Unsupervised Detection of Drivers' Behavior Patterns

Zhang (Gabriel) Li, Chen Cai

ATP Research Laboratory, NICTA, 13 Garden Street Eveleigh NSW 2015 Australia

Gabriel.Li@nicta.com.au

Abstract

Probes with GPS devices reveal useful information for traffic conditions and patterns. The high level of noises and lack of details make it challenging to mine behavioral patterns from the raw data collected. Behavioral patterns are essential for understanding the underlying structures of data sources and various real-world interests such as traffic planning, vehicle operations and anomalous/popular area detections. This paper proposes an unsupervised approach for mining behavioral patterns from naïve GPS data any devices can collect. The focus on the study is to apply the method for Taxi Drivers' behavioral analysis, which is essentially different from other road users. Unsupervised clustering algorithm successfully detects the cohesion of points in the 3-D space. Through comparison with other more fine-grained data set, the result is evaluated and justified for the key roads across Sydney areas.

1. Introduction.

Behavioral patterns have its significant roles in traffic related analytics. Vehicles-specific behaviors can be detected and mined to analyze spatial or temporal travel patterns and location/trajectory popularity. The analysis is capable to provide the theoretical foundations for modern traffic services, such anomalous traffic patterns detections or travel planning and route recommendations.

GPS devices, with increasing coverage and popularity nowadays, are widely used to collect traffic data. However, they are essentially suffering two problems:
1. High level of electronic noises due to radioactive decay among barriers.
2. Heterogeneous data sources from types of vehicles of various underlying patterns, which are required to be treated respectively.

Consequently, it is challenging but of great significance to have a robust and generic approach to mine behavioral patterns from the GPS data with increasing popular mass.

1.1 Background.

Behavioral pattern analysis using GPS data has received many attentions. Patterson et al. (2003) presented a method of learning Bayesian model for transportation mode as well as route preference. Zheng et al. (2008) proposed a supervised learning approach to infer people's motion modes from their GPS logs, collected by 65 people over a period of 10 months. By extracting a set of sophisticated features from GPS data, such as head turning rate and stop rate, the methods can predict the human travel mode at around 76% correct rate. Liu et al. (2010) developed an unsupervised approach to classify cabdrivers as top ones and ordinary ones based on a set of derived features.

Pang et al. (2012) used GPS data from taxis to monitor the emergence of unexpected behavior in the Beijing metropolitan area. Spatial-temporal outlier model is derived from likelihood ratio test statistic (LRT), which is first applied in the traffic domain for anomalous behavior patterns. Cao et al. (2010) presented techniques capable of extracting semantic

locations from GPS data. Through random walk on the graphs propagates significance among the locations, significance of locations are exploited such as the number of visits to the locations, the durations of the visits and the distances to reach locations.

1.2 Objective and contributions.

All of the models above require either labeled GPS data or fine-tuned GPS data provided by voluntary individuals or commercial companies. However, the requirement is hard to be satisfied in the real-world traffic networks. The distribution of GPS device carriers (e.g. taxis, courier trucks and private cars) is uncontrolled and tends to be sparse in observation and biased in vehicle types.

On the other hand, raw GPS with only basic information is extensively collected not only from on-boards devices but also from mobile phone of individuals thanks to the mobile century. Our aim is to discover some potential approach could be applied to raw GPS data in unsupervised manners.

The main contributions of this work are as following:
1. A generic unsupervised approach is applied to raw GPS data, which can be collected by any kinds of coarse-grained GPS devices.
2. A set of transformed features is derived for successful separation of taxi drivers' behaviors without any prior knowledge in terms of how they operate.

For example, in our user case, it provides us with the data usability estimation for travel time estimation. Different GPS data sources are inevitably fused together to provide more reliable traffic condition overview. It is significant to have a generic tool to analyze how the data sets are comprised of and whether they are unrealistically biased.

2. Data Resources.

2.1 Data Description.

In this paper, the data is collected from GPS devices on NSW probes biased by taxis, which records timestamps, coordinates in longitude and latitude format, bearings along with vehicle identifier and vehicle spot speed. The data is raw and has no road network information. The GPS points are mapped to Sydney traffic networks in NSW region. Road identifiers can be determined by projecting the points onto nearest links of NSW key roads. The travelling directions of vehicles are computed by vehicle bearings (angle measured in degrees in a clockwise direction from the north line).

GPS data is historically suffering from electrical noises, especially for largely urbanized city like Sydney. However, the substantial amount of observations (Table 1.) makes it possible to mitigate the impact of outliers. Besides, some customized pre-processing will be carried out to further reduce level of noises. GPS coverage are the main roads concerned in the Sydney area (Fig 1.), which are imperative roads connecting significant transportation hubs. Additional GPS data sources are introduced for comparison purpose, to justify the effectiveness of our approach.

GPS Data Observation	Number of Vehicles	Number of Records
Peak (Mon to Sat)	3,900 – 4,350	150,000 – 215,000
Off-Peak (Sun)	3,500 – 3,600	100,000 – 110,000

Table 1. Observations of GPS Data.

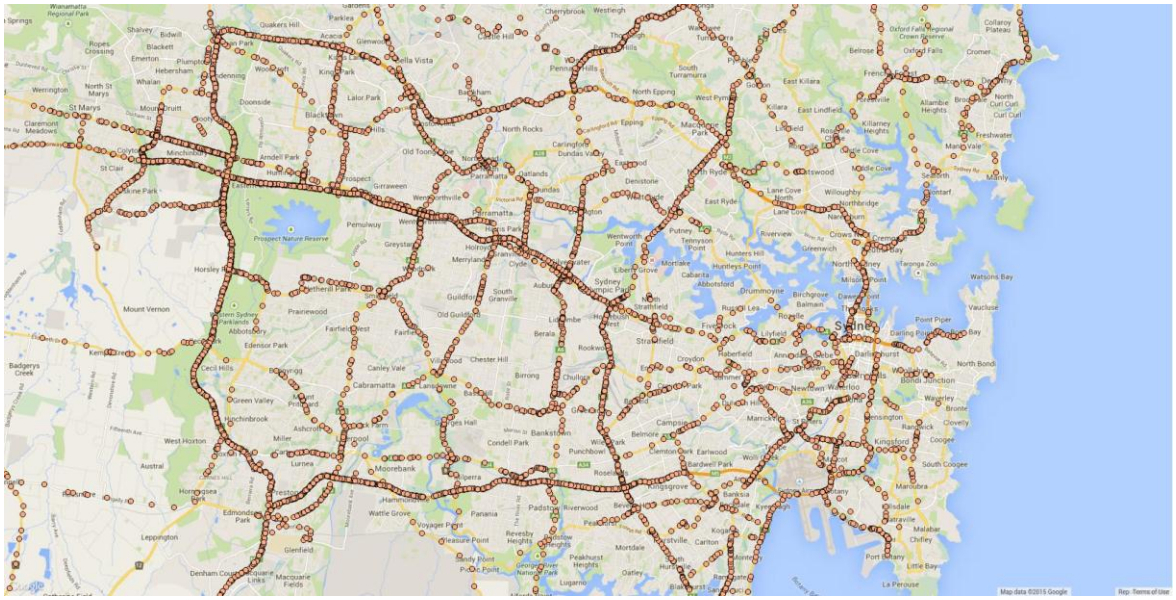


Fig 1. Coverage of GPS Data.

2.2 Data Pre-processing – Granularity Deduction.

A counter-intuitive approach is applied to improve the effectiveness of classification. As a generic approach for GPS data, it is unpractical to expect the coordinates and spot speeds to be reliably measured by various types of GPS devices in severe noises and interferences of modern urban environment. Instead of seeking for fine-grained data, we transform the data to lower granularities and are more representative of the vehicle behavioral characteristics. The transformation is done for input features as followings.

2.2.1 Spot Speed.

The accuracy of GPS spot speed is highly depending on the sampling time frames of devices, which are usually only short enough to reflect a snapshot of vehicle movements. Besides, as the time frames are short, small noises may result in a large discrepancy in the final estimation of spot speed. However, behavior patterns become more obvious and reliable when long time frames are taken into account. Data of short time frames is too sensitive and subjected to outliers. Spot speed is re-calculated using adjacent GPS observations by the equation below:

$$S_n = \frac{link\_distance(p_{n+1}, p_n)}{t_{n+1} - t_n},\ for\ S_n < \varepsilon\ and\ t_{n+1} - t_n > \zeta \qquad (1)$$

Where *link_distance* is the distance vehicles travelled between two adjacent positions; p_{n+1} and p_n are locations of two successive observations for the same vehicle along the same direction within the day; t_{n+1} and t_n are corresponding sample time for the above observations. Speed levels that are much higher than the speed limit are either rare or due to too short time windows, they are considered to be outliers and filtered out.

2.2.2 Probe Location.

GPS coordinates have been projected to links of Sydney key roads. Then probes' positions can be represented and identified by the locations of road links. As locations of interest only appear on the link levels, such as shopping malls, universities and train stations.

2.2.3 Time Stamp.

Though our data size is substantial, GPS data source may not have enough observations during off-peak hour or on less important road segments. Also, traffic patterns will often vary for relatively longer time frames. Therefore, for our cases, the time stamps are further mapped into hourly window index, to provide more robust and representative features of observations.

3. Methodology.

3.1 Unique Behaviors Patterns.

To illustrate how the data set is of inherently unique characters. Various data sources are selected for comparison on Parramatta Road:
(1) Our data dominated by taxi probes.
(2) GPS data dominated by courier truck probes.
(3) SCATS data has no discrimination of all types of vehicles.

Plotted and compared in Fig 2. is the average harmonic mean speed for each 15-minute window of the day.

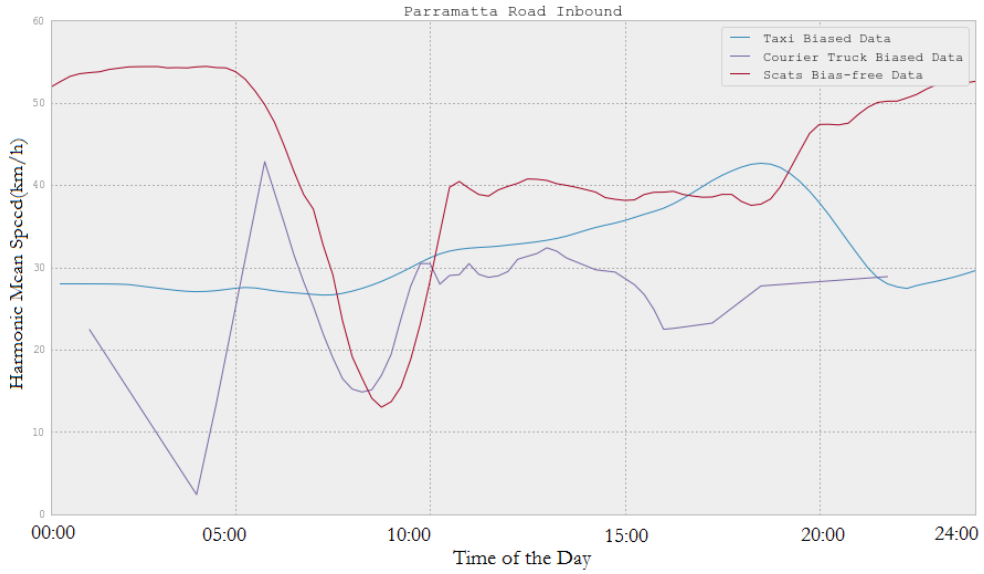


Fig 2. Harmonic mean speed comparisons of Tues 2014. (Gaps means no observations)

Seen from Fig 2., some inconsistencies are found for daily speed pattern:
P1. During peak hours, taxis have higher speed than courier trucks.
P2. Taxis have a tendency to travel slower than the average during off-peak hours.

Below is a box plot within the day about speed distribution. The six plots represent 4 hours time windows of the day, and the lines in the box indicate the mean speed. In addition, the density of probes is plotted as the blue line. There is an obvious trend that density is

negatively related to the mean speed, which is intuitive and understandable. However, patterns P1 and P2 are not consistent with this negative correlation.

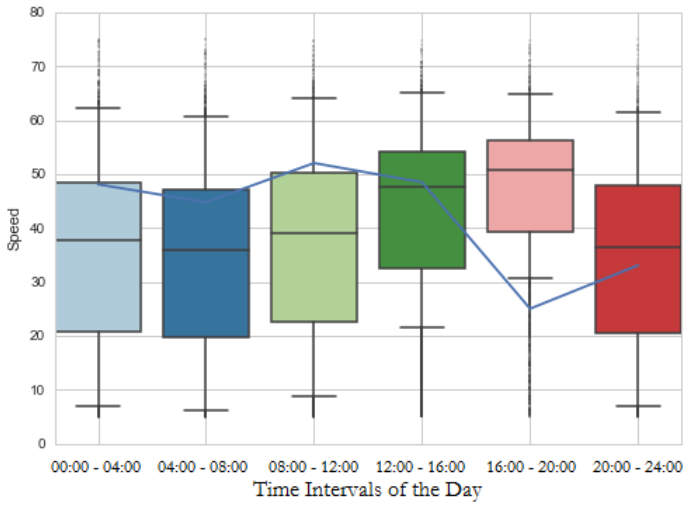


Fig 3. Probe Density and Speed Distribution.

3.2 Unsupervised behavioral classification.

In order to reveal the answers, k-means algorithm is applied to obtain clusters with unique driving characteristics. GPS data, after pre-processed as above, are fitted into the model with the set of derived features (l, t, s), where l is the link location, t is the 15-minute time window index of the day, and s is the adjusted spot speed by adjacent observations.

Afterwards, cluster number N was optimized for the largest Silhouette coefficient (Rousseeuw, P. J. 1987), a measurement for inter-cluster dissimilarity. Euclidean distance is chosen as the similarity measurement between individual GPS observations. Algorithm has been run 100 times by randomly assigning the initial centroids and the best result is selected to avoid local optima.

4. Clustering Result.

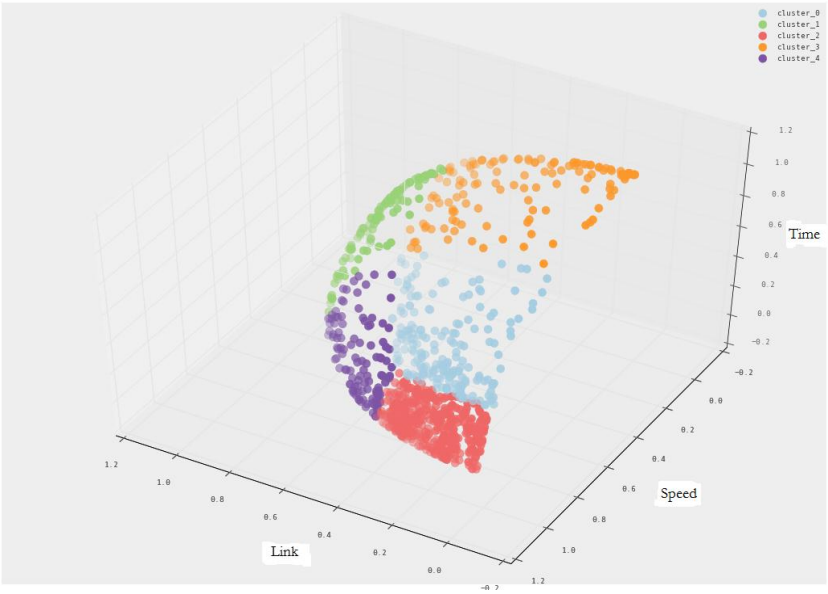


Fig 4. Clusters Obtained in 3-D Space (Scales normalized).

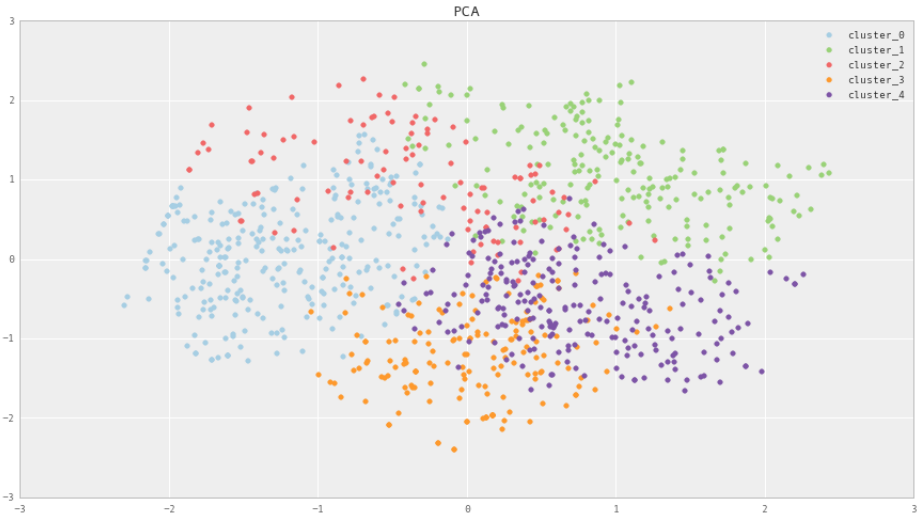
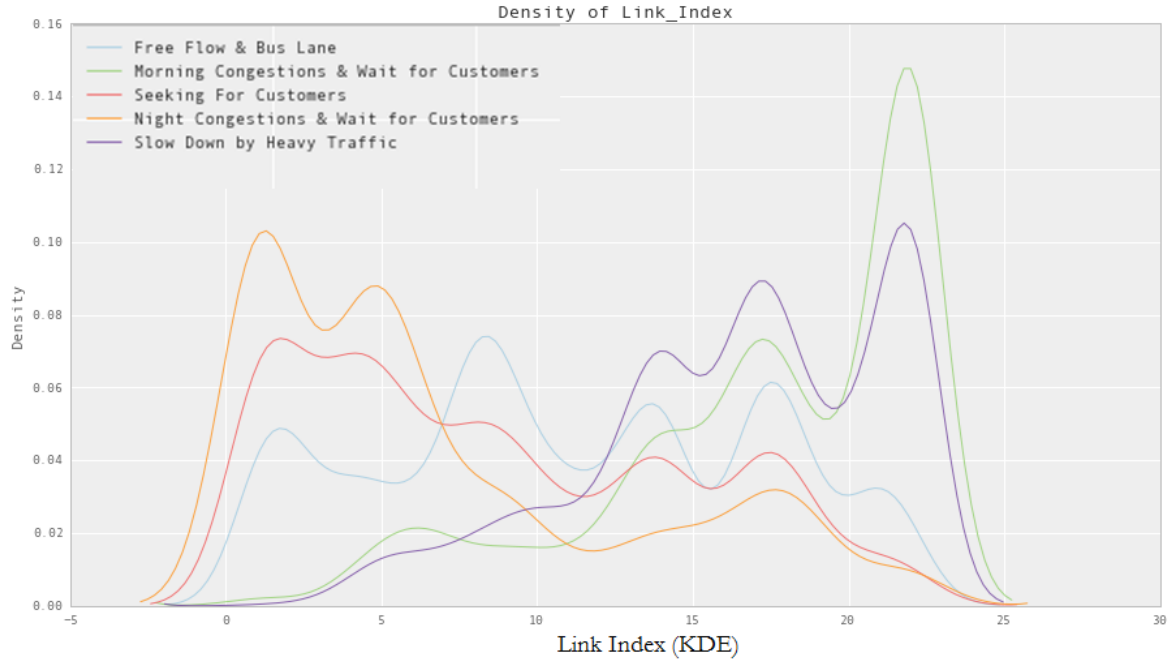
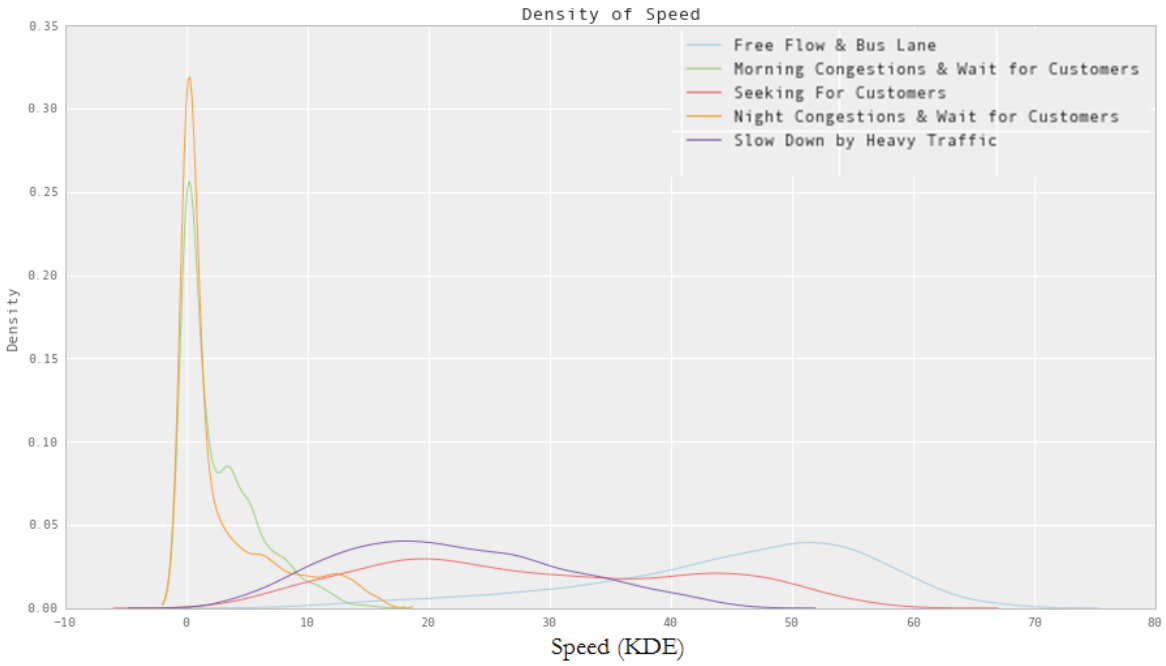


Fig 5. PCA Visualization of Clusters (Scale standardized).

According to the k-means algorithm, five clusters are obtained as shown in Fig 4. Fig 5 is the 2-D visualization after dimension reduction using Principle Component Analysis (PCA). Clearly, patterns are identified in the spatial-temporal spaces. Each cluster possesses unique properties and will be illustrated below:



(a) Link Location Density (left - downtown, right - uptown)



(b) Spot Speed Density

(c) Hourly Time Window Density

Fig 6. Kernel Density Estimation of Cluster Features.

(1) Free Flow & Bus Lane Borrowing: evenly distributed across all the links, travelling very close to the speed limits, and most of the observations are in the morning. (Pattern P1 in 3.1)

This pattern illustrates when the probes are travelling fast, usually during the off-peak hours. However, pattern P1 mentioned in section 3.1 shows that taxis have higher speed during the peak hours, corresponding to the shallow dip in the curve for the group. The reason is that taxis with passengers can travel along the bus lanes, which will alleviate the impact of heavy traffic flows.

(2) Morning Congestions & Waiting for Customers: concentrates on the uptown sections of roads, travelling at a slow speed below 10 km/h and majority of observations are in the morning.

This pattern has its counter-intuitive parts: why there are more slow probes before morning peak hours? Seen from the link location densities, lots of taxis are waiting for customers heading to the CBD area at uptown sections of roads. Undoubtedly, profit can be maximized through this way during the early morning hours by seeking customers along the inbound direction of roads.

(3) Slowed down by heavy traffic: concentrates on the uptown sections of roads, travelling at slow speed 10 - 30 km/h, and majority of observations are in the morning.

This is the "medium" speed group in the morning hours. They are neither so slow as congestions, nor close to the free flow speed. Seen from the speed density, even after isolated the congestions, most of the observations still lie in this group, which reveal the slow speed probes are dominated during morning hours.

(4) Night Congestions & Waiting for Customers: concentrates on the downtown sections of roads, travelling at slow speed below 10 km/h, majority of observations are after noon.

This is the counterpart of group (2). Notice that congestions are always grouped with behavior of waiting for customers. Since there is no lane formation in the GPS data, taxi probes waiting customers on the road side cannot be distinguished from taxi probes caught in the traffic jams. Lane separation definitely can provide us with more details, which is not in the topic of this paper.

(5) Seeking for Customers: mainly at the downtown sections of roads (CBD bound), travelling at medium speed 20 - 40km/h, surprisingly in the 8PM - 12AM periods. (Pattern P2 in 3.1)

This is the most interesting group, why probes do not travel close to speed limit when there is only light flow on the road? The locations indicate that the taxi probes in this group gather around places of interests with large customer bases, such as the central stations located at the CBD bound. This can be explained by that during midnight, people are usually active around CBD areas for nightlife, like bars, shopping malls and performance centers.

Comparing to peak hours, the number of customers decreases during off-peak hours. Amount of "Empty Drive" taxi probes will correspondingly increase. Therefore, slowing down around places of interests is a common strategy for taxi drivers to find more potential customers, which often happens at "night life" hours. Fig 7 demonstrates, though group 3 and group 5 both suffer from the slow speed, but the locations of group 5 are quite different compared to normal congestions. Not only they are sometimes in different spots, the empty drive behavior will not cause long tails since only certain lanes are occupied.

(a) GPS samples in the "Empty Drive" cluster

(b) GPS samples in the "Slowed by Heavy Traffic" cluster

Fig 7. Comparison between "Empty Drive" and "Slowed by Heavy Traffic" cluster. High temperature area indicates higher density of probe vehicles.

5. Conclusion and Future Work.

As above, we have demonstrated the approach of unsupervised classification for drivers' behavioral patterns in raw GPS data. A set of features is derived from basic information provided by coarse-grained GPS devices. The method is generic and can be applied to any GPS source as desired.

After identified the characters of behavioral groups in traffic data, several potential directions are open for further exploration. For example, how we fuse the information collected from heterogeneous sources together as data sets of synthesized components of behaviors, so they can serve different purposes of research or industrial applications. Also, the behavioral patterns can provide insights for taxi operators in terms of revenue optimization and planning.

References

Patterson, D. J., Liao, L., Fox, D., & Kautz, H. (2003, January). Inferring high-level behavior from low-level sensors. In *UbiComp 2003: Ubiquitous Computing*(pp. 73-89). Springer Berlin Heidelberg.

Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. Y. (2008, September). Understanding mobility based on GPS data. In Proceedings of the 10th international conference on Ubiquitous computing (pp. 312-321). ACM.

Liu, L., Andris, C., & Ratti, C. (2010). Uncovering cabdrivers' behavior patterns from their digital traces. Computers, Environment and Urban Systems, 34(6), 541-548.

Pang, L. X., Chawla, S., Liu, W., & Zheng, Y. (2013). On detection of emerging anomalous traffic patterns using GPS data. Data & Knowledge Engineering, 87, 357-373.

Cao, X., Cong, G., & Jensen, C. S. (2010). Mining significant semantic locations from GPS data. Proceedings of the VLDB Endowment, 3(1-2), 1009-1020.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.