Taxi Driving Fraud Detection

Pratik Chapadgaonkar David Troupe

Rutgers University

ABSTRACT

Due to many taxi cabs now having an embedded Global Positioning System (GPS) we can collect massive amounts of taxi trajectories throughout urban environments. These GPS records provide an opportunity for us to uncover taxi driving fraud events. In this paper we describe a method of detecting anomalous taxi trajectories. Sometimes taxi drivers can purposefully take longer routes to the destination in an attempt to get a higher fare. This can be a problem for the passengers who are forced to pay higher fares as well as the taxi company who might lose customers to competitors if such fleecing is discovered. Hence detection of such anomalous trajectories can be of paramount importance. Additionally, this technology could possibly be extended to detecting anomalous traffic patterns in general which could help identify unusual road conditions caused due to accidents, weather, construction, or other events. We use machine learning to detect when a trajectory between an origin and destination differs extensively from other trajectories during the same time frame. This allows us to rule out the previously mentioned outside events and classify the driver as malicious. We evaluated our method against real-world taxi trajectories from the Chinese city of Shenzhen.

1. INTRODUCTION

Taxi driving fraud is committed by greedy taxi drivers who deliberately take unnecessary detours in order to overcharge passengers. Many taxi service complaints are directly related to taxi driving fraud. Therefore, it is extremely valuable for taxi companies to be able to access this information so that customers have a high satisfaction rate. However, fraud detection is a challenging problem to solve since experienced drivers often know the city better than their passengers.

Fortunately, the GPS device equipped on modern taxis allows us to examine traces throughout the city. These traces provide the necessary information for large scale fraud detection. This paper proposes the use of machine learning to find anomalous taxis trajectories. We focus our research on the areas that we believe taxi drivers are most likely to take advantage of customers. We hypothesis that tourists are the most likely victims of a taxi driver taking a sub-optimal route. This is because tourists are unfamiliar with the area, and therefore unlikely to notice that the driver is taking an unnecessary detour. However, clever taxi drivers can still manage to take unnecessary detours even with local passengers, since they often know the city better than their passengers. In order to find trajectories likely to be traveled by a tourist we created fixed rectangular areas around key tourist destinations in Shenzhen, such as the airport and train station. We then focus on taxi trajectories that start and end at those key destinations. For each trajectory we calculate the total distance traveled and the total duration of the trajectory. Anomalies can then be detected by comparing the total distance and total time to historical trajectory data. Anomalous routes will have a larger total distance covered, a larger duration or both. Even though all taxi drivers committing fraud have the same motivation, to overcharge passengers solving this problem is by no means trivial.

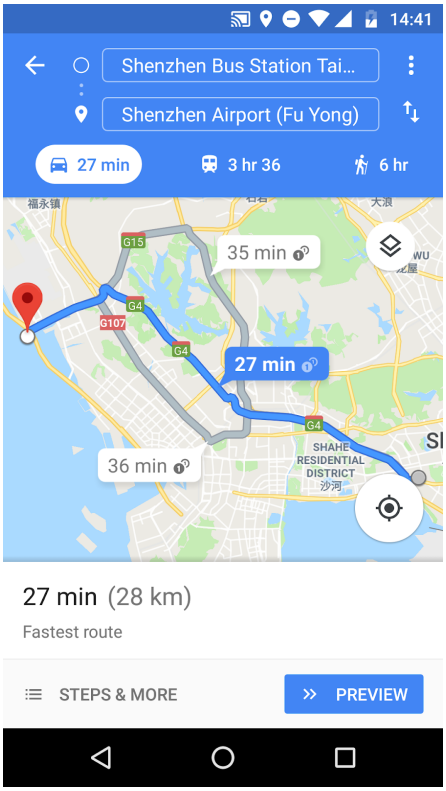


Figure 1: Alternate paths between Shenzhen airport and bus station. Source: Google Maps

A few of the main challenges:

There can be multiple paths between a source and destination. It would be hard to classify a path as anomalous especially if two or more paths are similar in terms of distance and time. As seen in Figure 1 the route highlighted in blue and the route passing through road G107 have similar lengths though the time taken is substantially different. This makes it harder to detect routes that detour locally, but still fall within an overall acceptable distance measurement.

Longer paths might take less time than the shortest route due to traffic conditions and therefore should not be counted as anomalies. As seen in figure one route through G15 takes less time than the one through G107 despite being longer.

Some routes might appear to be anomalies when they are in fact shortcuts. The distance and GPS reading during this route may vary from "normal" routes, but it might be a legitimate shortcut.

If a segment of a road is blocked off a route that was previously classified as anomalous would now have to be classified as the proper route. This detection can't rely on historic data alone and will need multiple taxi trajectories in the same time frame with the same source and destination to be classified properly. However, such data is not always available. Additionally, this problem is more pronounced when dealing with data in real-time.

Some drivers many not be intentionally committing fraud they could truly be unfamiliar with the local area. Additionally, some suspicious activity might be the result of changes in traffic conditions that are difficult to assess in historical and real-time data.

In this paper we develop a taxi fraud detection system equipped with several components to overcome these challenges. First we will identify interesting sites from Shenzhen in order to focus our detection. These sites are locations that are frequently visited as pick-up and drop-off locations by tourists. Between these locations we perform taxi driving fraud detection. In order to provide detection information, we make use of two primary features of each trajectory distance traveled and duration in time. After evaluating all trajectories between these interesting locations we can identify typical routes taken by taxis based on a probabilistic model. Moreover, we can determine the actual route by reconstructing GPS readings during that trajectory. This will allow us to find the most commonly taken routes by taxis in Shenzhen. Trajectories that are not commonly taken will be our initial candidates for fraud. Next, we determine if this suspicious route has a larger than usual time or distance. If either of those values is larger than usual we compare that suspicious route to other routes taken between the same origin and destination at the same time. This allows us to filter out anomalies caused by traffic conditions since other trajectories during that time will be taking the same or similar abnormal routes. Last we look at the data for that specific cab if that cab has been in the area before we determine that the driver is not new to the area and is fact acting maliciously.

We will be using taxi trajectory data from the Chinese city of Shenzhen. The data about the GPS location is updated every few seconds and there are approximately 2,500 records for a single taxi and a total of 350 taxis in the dataset. The dataset also contains occupancy status of the taxi. Which is used to determine whether the taxi is currently serving a customer. For our purposes we are only interested in anomalous routes while a passenger is in the taxi. However, the taxi company might also be interested in anomalous routes when the taxi does not have a passenger. Additionally, the dataset contains the speed of the taxi at every GPS reading.

2. RELATED WORK

Isolation Based Anomalous trajectory (iBAT)[5] was proposed as a solution for detecting anomalous trajectories and detecting changes in traffic network. This method also relied on GPS trajectory data. An improvement on this called iBOAT[4] (Isolation Based Online Anomalous Trajectory) can not only detect the anomalous trajectories but also detect which portion of the trajectory was anomalous in real time.

Others [2] have done work related to outlier detection. Similar methods were used to evaluate the distance traveled by a taxi during a route by evaluating GPS data. Additionally, this research provides guidelines for dealing with some potential excuses for fraud such as being unfamiliar with the area, and traffic conditions. However, their research did not make use of the duration in time of the trajectory. A passenger might prefer to take a physically longer route if it's more time efficient.

Research has also been done on how speed might play a role in detecting taxi fraud [6]. However, this research was based on the assumption that all taxis drivers committing fraud have in some way modified the taxi

meter. While our work does not focus on this specific type of fraud their clustering and route analysis is insightful.

3. MOTIVATION

A study by the National Bureau of Economic Research found that taxis take unnecessarily long detours on about seven percent of routes that originate from an airport [3]. Detection of anomalous taxi patterns can have a variety of applications.

Detection and prevention of taxi drivers intentionally taking customers on longer routes.

Detection of blocked roads and unusual traffic patterns or conditions. Which could be done in real-time if the data is available.

Increased customer satisfaction and driver accountability. The

This method could also be used to identify the routes that are anomalies in an inverse perspective. That is, we could identify the routes between interesting locations that have the shortest distance and duration in time. This would allow for taxi cab companies to optimize their routes.

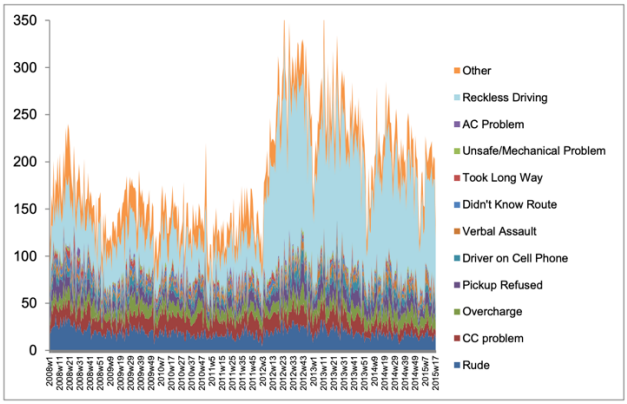


Figure 2: Taxi complaint by subject Chicago. We see that overcharging and taking the long way are both common customer complaints [7].

Additionally, the introduction of ride sharing applications such as Uber and Lyft has had a significant impact on taxis businesses as shown in Figure 3.

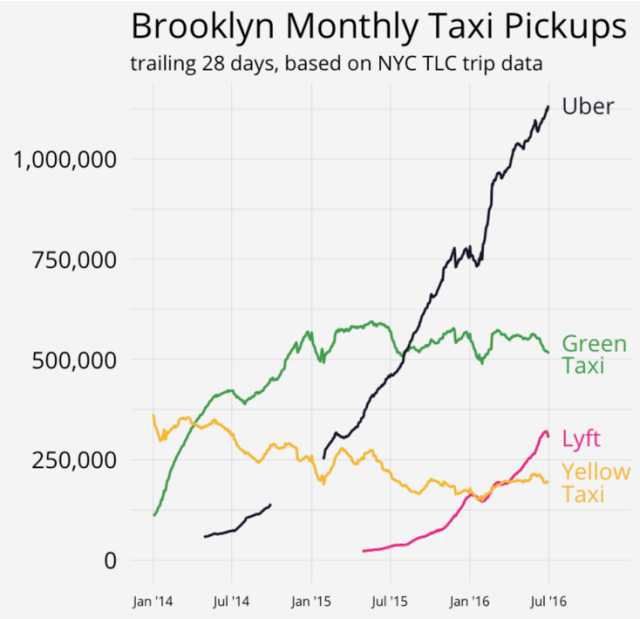


Figure 3: Ride sharing vs taxi pickups in Brooklyn, New York between January 2014 and July 2016 [8].

This decline in taxi ridership has gotten so bad in New York that drivers and committing suicide at rising rates [9] [10] and the city recently placed a limit on the number of cars applications like Uber and Lyft can have in the city [11]. Taxi companies being able to prevent fraud on the part of their drivers would provide taxi companies an advantage over their ride sharing competition and perhaps help combat this growing problem for millions of drivers across the United States.

4. MAIN DESIGN

4.1 Challenges

For detection of anomalous trajectories, we first must fix a starting point and destination. We calculate a range of GPS points corresponding to the source and destination. To ensure we have plenty of data for analysis for a source-destination pair we will only be looking at major hubs like train stations or airports additionally these locations are often hotspots for tourists, who are the most likely victims of taxi fraud. Then we will analyze all trajectories with this source and destination to discover the route they took. By examining intermittent GPS readings, we are able to determine the route taken. Analyzing data over multiple taxis and multiple trips by the same taxi for a source-destination (S-D) pair we should be able to determine which routes are not fraudulent for each S-D pair. A valid route will be one that is taken by numerous taxis traveling from the same (S-D). Then we will calculate the distance and time distribution for these typical non-fraudulent routes.

Trajectories which deviate a lot from these routes are potential candidates for fraud. However, we still need to determine the cause of these anomalies. They could be caused by: 1) Taxis trying to cheat their customers, 2) The possibility that a taxi driver new to the area doesn't know the optimal route or 3) If the road segment is blocked during that time due to traffic, accidents or any other reasons.

To tackle point (2) we can simply check the previous trajectories of the same taxi. Our analysis is only performed on one day worth of data and so this will not be possible in this study, but can be easily done with more data. If this driver has operated in this area previously and not taken the same route we are able to determine that they are in fact acting maliciously. For case (3) we can check the trajectories of other taxis for the same S-D pair during same time period. If other drivers are also taking the same route, then we conclude that there must be some traffic condition causing the change in route. However, finding two or more taxis with the same source and destination at the same time can be hard except between certain S-D pairs and that too only during peak hours. However, this still isn't enough information to classify the route as fraud. We still have to look at the overall time taken for the route. We can see that some suspicious trajectories detected by distance alone, may just be shortcuts instead of frauds. Such as the cyan route in Figure 4 below.

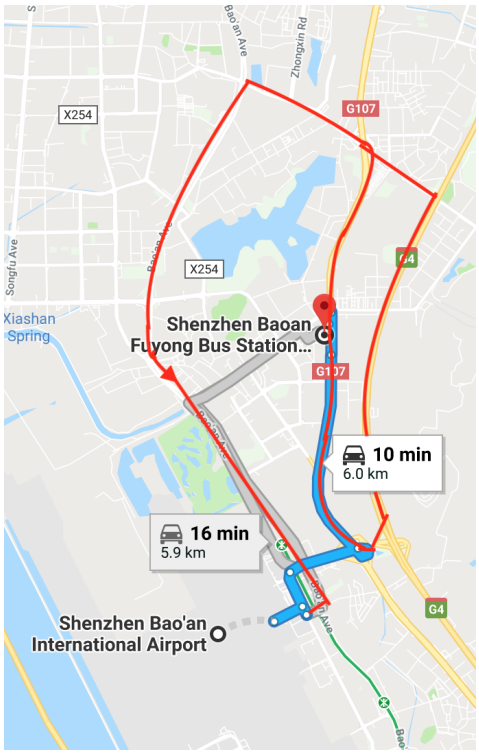


Figure 4: Illustration of potential anomalous trajectories (in red) in Shenzhen between the airport and bus station. Recommended trajectories are (in blue and gray) are provided by Google Maps.

The distance covered in that short-cut might be larger and the path might be different, but if the time taken during the route in within our margin of error it is not fraud. If the time taken is more than our margin, then we can conclude that the driver is off route and they are in fact acting with malicious intent.

Additionally, a trajectory could have normal time, but take much longer than other routes. Since taxi fare is based on a combination of time and distance this is another potential method for fraud. However, since we are replicating routes via their intermittent GPS readings this trajectory would already be flagged as an anomaly without having to look at its time. Thus our method will cover both potential sources of fraud.

4.2 Data source

Our data consists of one day of detailed trajectory data for the Chinese city of Shenzhen with 449,139 passenger trajectories. It consists of 14,728 taxis with about 3,000 GPS points for each taxi. The data is in CSV file with the format: Taxi ID, Time, Latitude, Longitude, Occupancy Status, Speed. Occupancy Status is a binary variable whose value is 1-with passengers and 0-without passengers**.**

Some of these trajectories however have very few or infrequent GPS readings. This makes it challenging to reconstruct the true trajectory. Therefore we filtered our data to discard such trajectories.

In spite of the seemingly huge amount of data the number of trajectories for a S-D pair is low. Between the airport and the train station we found only 35 trajectories that qualified for analysis (21 from the airport to the train station and 14 from the train station to the airport). We did however manage to find 479 trajectories between the north train station and the west train station (283 from the west train station to the north train station and 196 from north train station to the west station).

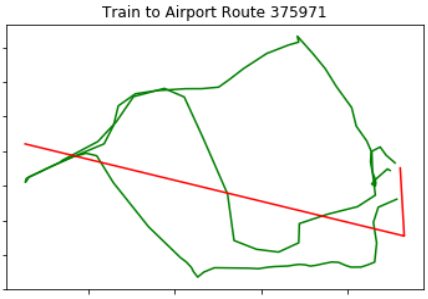


Figure 5: Route with infrequent GPS readings (red) appears as two straight lines. A trajectory that's clearly impossible. Google Maps recommended routes (green)

4.4 Mapping routes to grids

Given the different times and routes possibly taken by a taxi comparison of raw GPS trajectories proposes to be a significant challenge. For simplifying this task we discretize the world into grids. Each grid cell corresponds to a 0.05 degree change in latitude and longitude. It is possible for a taxi to send multiple GPS readings from a single cell hence we consider each cell only once in a trajectory.

Once routes are mapped to grids we try to find anomalies among them using 2 approaches:

- Detecting Common Subsequence of Grids among routes (DCSGR)
- Comparison with Google-Maps Recommended routes (CGMRR)

4.4 Detecting Common Subsequence of Grids among routes (DCSGR)

We discretize each route into a unique sequence of grids. These sequences are compared against each other to find a measure of similarity between them. Routes which are different from others beyond a certain threshold are classified as anomalous. There will always be slight variations in the start and destination locations for every trip which will cause every trip to be nearly unique. So a similarity threshold is defined instead of considering a perfect one is to one matching of grid cells. After performing experiments a threshold of 80% similarity was chosen. Any route having less than 80% similarity to other routes gets classified as anomalous. This method assumes that fraudulent routes are few and most of the detours might be due to road network restrictions or traffic. Since a detour caused due to either of these causes would be taken by multiple drivers they are considered as not anomalous. Varying the similarity threshold can generate different results. Increasing it guarantees that all anomalies are detected but also increases the number of false positives. Lowering it guarantees the absence of false positives but also misses out on actual anomalies. The threshold must be chosen after careful experimentation so as to maximize the true positive rate while minimizing the false positive rate.

4.5 Comparison with Google-Maps Recommended routes (CGMRR)

In this approach we use Google maps to recommend the best routes between a S-D pair. Google maps usually recommends more than one route. We extract the GPS coordinates of these routes and map them to grids like we do for the taxi trajectories. Then we compare the sequence of grids for routes from Google maps to the sequence of grids for our trajectories. Similar to the previous method a similarity threshold is defined keeping in mind slight variations in starting and ending coordinates as well as traffic and road network restrictions. As with DCSGR increasing similarity threshold guarantees that all anomalies are detected, but also increases the number of false positives. Lowering it guarantees the absence of false positives but also misses out on actual anomalies. The threshold must be chosen after careful experimentation so as to maximize the true positive rate while minimizing the false positive rate. This method can account for anomalies caused due to traffic however it cannot account for all the anomalies caused due road network restrictions. Eg. a taxi might be on the wrong side of the road and might have to take a u-turn before proceeding towards the destination. Or to access a parking lot near a train station or airport a longer detour might be necessary. This serves as our baseline method.

5. PERFORMANCE EVALUATION

5.1 Determination of ground truth

Which trajectories are really anomalous? To evaluate the performance of our algorithms we need to know which trajectories are really fraudulent. A few studies [3][4] of this kind used humans with intimate knowledge of the road network to manually label the trajectories as

anomalous. Without access to a intimate knowledge of the road network we try to best approximate the ground truth by plotting the GPS trajectories and marking trajectories straying off the path for no obvious reason as anomalous.

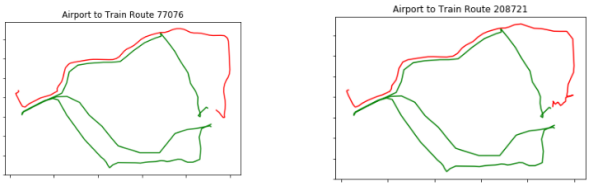
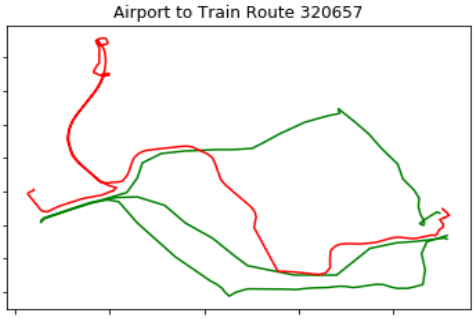


Figure 6

The figure above shows 2 similar trajectories taken by 2 different taxis. Red shows the trajectory taken by the taxi and green shows the routes recommended by Google maps. While both these routes are much longer than the recommend routes it would be highly improbable for two different taxis to take the same fraudulent routes. Hence it might be safe to conclude that these routes are the not fraudulent but the result of road network restrictions. The presence of more similar looking routes in our data would support this hypothesis.



Here we observe a route that takes a long and unnecessary detour. However, observing that the detour starts and ends at nearly the same location we must also consider the possibility that it was requested by the customer for some purpose of their own such as picking up someone or buying something. In the absence of any data regarding customer complaints we regard this as a fraudulent trajectory.

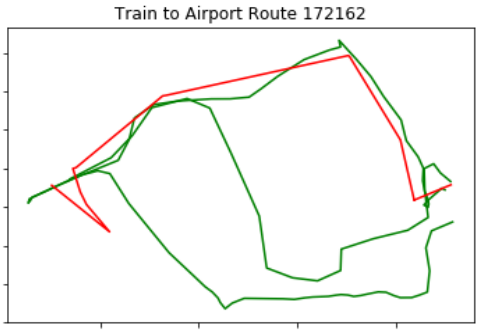


Figure 7

Here we observe a trajectory with a sharp unnecessary deviation which later follows the recommended route. This deviation could possibly be caused by being on the opposite side of road with no possible u-turn. However in the absence of other similar trajectories we treat it as an anomaly.

5.2 Evaluation of Algorithms

To evaluate the performance of our algorithms we consider 2 metrics:

- True Positive Rate (TRP): Number of anomalies detected correctly / Total number of anomalies. Ideal value=1
- False Positive Rate (FRP): Number of valid routes classified as anomalies / Total number of valid routes. Ideal value=0

In our experiments we consider two S-D pairs:

- Shenzhen airport to north Shenzhen train station
- North Shenzhen train station to west Shenzhen train station.

We found 35 trajectories for the first S-D pair and 479 for the second one.

CGMRR evaluation

Train station to airport

		True	False
80% similarity	Positive	0	0
	Negative	13	1
85% similarity	Positive	1	2
	Negative	11	0
90%similarity	Positive	1	4
	Negative	9	0

Shenzhen airport to Shenzhen train station

For 80% similarity

TPR = TP / TP + FN = 0/0+1 = 0
FPR = FP / FP + TN = 0/0+13 = 0
For 85% similarity
TPR = TP / TP + FN = 1/1+0 = 1
FPR = FP / FP + TN = 2/2+11 =2/13 = 0.1538
For 90% similarity
TPR = TP / TP + FN = 1/1+0 = 1
FPR = FP / FP + TN = 4/4+9 = 4/13 = 0.3076

Airport to train station

		True	False
80% similarity	Positive	1	0
	Negative	14	6
85% similarity	Positive	1	0
	Negative	14	6
90%similarity	Positive	5	6
	Negative	8	2

For 80% similarity

TPR = TP / TP + FN = 1/1+6 = 1/7 = 0.1428
FPR = FP / FP + TN = 0/0+14 = 0
For 85% similarity
TPR = TP / TP + FN = 1/1+6 = 1/7 = 0.1428
FPR = FP / FP + TN = 0/0+14 = 0
For 90% similarity
TPR = TP / TP + FN = 5/5+2 = 5/7 = 0.7142
FPR = FP / FP + TN = 6/6+8 = 6/14 = 0.4285

North Shenzhen train station to west Shenzhen train stations.

West to north train station for similarity threshold 85%

	True	False
Positive	7	0
Negative	275	1

TRP=0.875

FRP=0

North to west station for similarity threshold 75%

	True	False
Positive	2	8
Negative	182	4

TRP=0.33
FRP=0.0421

DCSGR evaluation

Airport to train station

	True	False
Positive	3	0
Negative	17	1

TPR = TP / TP + FN = 3/3+1 = 3/4 = 0.75
FPR = FP / FP + TN = 0/17+0 = 0

Train station to airport

	True	False
Positive	1	0
Negative	13	0

TPR = TP / TP + FN = 1/1+0 = 1
FPR = FP / FP + TN = 0/7+0 = 0

West to north train station

	True	False
Positive	6	2
Negative	273	2

TRP=0.75
FRP=0.0072

North to west train station

	True	False
Positive	0	0
Negative	190	6

TRP=0
FRP=0

Our observations showed that on average the DCSGR outperforms CGMRR. This because the former can

account for most of the detours happening due to road network restrictions. As observed in the results our algorithms worked really well in certain situations and not so well in others. This is likely the result of the diverse nature of road networks. Both the algorithms give a low performance for routes ending at the North Shenzhen train station. The north Shenzhen train station is near a residential area which has many small parallel roads which could be taken by a taxi. It also has multiple drop-off points at which passengers may choose to get off. To find similar routes we only compared the starting and ending grid, which provides a lot of wiggle room. Perhaps the side of the grids is appropriate for comparing routes, but a smaller grid size is needed for S-D comparison. Additionally, better information about road network itself would be helpful. Perhaps Google Maps isn't the best resource to direction in this area. Moreover, an equivalence relation needs to be established such that two parallel roads would be classified as the same route (ideal grid size). Alternatively, this problem could be addressed with the help of more data which would ensure multiple trajectories for valid detours, shortcuts, and perhaps near real-time obstructions.

6. FUTURE WORK

This study was done using only one day of trajectory data. Most of the related studies used about a month of data for their analysis. Access to more data would definitely help improve our results and might also allow the use of more sophisticated analysis methods such as clustering. A major hurdle in this study was determination of what qualifies as the ground truth. Knowledge about road network constraints and real time traffic data could help solve this problem. Additionally, data about customer complaint will also be useful in the classification of ground truth. In the future we plan to try and make use of customer complaints to label fraudulent trajectories. Additionally, we will work with a much larger dataset 30 to 60 days of data would allow for more precise fraud detection.

7. CONCLUSION

This study shows a way to detect taxi driving fraud especially in the presence of limited data. It showcases some of the difficulties encountered in this task along with their solutions and possible improvements to this study given the availability of further resources and data.

8. REFERENCES

[1] Ge, Y., Xiong, H., Liu, C., & Zhou, Z. (2011). A Taxi Driving Fraud Detection System. *2011 IEEE 11th International Conference on Data Mining*. Retrieved October 27, 2018.

[2] Liu, M., Brynjolfsson, E., & Dowlatabadi, J. (2018). Do Digital Platforms Reduce Moral Hazard? The Case of Uber and Taxis. Retrieved October 27, 2018.

[3] iBOAT: Isolation-Based Online Anomalous Trajectory Detection Chao Chen, Daqing Zhang, Member, IEEE, Pablo Samuel Castro, Nan Li, Lin Sun, Shijian Li, and Zonghui Wang

[4] iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces Daqing Zhang , Nan Li , Zhi-Hua Zhou , Chao Chen , Lin Sun , Shijian Li

[5] Taxi Trajectory Data Source: http://www-users.cs.umn.edu/~tianhe/BIGDATA/Feeder/TaxiData/TaxiData

[6] Liu, S., Ni, L. M., & Krishnan, R. (2014). Fraud Detection From Taxis Driving Behaviors. *IEEE Transactions on Vehicular Technology,63*(1), 464-472.

[7] Wallsten, S. (2015). The Competitive Effects of the Sharing Economy: How is Uber Changing Taxis?The Competitive Effects of the Sharing Economy: How is Uber Changing Taxis? *Technology Policy Institute*. Retrieved October 28, 2018, from https://www.ftc.gov/system/files/documents/public_comments/2015/06/01912-96334.pdf.

[8] Schneider, T. (n.d.). Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. Retrieved from http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

[9] Fitzsimmons, E. G. (2018, October 02). Suicides Get Taxi Drivers Talking: 'I'm Going to Be One of Them'. Retrieved from https://www.nytimes.com/2018/10/02/nyregion/suicides-taxi-drivers-nyc.html

[10] Suicides shake cabbie community as ride-hailing companies take over. (n.d.). Retrieved from https://www.nbcnews.com/news/us-news/shadow-uber-s-rise-taxi-driver-suicides-leave-cabbies-shaken-n879281

[11] Fitzsimmons, E. G. (2018, August 08). Uber Hit With Cap as New York City Takes Lead in Crackdown. Retrieved from https://www.nytimes.com/2018/08/08/nyregion/uber-vote-city-council-cap.html

[12] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. UrbanCPS:a Cyber-Physical System based on Multi-source Big Infrastructure Data for Heterogeneous Model Integration. In the 6th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS'15), 2015.