Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems | School of Information Systems

11-2013

Challenges and Opportunities in Taxi Fleet Anomaly Detection

Rijurekha SEN
Singapore Management University

Rajesh Krishna BALAN
Singapore Management University, rajesh@smu.edu.sg

Citation

SEN, Rijurekha and BALAN, Rajesh Krishna. Challenges and Opportunities in Taxi Fleet Anomaly Detection. (2013). *SENSEMINE'13: Proceedings of the 1st International Workshop on Sensing and Big Data Mining (SenseMine)*. 1-6. Research Collection School Of Information Systems.
Available at: http://ink.library.smu.edu.sg/sis_research/2241

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, logistics companies 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. These logistics 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 logistics companies themselves are important to attract and engage these companies for collaborative data analysis.

In this paper, we show how GPS traces from taxis can be used to answer three different questions that are of great interest to a taxi operator. These questions are 1) What is the occupancy rate of the taxi fleet?, 2) What is the effect of route selection on the distance and time of a chosen route?, and 3) Does an analysis of travel times show deviations from the posted speed limits? We provide answers to each of these questions using a 2 month dataset of taxi records collected from over 10,000 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 work is supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 under the research grant T208B2301. 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, National University of Singapore, or Singapore Management University.

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 a taxi fleet (comprising 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 [6] 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 logistics companies in Singapore that operate fleets of vehicles has convinced us that identifying questions of *mutual* interest that have both exciting research potential and important operational consequences for the 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 company to determine if its fleet is fully utilised and / or sufficient. 2)

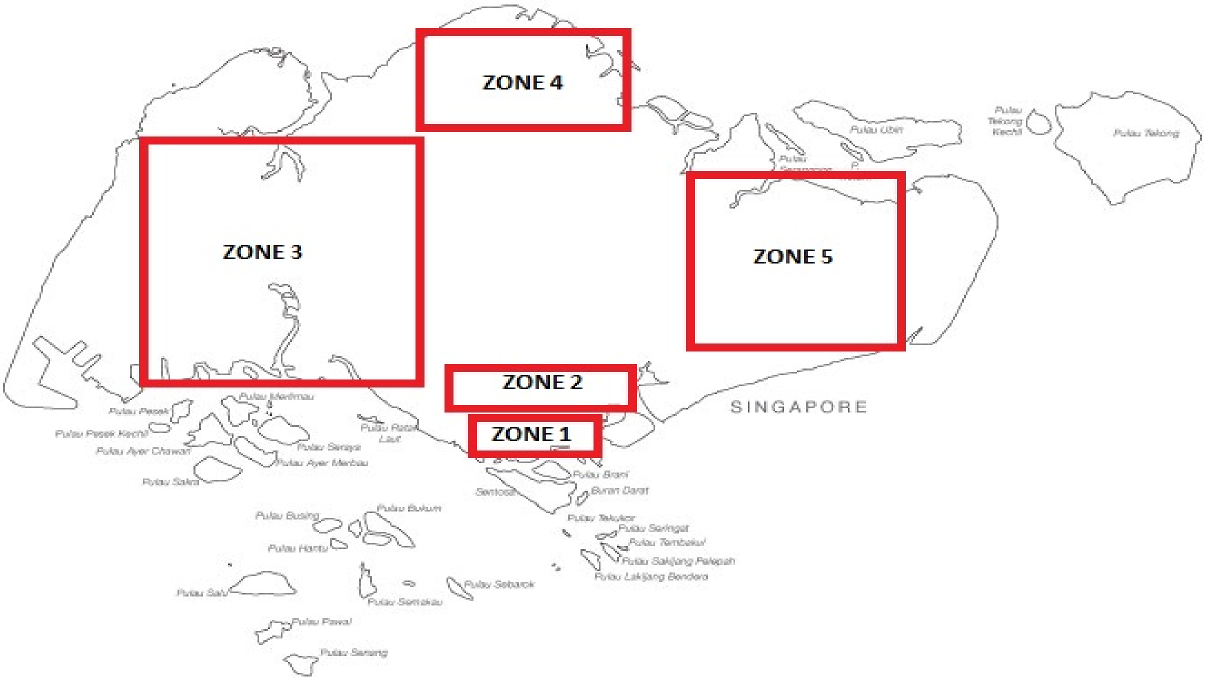

Figure 1: Rectangular zones under observation

What is the effect of route selection when deciding how to reach a passenger's desired destination? In particular, are the routes selected anomalous with respect to distance and time? 3) What is the correlation between trips times and posted road speed limits? In particular, are the speed limits being ignored by a large number of drivers?

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 taxis 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 AND 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.

For this paper, we used two months of GPS-enhanced data from over 10,000 taxis operating in Singapore. In particular, we use the following two types of data, one or both of which are necessary for exploring the three questions in this paper.

a) Log Data: This data contain records about the instantaneous state of a particular taxi. This data 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 data is also used in our analysis of travel times (Section 5), to compute how long it took a particular taxi to cover a certain distance.

Each log data record 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 many states at any point in time. However, for the purpose of this work, we only use the following 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), *ONCALL* (taxi is on the way to service a special booking call), and *POB* (passenger on board – taxi is taking a metered customer to a destination).

b) Trips Data: This data 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 travelled during the trip. This data is used to detect anomalous trips in our route selection 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 inefficiencies are important for transportation companies 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 labelled 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 utilisation 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*) in the given zone and hour, and $A =$ the number of minutes it was available (in state *FREE*). 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 activities 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.

  In addition, the number of reporting taxis in zone 2 ((Fig. 4) is significantly higher than in zone 1 ((Fig. 2), even though the 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 the 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 the supply of available taxis does not appear to match the demand shown by occupancy levels. However, further analysis is necessary to identify the root causes of these possible mismatches and to develop an appropriate solution.

- 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 most buses and train services do not run in the early hours of the day (between 1 a.m. to about 5.30 a.m.), 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. EFFECT OF ROUTE CHOICE

In this section, we identify the effect of taxi route selection on two different efficiency metrics; We start with distance – a route between two end points is flagged as anomalous if it is significantly longer than another route between those same end points. We then expand our analysis to understand whether those distance-anomalous routes are actually efficient for one other metric; namely speed.

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]). We then identify the potentially anomalous trips that have the highest discrepancies between the actual distance and time taken, and the Google Maps-reported estimated distance and time.

With the Google Maps data, it is possible to calculate differences between the actual trip data and the Google Maps suggested values. However, what thresholds should be used to determine inefficiencies? To answer this, we compared the differences obtained at different thresholds. These results are shown in Table 1. The table provides results for the distance (trip took a longer route than it should have) and time (trip took more time than it should have). For each comparison point, we looked at different thresholds ranging from greater than -50% to greater than -5% – i.e., Google Maps values were at least 50% lower than the corresponding trip value to Google Maps values were at least 5% lower than the corresponding trip value.

From the table, we see that at the 20% threshold (-20) (Google Maps values are at least 20% lower than the actual trip value), 6.37% of the trips were longer than expected while 43.39% of the trips took longer than expected. When the threshold was raised to 50% (-50), the anomalous trip percentage dropped to 2.30%, and 14.98% with respect to distance and time (in that order).

Next we attempted to validate if longer routes are faster, i.e. due to road conditions etc., the longer routes turn out to be faster in practice. We extract all the trips (at both -20% and -50% thresholds) that were anomalous 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 predicted than Google.

However, it should be noted that Google's estimates for time and distance (at least for Singapore at the time we did this study) do not accurately consider factors such as traffic conditions, road pricing (at certain times of the day, using specific roads can result in additional charges which passengers might want to avoid), accidents etc. Hence, this analysis is just a starting point for a deeper study into the effects of route selection.

5. ANALYSIS OF TRAVEL TIMES

In this section, we investigate the issue of travel times compared to the posted speed limits. This is an important consideration for transport network operators as exceeding specified speed rules can increase fuel consumption and related costs.

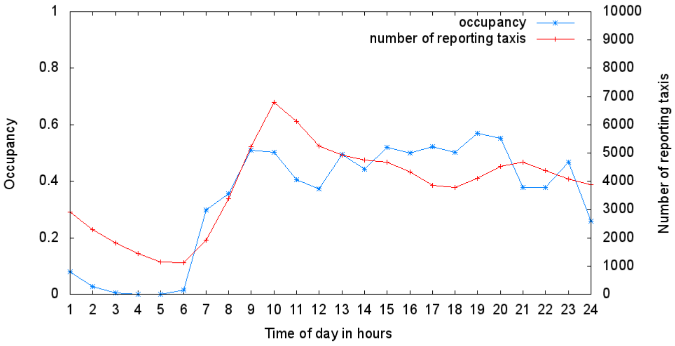

Figure 2: Zone 1 on weekdays in Oct, 2012

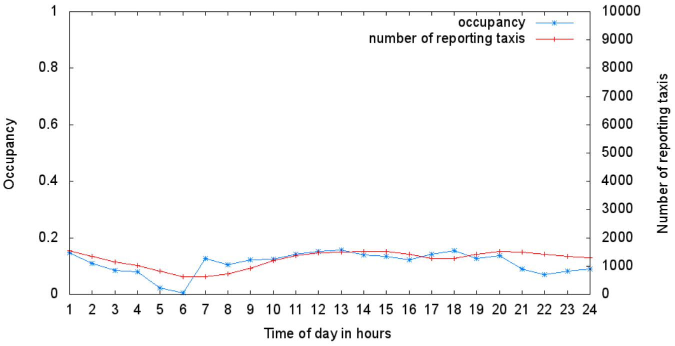

Figure 3: Zone 1 on weekends in Oct, 2012

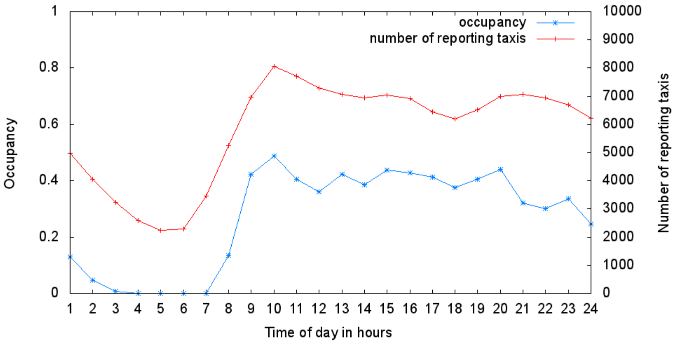

Figure 4: Zone 2 on weekdays in Oct, 2012

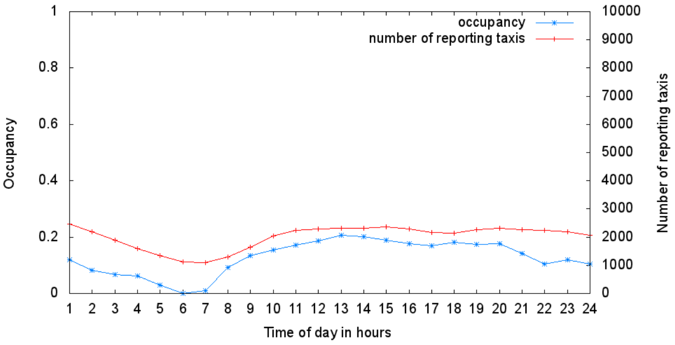

Figure 5: Zone 2 on weekends in Oct, 2012

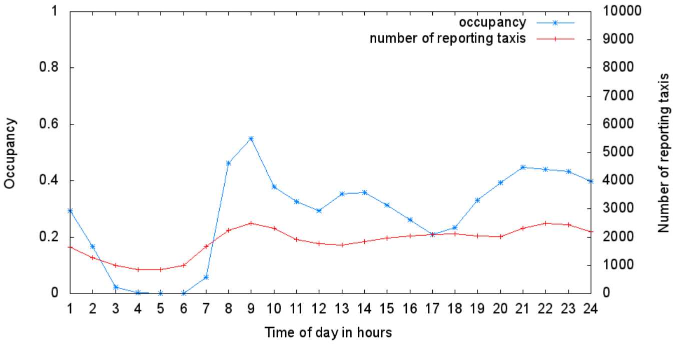

Figure 6: Zone 3 on weekdays in Oct, 2012

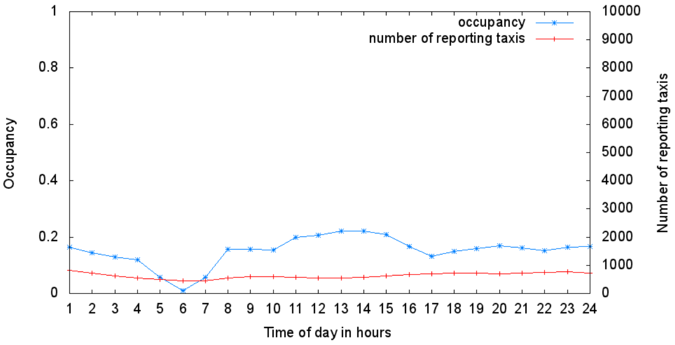

Figure 7: Zone 3 on weekends in Oct, 2012

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.

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 120 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.

The red box shows the point at which taxi speeds are at or greater than 100 km/h. This is faster than the maximum speed limit on any public road and is thus considered as “high”. In particular, we observe that there are a few taxis that spent 40% or more of that day travelling 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 in the “high” speed range (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 some taxis in Oct 2012. For

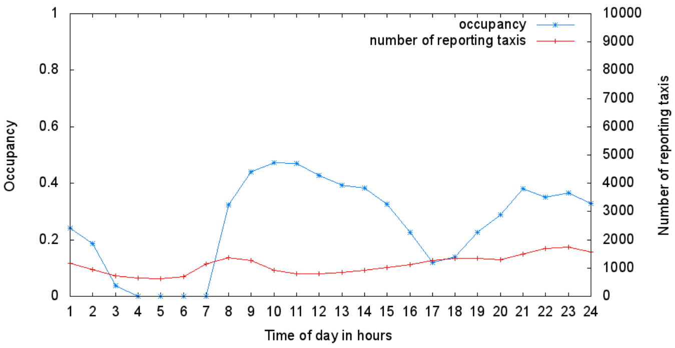

Figure 8: Zone 4 on weekdays in Oct, 2012

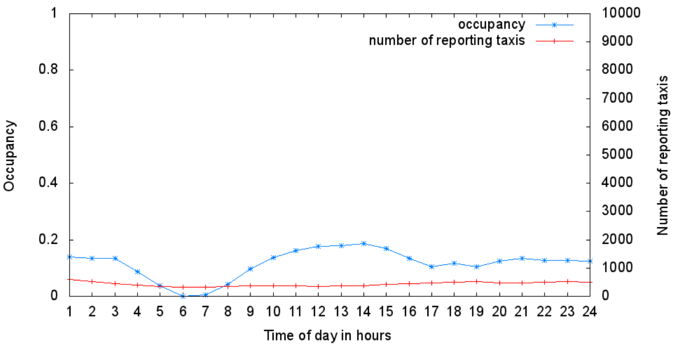

Figure 9: Zone 4 on weekends in Oct, 2012

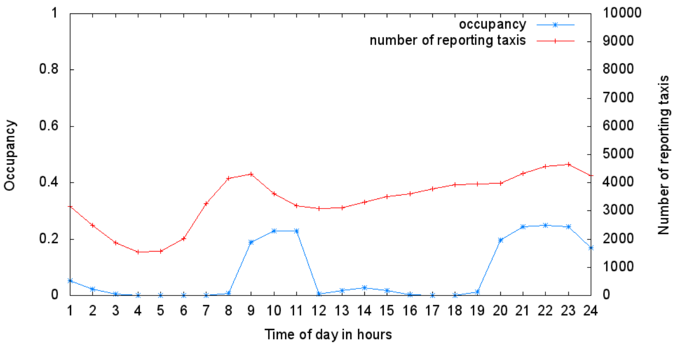

Figure 10: Zone 5 on weekdays in Oct, 2012

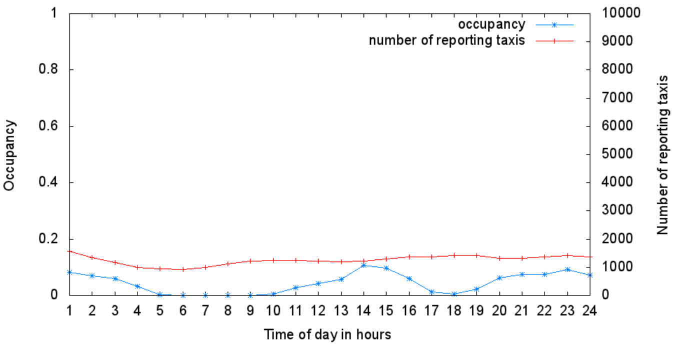

Figure 11: Zone 5 on weekends in Oct, 2012

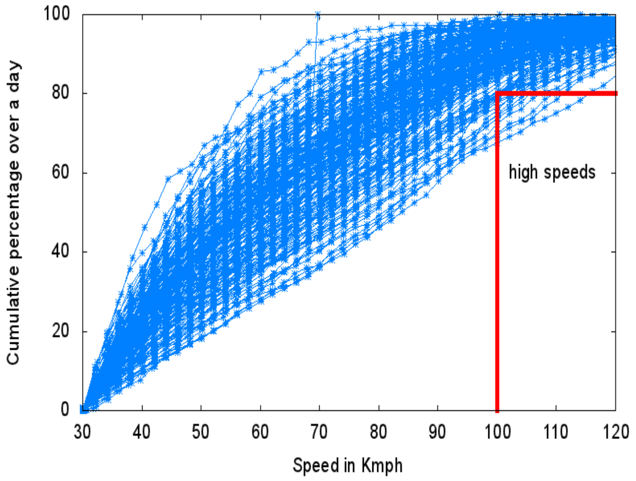

Figure 12: Cumulative percentage of speeds in km/h for different taxis on Oct 8, 2012

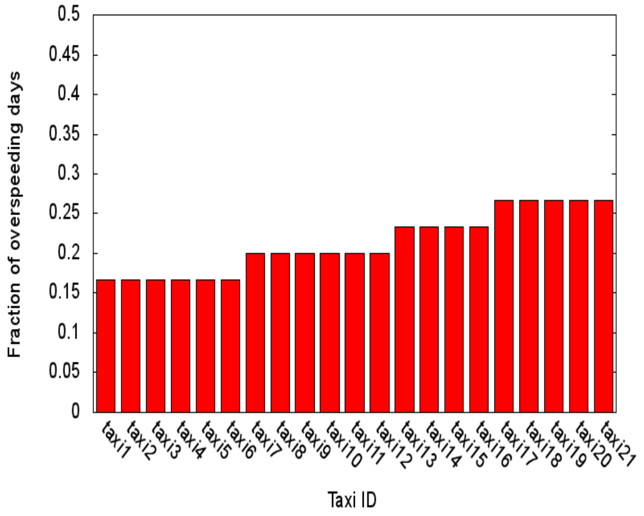

Figure 13: Fraction of days in Oct, 2012 during which high speeds were detected for these taxis

example, *taxi21* forms the rightmost bar in the graph, with 9 days of high speeds (0.29 fraction of the 31 days in the month).

Figure 14 shows similar results as Figure 13 except that we now look at a larger range, Oct, Nov, and Dec 2012, when computing the analysis. Note: we can also compute the fraction of hours or even minutes each taxi spends at a high speed. However, we decided to use fraction of days as that was easier for conveying the idea behind this method.

The analysis provided here is just a starting point for any system. Various reasons such as GPS errors etc., can temporarily cause a

Metric	Percentage Difference Between Google Maps and Trip Data ($\frac{GoogleMaps-TripData}{GoogleMaps} * 100$)									
	>-50	>-45	>-40	>-35	>-30	>-25	>-20	>-15	>-10	>-5
Distance	2.30	2.66	3.04	3.58	4.25	5.13	6.37	8.14	10.59	14.35
Time	14.98	18.87	20.90	24.43	28.37	31.67	36.82	43.39	49.14	55.19

Each entry shows the percentage of the trip records that matched that criteria. Thresholds are negative because these results are for those trips where the Google Maps value is smaller than the Trips data value.

Table 1: Picking the Right Cutoffs for Anomalous Routes

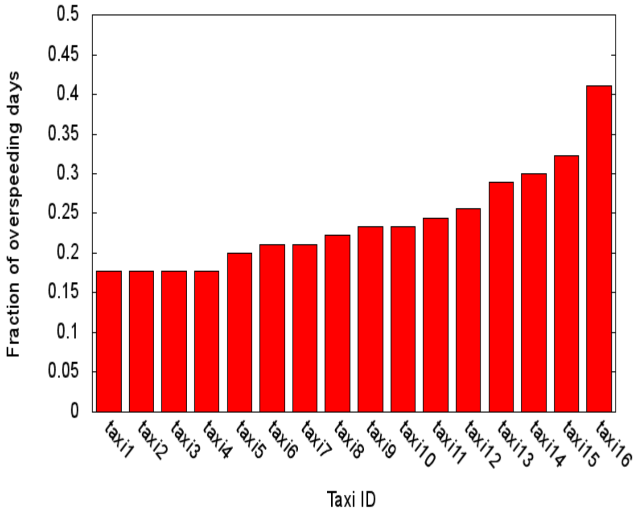

Figure 14: Fraction of days over Oct, Nov and Dec, 2012 during which high speeds were detected for these taxis

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 is probably a real anomaly. In addition, our current method only detects seeds over 100 km/h which are most likely to occur only on the major expressways in Singapore. This analysis is not able to detect speed anomalies that occur on minor roads (that have a 40 or 50 km/h speed limit) and do not exceed 100 km/h. In the future, we hope to integrate a dynamic reverse geo-coding system that will automatically 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 fleet companies.

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.

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