Challenges and Opportunities in Taxi Fleet Anomaly Detection

Rijurekha Sen
Singapore Management University
rijurekhasen@smu.edu.sg

Rajesh Krishna Balan
Singapore Management University
rajesh@smu.edu.sg

ABSTRACT

To enhance fleet operation and management, taxi operators instrument their vehicles with GPS receivers and network connectivity to servers. Mobility traces from such large fleets provide significant information on commuter travel patterns, traffic congestion and road anomalies, and hence several researchers have mined such datasets to gain useful urban insights. The fleet companies, however, incur significant cost in deploying and maintaining their vast network of instrumented vehicles. Thus research problems, that are not only of interest to urban planners, but to the fleet companies themselves are important to identify, to attract and engage these companies for collaborative data analysis.

In this paper, we show how GPS traces from a taxi company can be used to answer three different questions that are of great interest to the taxi operator. These questions are 1) What is the occupancy rate of the taxi fleet?, 2) Do taxi drivers often take inefficient routes when serving passengers, and 3) Are there a large number of taxi drivers who are traveling significantly faster than the posted speed limits? We provide answers to each of these questions using a 2 month dataset of taxi records collected from about 15,000 taxis located in Singapore.

The goal of this paper is to stimulate interest in the questions listed above (as they are of high interest to fleet operators) while also soliciting suggestions for better techniques to solve the problems stated above.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Information System Applications

Keywords

GPS, taxi fleet, anomaly detection

This research is supported by the Singapore National Research Foundation under its IDM Futures Funding Initiative and administered by the Interactive Digital Media Programme Office, Media Development Authority. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agency, or Singapore Management University.
 SENSEMINE'13, November 14 2013, Roma, Italy
http://dx.doi.org/10.1145/2536714.2536715

1. INTRODUCTION

The need to continuously improve their service offering while reducing their costs has caused many logistics and transportation companies to install GPS devices in their entire vehicular fleets. These devices allow the companies to monitor the movement of their fleet in real-time and use that data to improve the efficiency, reliability, and safety of their fleet. The availability of GPS traces from these fleets has also allowed researchers to investigate various topics related to planning, mobility, efficiency, and other areas involving location traces.

For example, in our prior work [2], we built and deployed a system that uses historical records from an entire taxi fleet (comrising of over 15,000 taxis) to provide commuters with the expected travel time and fare for any taxi trip that they might take. Easy Tracker [3], used 3 months of traces from the Chicago Transit Authority, to discover both routes and schedules of transit vehicles from their GPS traces and to also predict the arrival times of these vehicles. Liu et al. [5] used a 3 month trace from 33,000 taxicabs in Beijing to investigate the spatial and temporal causes behind anomalies in traffic situations while Zheng et. al [7] used the same dataset to find faults in urban planning and design, by simultaneously mining people's travel patterns and commonly occurring urban hotspots. Thus fleet GPS traces can be a great starting point for many different research avenues, that either offers important insights into transportation engineering problems or provides exciting commuter applications.

However, even though the research community benefits from these GPS traces, do transportation companies have sufficient incentives to share their GPS traces with researchers? In particular, the companies incur significant installation costs in deploying GPS devices in a large fraction of their vehicles, and also considerable communication costs in transferring the GPS information from their road-going vehicles to their back-end servers. Thus it is quite unlikely that these transportation companies would be willing to share their data with researchers who do not produce results that are of direct relevance to the companies themselves. Rather, our multi-year experience with working with a large taxi company in Singapore has convinced us that identifying questions of *mutual* interest that have both exciting research potential and important operational consequences for the transportation company, is the key to a successful long-term collaboration.

In this paper, we explore a set of these *mutual* interest questions, that have both exciting data analysis aspects on one hand, and important operational efficiency results for the collaborating taxi company on the other. In particular, we mine millions of taxi records spanning several months to answer the following three questions - 1) What is the occupancy rate of the taxi fleet at different times of the day? The answer to this question will allow the taxi com-

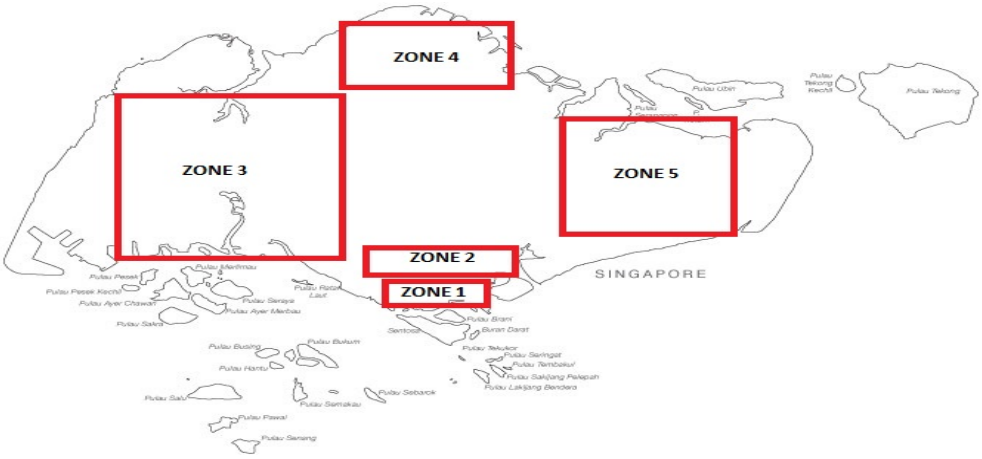

Figure 1: Rectangular zones under observation

pany to determine if its fleet is fully utilised and / or sufficient. 2) Do taxi drivers often take inefficient routes when servicing a passenger, increasing both travel time and fares for passengers? This in turn turn causes bad publicity for the company. 3) Are there a large number of taxi drivers who are traveling significantly faster than the road speed limits? This can potentially cause the company increased maintainence and accident related costs.

We show, in the rest of the paper, how we can answer these three questions using a variety of fairly standard data mining and analysis techniques. Our goal in this work is to a) explain three different problems that are of high interest to taxi fleet operators and which can spur future research, b) demonstrate possible solutions to these questions using standard methods, c) provide more insights into the operational characteristics of a large taxi fleet in Singapore, and finally, d) solicit suggestions for other techniques, with different accuracy, computational efficiency, and complexity tradeoffs, that can provide solutions for these types of problems.

2. BACKGROUND & DATASET

Singapore has a world-class public transportation system with an extensive network of taxis, buses, and rapid transit rail lines that provide convenient and affordable services to the city-state's population of 4.5 million. In particular, taxis are widely available and relatively low-priced (metered fares rarely exceed US $15). This affordable and accessible public transportation network, coupled with high taxes on both private cars and petrol, result in many Singaporeans choosing not to own a car.

We have been collaborating, for many years, with one of the taxi companies which operates a fleet of slightly more than 15,000 taxis (about 60% of the total taxis on the roads) in Singapore. We use two months of GPS-enhanced data from the taxis operated by this company for the results shown in this paper. In particular, we use the following two datasets, one or both of which are necessary for exploring the three questions in this paper.

a) Log Data: This dataset contain records about the instantaneous state of every taxi in the fleet. This dataset is used in our analysis of demand supply mismatch issues (Section 3) to know where taxis are located geographically and to determine the occupancy ratio for each taxi. This dataset is also used in our speeding analysis (Section 5), to compute how long it took a particular taxi to cover a certain distance.

Each record in this dataset consists of a timestamp, the taxi ID, the GPS coordinates of the taxi, and the state of the taxi. Taxis can be in one of eleven states at any point in time. However, for the purpose of this work, we only the following eight states: *OFFLINE* (taxi is not operational), *BREAK* (taxi driver is on a break and the taxi will not pick up passengers), *FREE* (taxi is active, empty, and looking for passengers), *BUSY* (taxi is available for street pickup but but not for booking call pickups), *ONCALL* (taxi is on the way to service a special booking call), *POB* (passenger on board – taxi is taking a metered customer to a destination), *PAYMENT* (meter has just been stopped and driver is receiving payment, meter will be reset after payment is made) and *STC* (taxi is soon to drop the current passenger and ready to accept booking calls).

b) Trips Data: This dataset contains information of every paid trip that a taxi made. It contains the starting (where the passenger was picked up) and ending (where the passenger alighted) GPS coordinates of the trip, start and end times of the trip, and the distance the taxi traveled during the trip. This dataset is used to detect outlier trips in our route inefficiency analysis (Section 4), by comparing the expected time, and distance of a trip with the recorded time and distance.

3. OCCUPANCY RATE OF TAXIS

In this section, we analyse the occupancy rate of taxis to identify potential mismatches between a) the geographical areas having demand for taxis or where the potential passengers are and b) the areas with abundant supply of available taxis or where empty taxis are hunting for passengers. These inefficiences are important for the transportation company to identify as it increases both taxi driver and passenger unhappiness as passengers cannot find a free taxi while drivers are unable to earn fares.

To perform this analysis, we divided Singapore into five rectangular zones as shown in Figure 1. These zones were chosen to contain unique characteristics, based on the types of business and/or residential areas present in each zone.

Zone 1, the Central Business District (*Central*), contains most of the high-rise office buildings in the city, with relatively few residential areas. **Zone 2** is labeled as *Condo* because of its concentration of private condominiums and landed housing. This zone also includes Orchard Road, Singapore's main retail shopping district. **Zones 3, 4 and 5**, labelled *West*, *North* and *East* respectively,

include most of Singapore's public housing, with *West* containing a mix of public housing and industrial estates, *North* containing a mix of public housing and unpopulated areas, and *East* containing mostly public housing.

To determine utilization of the taxi fleet, we first define an *occupancy* metric as $O/(O+A)$, where $O =$ the number of minutes the taxi was occupied (in state *POB* or *ONCALL* or *PAYMENT* or *STC*) in the given zone and hour, and $A =$ the number of minutes it was available (in state *FREE* or *BUSY*). Note that we exclude from the denominator periods in which the driver is on break or otherwise not actively servicing or seeking to service passengers.

Figures 2 to 11 show, for each zone, the median occupancy values on the left y-axis and the number of taxis in that zone (for which the median occupancy was computed) on the right y-axis, averaged over 31 days in the month of Oct 2012. We computed separate values for weekdays and weekends. The x-axis reports the time of the day in hours. Some of the key observations that we can make from the graphs are as follows.

- The occupancy values are higher in zones 1 and 2 as they have more commercial activities than the other three zones, which are primarily residential with some industrial estates.

- The morning peak in occupancy on weekdays is visible for all the residential zones (especially zone 5), when people will leave from these areas in taxis and come to offices in zone 1 and zone 2. On weekends, this morning increase in occupancy is again present in the residential areas, probably for shopping or leisure activities, but starts later than office going activites on weekdays.

- The number of reporting taxis are higher in zone 1 and zone 2, where occupancies are higher as well. But for some zones at some times, for example zone 5 (Fig. 10), number of reporting taxis do not go down in late noon and afternoon, though occupancies are very low. This might be a situation, when empty taxis are roaming in these areas, looking for passengers while the demand is not that high.

  Also number of reporting taxis in zone 2 ((Fig. 4) is significantly higher than in zone 1 ((Fig. 2), though occupancies in zone 1 are equal, if not greater than in zone 2. Thus if some empty taxis move from zone 2 to zone 1, the occupancies for both zones might improve.

  A similar discrepancy can be observed between zone 4 ((Fig. 8) and zone 5 ((Fig. 10), with the former showing higher occupancies and lower number of reporting taxis, while the latter shows the opposite.

  A third such mismatch in demand-supply is observable for zone 4 ((Fig. 8), where number of reporting taxis do not change according to the increase and decrease in occupancies.

  All these cases exhibit a possible demand-supply mismatch issue, where supply of available taxis do not match the demand shown by occupancy levels.

- The occupancy drops in the late hours of the night, more so on weekdays than on weekends, as on weekends people might indulge in late night activities. Also buses and MRT services are suspended in the late hours, so taxis are the only means of transport at that time. Other than those late hours, weekends show much less activities than weekdays, both for occupancy and number of reporting taxis.

4. INEFFICIENT ROUTES

In this section, we identify taxi routes that are inefficient across three different efficiency metrics; We start with distance – a route between two end points is flagged as inefficient if it is significantly longer than another route between those same end points. We then expand our analysis to understand whether those distance-inefficient routes are actually efficient for two other metrics – speed and pricing.

For every trip recorded in the dataset, we find the estimated distance and time taken for that trip as reported by Google Maps (using the Google Maps API [4]). Then we use this estimated time to compute an estimated fare for the trip using the same fare computations used by the taxi's meter system. Finally, we identify the potentially anomalous trips that have the highest discrepancies between the actual distance, time taken, and fare, and the Google Maps-reported estimated distance, time, and estimated fare.

A short description of Singapore taxi pricing is warranted here. The Singapore taxi fare structure [6] uses a fixed starting fare augmented with distance-based charges and a combination of extra charges depending on location and time. The major component of the metered fare is the distance travelled. In addition to the starting and distance-based fares, there are also some time and location-based surcharges. For example, to reduce peak-hour traffic congestion on major highways and in the Central Business District (CBD), Singapore uses an Electronic Road Pricing (ERP) system that uses RFID to automatically charge drivers when they enter an ERP zone. Passengers also incur a Peak Hour Surcharge during peak business hours. In this analysis, we removed the ERP charges from the fare as they are not extra earnings for the driver – the driver pays the full ERP charge and the passenger reimburses the charge to the driver. We also remove the starting fare and all other time-based and location-based charges as they will apply equally to both efficient and in-efficient routes (as they start at the same time at the same place). We thus only consider the distance-related component of the fare in this analysis.

We see that at the 20% threshold (Google Maps values are at least 20% lower than the actual trip value), 6.37% of the trips were longer than they needed to be while 43.39% of the trips and 7.42% of the trips took longer and charged a higher fare respectively. When the threshold was raised to 50%, the inefficient trip percentage dropped to 2.30%, 14.98%, and 2.90% with respect to distance, time, and fare (in that order).

Next we seek to validate if longer routes are faster, i.e. taxi drivers are intentionally choosing an inefficient route to reduce travel time. We extract all the trips (at both -20% and -50% thresholds) that were longer than they needed to be and compared the difference between Google Map's suggested trip time and the actual trip time. We found that for the -20% threshold, 88.75% of the longer trips also took more time while at the -50% threshold, 94.35% of the trips took more time than they should have. Hence, we reject this hypothesis and state that longer routes do not generally result in shorter travel time.

Finally we analyse the increase in fare that passengers have to pay due to route inefficiencies. Table 1 shows the results of this analysis. The Mean values show the average % difference between what the passenger could have paid with the shorter route and what they actually ended up paying ($\frac{Shorter-Actual}{Shorter}$). The standard deviations (in %) are given in brackets. All results are not cumulative (i.e, no intersection between distance buckets).

For both thresholds (20% and 50%), it shows how many of those longer routes were for a particular distance. For example, 15.55% of the longer routes (at 20% threshold) were between 2 to 4 kilometres in length. For each of these route buckets, we computed

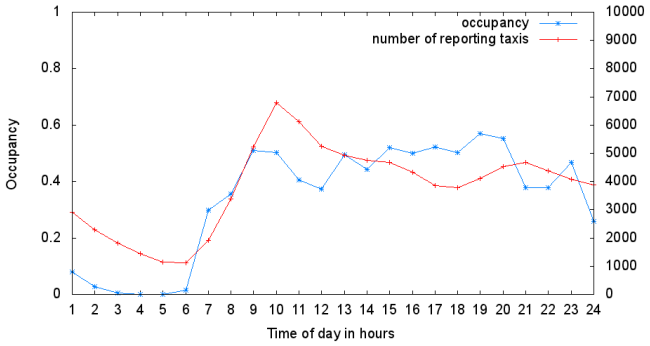

Figure 2: Zone 1 on weekdays in Oct, 2012

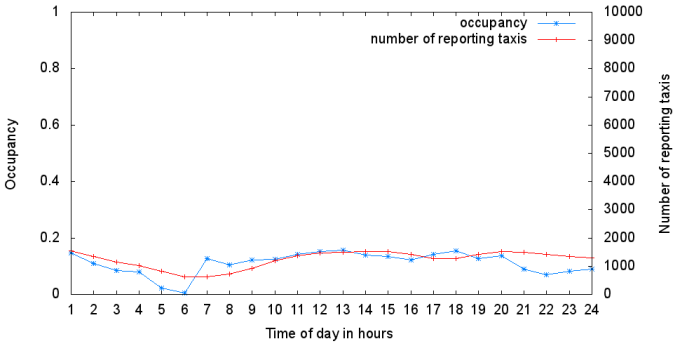

Figure 3: Zone 1 on weekends in Oct, 2012

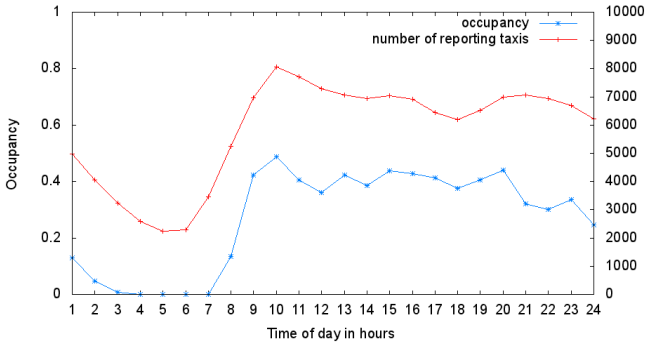

Figure 4: Zone 2 on weekdays in Oct, 2012

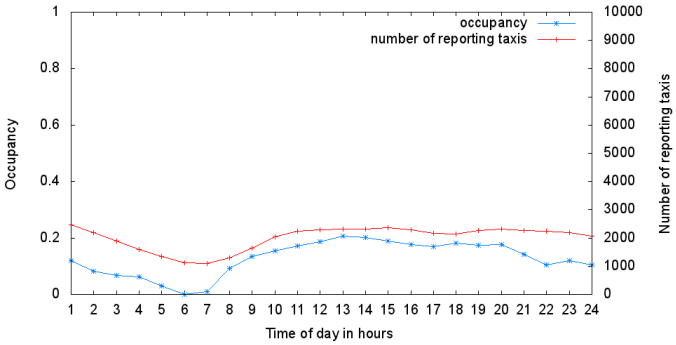

Figure 5: Zone 2 on weekends in Oct, 2012

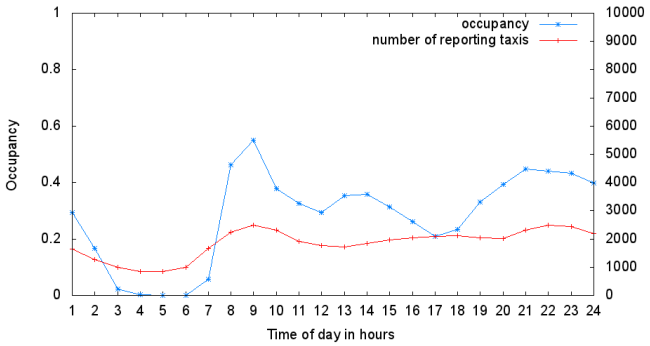

Figure 6: Zone 3 on weekdays in Oct, 2012

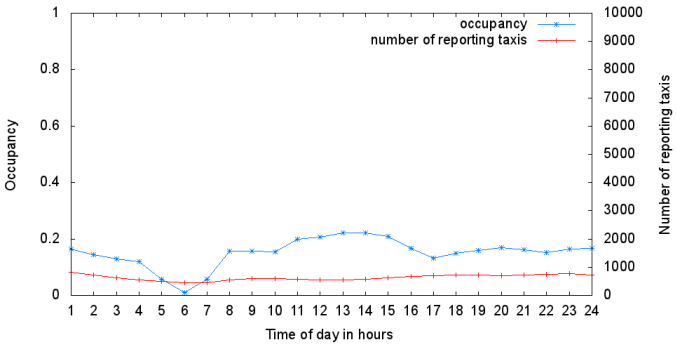

Figure 7: Zone 3 on weekends in Oct, 2012

the average percentage difference in fare that a passenger paid as a result of the inefficient route. For example, for routes (at threshold 20%) between 6 to 8 kilometres in length, the passenger could have paid 63% less on average (with a standard deviation of 74%).

These results suggest that inefficient routes tend to increase as the distance of the trip increases – note the increasing trend in the percentage of trips that are inefficient as the distances increase (31.22% of the inefficient routes at 50% threshold were greater than 16 kilometres in length). In addition, the fare difference is relatively higher as the distance increases. This suggests that passengers travelling longer distances do need to pay more attention to the routes chosen by the taxis.

5. EXCESSIVE SPEEDING

In this section, we investigate the issue of excessive speeding. This is an important consideration for transport network operators as exceeding specified speed rules can not only increase fuel consumption and related costs but also increase the probability of traffic accidents, raising cost in the form of repairs, possible litigation and higher insurance premiums.

Singapore has six different speed limits of 40, 50, 60, 70, 80, and 90 km/h respectively, for different road stretches. The document "Speed Limits of All Roads" at [1], gives road specific speed limit values, as of July 2013.

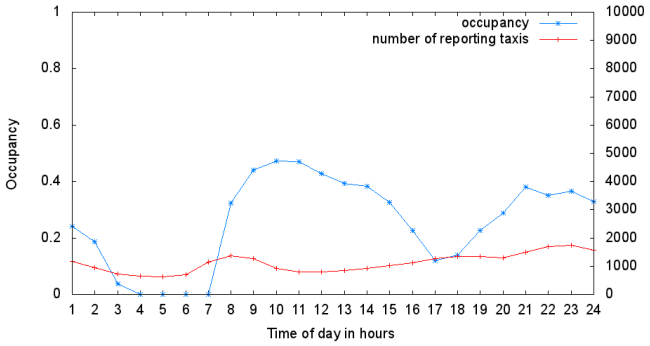

Figure 8: Zone 4 on weekdays in Oct, 2012

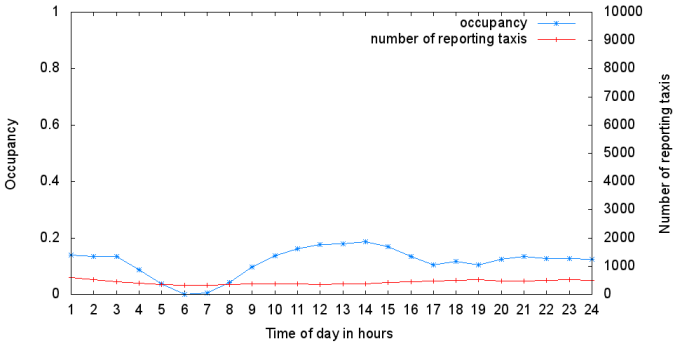

Figure 9: Zone 4 on weekends in Oct, 2012

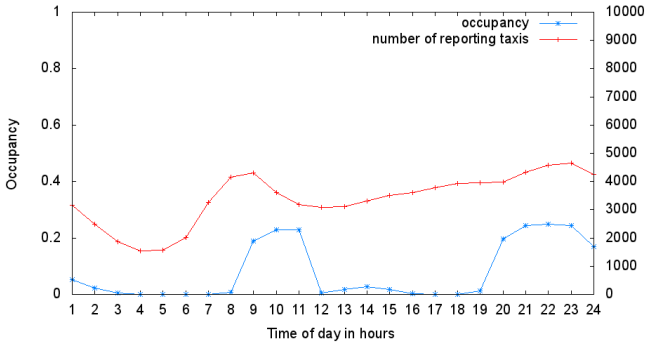

Figure 10: Zone 5 on weekdays in Oct, 2012

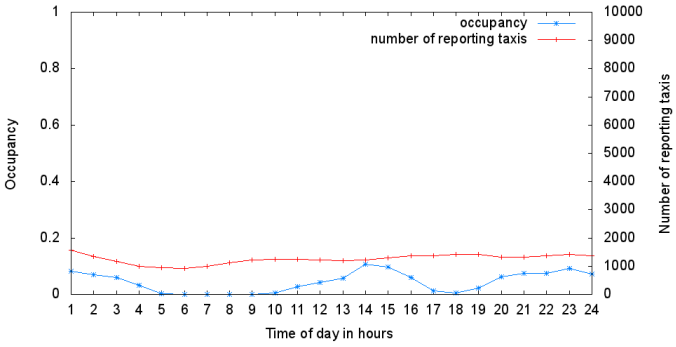

Figure 11: Zone 5 on weekends in Oct, 2012

Route	20% Thres.		50% Thres.	
Distance (km)	%	Mean (Std.)	%	Mean (Std.)
0 - 2	3.67	-36 (14)	1.56	-62 (8)
2 - 4	15.55	-49 (33)	14.14	-87 (36)
4 - 6	12.73	-59 (56)	12.76	-111 (70)
6 - 8	11.03	-63 (74)	11.02	-125 (101)
8 - 10	10.32	-64 (88)	9.74	-134 (129)
10 - 12	8.75	-54 (97)	7.94	-120 (148)
12 - 14	6.93	-69 (117)	6.31	-154 (183)
14 - 16	5.55	-83 (143)	5.21	-175 (223)
>16	25.44	-138 (362)	31.22	-278 (524)
0 - 10	53.31	-56 (62)	49.23	-110 (87)
>10	46.68	-106 (281)	50.76	-229 (431)

Table 1: Fare Discrepancies for Inefficient Routes

The GPS traces from the taxis include an instantaneous speed value, but manual inspection found almost all those reported speed values to be zero. So instead, we extract consecutive location coordinates of a particular taxi from the Log data trace, compute the geodesic distance d between the two coordinates, note the difference t in timestamps at the two coordinates, and estimate the vehicle's speed between those two coordinates as $v = d/t$.

The CDF of all computed speeds, between 30 km/h and 160 km/h, for a subset of 500 taxis on Oct 8, 2012 is shown in Figure 12. Each curve represents the CDF of the speeds computed for an individual taxi.

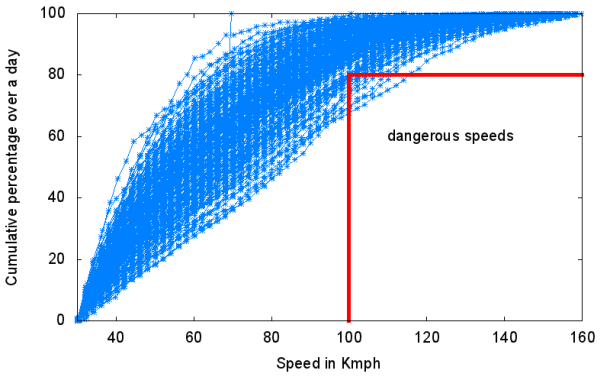

Figure 12: Cumulative percentage of speeds in km/h for different taxis on Oct 8, 2012

The red box shows the point at which taxi speeds are at or greater than 100 km/h. This is significantly faster than the maximum speed limit on any public road and is thus considered as "dangerous". In particular, we observe that there are a few taxis that spent 40% or more of that day traveling at these high speeds.

Using a similar technique to the one described above, we next computed per-day per-taxi speed CDFs for the entire month of Oct 2012. We then computed the number of unique days each taxi was

found to be "dangerously speeding" (at least once in that day) and divide that by the total number of days in the month (31 for Oct). Figure 13 shows the results for the top offenders in Oct 2012. For example, *taxi21* forms the rightmost bar in the graph, with 9 days of dangerous speeding (0.29 fraction of the 31 days in the month).

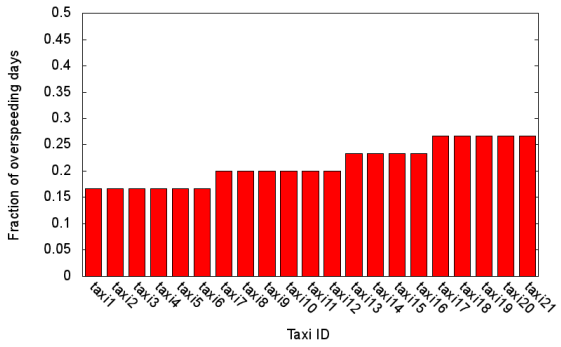

Figure 13: Fraction of days in Oct, 2012 during which excessive speeding was detected for these taxis

Figure 14 shows the top offenders when we include both Nov and Dec 2012 to the analysis. Note: we can also compute the fraction of hours or even minutes each taxi spends at a dangerously high speed. However, we decided to use fraction of days as that was easier for conveying the idea behind this method.

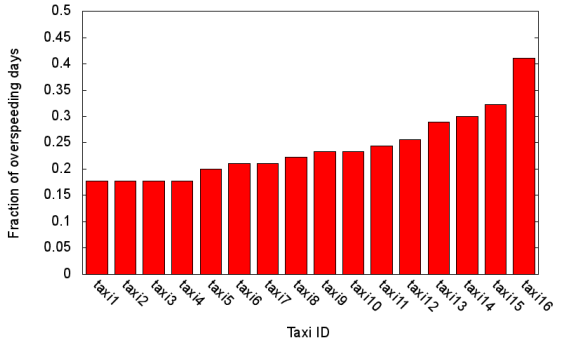

Figure 14: Fraction of days over Oct, Nov and Dec, 2012 during which excessive speeding was detected for these taxis

The analysis provided here is just a starting point for any system. Various reasons such as GPS errors etc., can temporarily cause a taxi's speed to appear much larger than it actually is. However, if the same taxi keeps re-appearing in the top offenders list, then it really might be speeding. In addition, our current method only detects extremely aggressive speeding that is likely to only occur on the major expressways in Singapore. It is quite possible for drivers to exceed the speed limit excessively on minor roads (that have a 40 or 50 km/h speed limit) without exceeding 100 km/h. As such, we plan to integrate a reverse geo-coding system to determine the actual speed limit of the road segment being traversed by the taxi.

6. DISCUSSION & CONCLUSION

Increasing operational efficiency of a transport company, mining GPS traces of their fleet, might apparently seem mundane. But as we will see next, several interesting research problems, involving both theoretical data analysis and experimental system building, can be crafted in this domain. In fact, the most interesting aspects of the analyses done in this paper are not the results presented, but the deeper research questions that can be designed from them.

For example, we observed a mismatch in demand versus supply for this taxi fleet at certain locations and periods of the day. When the occupancy rate is low, you end up with dissatisfied drivers while a high occupancy rate angers passengers who are unable to find a free taxi.

Some interesting research questions that arise due to this demand-supply tension include are free taxis not properly distributed in the geographical areas where there is potential demand for them? Is this distribution inefficiency causing long waiting delays for passengers as there are no taxis available near them? Is the distribution inefficiency also causing low occupancy for drivers, as there are no passengers available near them? Can we somehow predict demand for taxis in different geographical zones and route free taxis appropriately there? How can demand be identified, without requiring any input from passengers? Can taxi logs be mined to infer demand? Will historical information or information from a city events calendar help in demand estimation?

Even assuming that the demand estimation problem gets solved, and a real-time city demand map is built, what routing strategy should be followed to distribute the free taxis to the demand zones? In particular, what metrics should we use when determining a routing strategy? Should we minimise the average time the taxis spend without passengers or minimise the worst-case time without passengers? Or should we minimise the average or worst-case pickup distance? Maybe minimising the driving delay, considering traffic congestion delays, to the pickup point is a better metric? Experimenting with a combination of such metrics in a real setting will be interesting from both theoretical and experimental perspectives. Questions such as these and many more are currently being examined in collaboration with the fleet company.

7. ACKNOWLEDGEMENT

We would like to thank Darshan Santani and Prof. C. Jason Woodard for their experimentation and analysis efforts in the initial stage of the project. We are also grateful to Prof. Youngki Lee for his valuable comments on the paper. Last but not the least, we would like to thank the taxi company officials, without whose patient collaboration, this study would not have been feasible.

8. REFERENCES

[1] Speed limits on singapore roads. `http://www.onemotoring.com.sg/publish/onemotoring/en/on_the_roads/road_safety/speed_limits.html`.

[2] Rajesh Krishna Balan, Nguyen Xuan Khoa, and Jiang Lingxiao. Real-time trip information service for a large taxi fleet. In *Mobisys*, 2011.

[3] J. Biagioni, T. Gerlich, T. Merrifield, and J. Eriksson. Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones. In *SenSys*, 2011.

[4] Google Inc. *Google Maps API*, 2009.

[5] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie. Discovering spatio-temporal causal interactions in traffic data streams. In *SIGKDD*, 2011.

[6] Singapore Taxi Fare Chart. 2008. `http://www.cdgtaxi.com.sg/commuters_services_rates.mvn?cid=57870`.

[7] Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *UbiComp*, 2011.