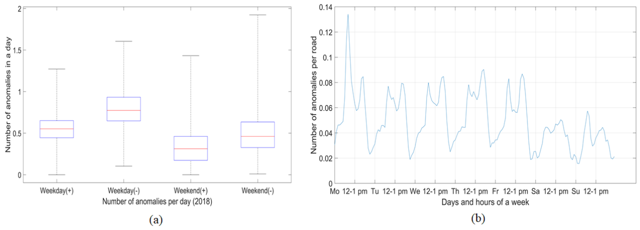
Figure 5: Different speed patterns in 2018 and 2019 for the same road segment.

Then we calculated the mean and standard deviation of these differences, denoted by μ_Δ and σ_Δ. Anomalies were defined as the hours when the difference between actual speed and predicted speed was more than two standard deviations away from the mean. The anomalies for which the actual speed was higher or lower than the predicted speed were defined as positive and negative anomalies respectively. We declared positive and negative anomaly at time t whenever

$$\Delta s(t) > \mu_\Delta + 2\sigma_\Delta$$

$$\Delta s(t) < \mu_\Delta - 2\sigma_\Delta$$

We repeated this process for each of the selected road segments to find positive and negative speed anomalies. Figure 6 (a) shows the number of positive and negative anomalies per day for weekdays and weekends. We found that both positive and negative anomalies are more likely to occur on weekdays than on weekends. Also, the travelers are more likely to experience negative anomalies than positive anomalies on both weekdays and weekends. Figure 6 (b) shows when travelers are more likely to experience anomalies. We found that similar to the weekly speed pattern, anomalies are also more likely to occur during morning and afternoon rush hours. This could be because the effect of unusual traffic conditions can be more intense during rush hours than other times. During off-peak hours, the traffic is lighter and the speed mainly depends on the speed limit rather than how heavy the traffic is. So a similar change



(a)

(b)

Figure 6: (a) Number of positive and negative anomalies per day for weekdays and weekends. The whiskers extend to the minimum and maximum data points. The bottom and top edges of the box indicate first and third quartiles, respectively. The central line indicates median value. (b) Number of anomalies per road segment on each hour of the week.

in the traffic condition will affect the speed more during rush hours, which may also contribute to a higher number of anomalies.

4 TIME: ROUTE STATISTICS

Uber Movement datasets provide hourly average speed data for individual road segments. This can be used to find useful travel time statistics for specific routes in real-time to help travelers make a more efficient travel plan. For example, we can calculate the mean travel times for different hours of the week and find the best and worst hours for traveling on that route. In the 2018 training data, each hour of each week is represented 52 times with a mean speed. From these samples, we can compute a discrete probability distribution of travel times for each hour of the week for individual road segments. By combining these probability distributions, we can further compute how likely a traveler would reach their destination in a given amount of time when traveling at a certain hour of the week. This can help the travelers decide how early they need to leave to reach the destination at a target time with some probability.

We used the example route in Queens, discussed in section 3.1 and shown in Figure 3, to find useful travel time statistics. To find the travel time, we first found the length of each road segment on the route from OSM data. Then we found the travel time for each road segment for each hour of the year in 2018 using the speed data and the length of the segment. For each hour of the week, we found the travel time probability distribution of the route by combining the travel time statistics of all the constituent road segments. The mean and 90th percentile travel time for each hour of the week is shown in Figure 7. We can see that the travel time is longer during the morning and afternoon rush hours on weekdays, and weekends have shorter travel times than weekdays. If we compare the travel time of the 90th percentile to the mean travel time, we can see the difference between them is also larger during weekday morning and afternoon rush hours. This indicates that not only the travel time is longer during those hours, but the variation in travel times is also higher during those hours. This is further demonstrated in Figure 8 which shows how the probability of reaching the destination on time changes at different hours of the week for leaving earlier than the mean travel time of that hour. We can see that the probability of reaching the destination within the mean travel time during weekday rush hours (Mon 8-9 am and Tue 4-5 pm) is lower than off-peak hours (Wed 2-3 am and Sun 10-11 am). This could be because the fastest travel time of a route is restricted by the speed limit of the road: during off-peak hours, the travel times do not decrease much from the mean travel time even if the traffic on the road is less than usual. But if the traffic at that hour is busier than usual, the travel time is more likely to be longer than usual. We do not see this phenomenon during rush hours, because the traffic at those hours is comparatively busier. So both heavier and lighter than usual traffic will most likely affect the travel time similarly. Also, the probability of reaching the destination on time only reaches to about 80% to 90% for those two rush hours if 5 minutes were added to the mean travel time. But it reaches 100% for two off-peak hours we examined if only 1 or 2 minutes were added to the mean travel time.

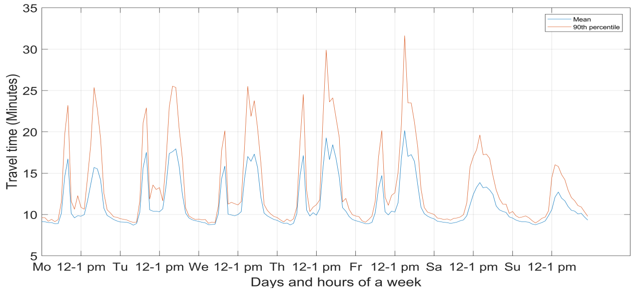


Figure 7: Travel time (mean and 90th percentile) for the second example route in Queens at different hours of the week.

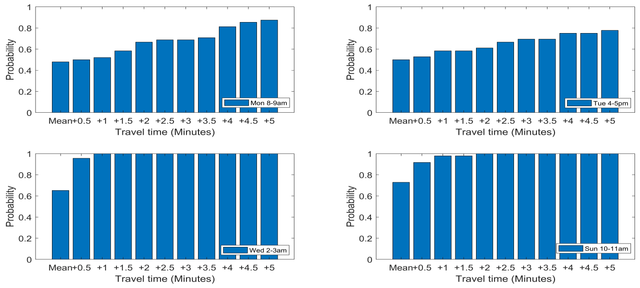


Figure 8: Probability of reaching destination within 1 to 5 minutes from mean travel time of that hour.

5 CONCLUSION

In this study, we explored some of the ways the Uber Movement datasets can be used to examine the traffic patterns of a city and different routes of the city. We also discussed how this analysis can be useful for city planners, ride-sharing service providers, and travelers. Our analysis on two example routes shows that the speed patterns are not the same for all the routes of an area. So finding useful traffic-related statistics like travel time at different hours of the week for individual routes can help the travelers plan their trips better.

REFERENCES

[1] [n.d.]. Data USA. https://datausa.io/profile/geo/boston-ma. Accessed: 2020-04-10.

[2] [n.d.]. Longitudinal Employer-Household Dynamics. https://lehd.ces.census.gov. Accessed: 2020-04-10.

[3] [n.d.]. Uber Movement. https://movement.uber.com. Accessed: 2020-01-29.

[4] Shan Jiang, Joseph Ferreira, and Marta C González. 2012. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery* 25, 3 (2012), 478–510.

[5] Ju Sam Oh, Yong Un Shim, and Yoon Ho Cho. 2002. Effect of weather conditions to traffic flow on freeway. *KSCE Journal of Civil Engineering* 6, 4 (2002), 413–420.

[6] Austin W Smith, Andrew L Kun, and John Krumm. 2017. Predicting taxi pickups in cities: Which data sources should we use?. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 380–387.

[7] Yeran Sun, Yinming Ren, and Xuan Sun. 2020. Uber Movement Data: A Proxy for Average One-way Commuting Times by Car. *ISPRS International Journal of Geo-Information* 9, 3 (2020), 184.

[8] Hubert Verreault and Catherine Morency. 2011. Transcending the typical weekday with large-scale single-day survey samples. *Transportation research record* 2230, 1 (2011), 38–47.

[9] Hao Wu. 2019. Comparing Google Maps and Uber Movement travel time data. *Transp. Find* (2019).

[10] Yaying Zhang and Guan Huang. 2018. Traffic flow prediction model based on deep belief network and genetic algorithm. *IET Intelligent Transport Systems* 12, 6 (2018), 533–541.

[11] Lingxue Zhu and Nikolay Laptev. 2017. Deep and confident prediction for time series at uber. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 103–110.