Surface Street Traffic Estimation

Jungkeun Yoon   Brian Noble   Mingyan Liu

Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2122
{jkyoon,bnoble,mingyan}@eecs.umich.edu

ABSTRACT

In this paper, we propose a simple yet effective method of identifying traffic conditions on surface streets given location traces collected from on-road vehicles—this requires only GPS location data, plus infrequent low-bandwidth cellular updates. Unlike other systems, which simply display vehicle speeds on the road, our system characterizes unique traffic patterns on each road segment and identifies unusual traffic states on a segment-by-segment basis. We developed and evaluated the system by applying it to two sets of location traces. Evaluation results show that higher than 90% accuracy in characterization can be achieved after ten or more *traversals* are collected on a given road segment. We also show that traffic patterns on a road are very consistent over time, provided that the underlying road conditions do not change. This allows us to use a longer history in identifying traffic conditions with higher accuracy.

Categories and Subject Descriptors

H.2 [**DATABASE MANAGEMENT**]: Database Applications; H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Online Information Services

General Terms

Algorithms, Design, Measurement

Keywords

Traffic, Estimation, GPS

1. INTRODUCTION

Driving a vehicle is an essential part of our daily lives, but traffic congestion increasingly leads to significant delays. According to a recent report [19], US drivers wasted 3.7 billion hours of time, 2.3 billion gallons of fuel, and a total cost of 63 billion dollars in 2003 due to traffic delays. If it were possible to detect congestion and notify drivers in real time, many drivers might have been able to avoid congested routes before they join the backup, further exacerbating the problem.

One already has a crude approximation of this, with congestion detected by eyewitness reports from news organizations or operators themselves. These are augmented by closed-circuit cameras and sensors that have been installed as infrastructure. Unfortunately, the coverage of those systems is extremely limited due to high installation and maintenance costs. For example, a vehicle loop detector costs $700 for a loop, $2500 for a controller, $5000 for a controller cabinet, $300000 for fiber optic cable per mile, and 10% of the original installation cost for annual maintenance as of 1999 [2]. As an example of the limited coverage, New York City Department of Transportation provides only 22 traffic monitoring cameras for 8 million New Yorkers [4]. It is practically impossible to install traffic monitoring systems densely enough to cover the entire network of surface streets.

Instead, we propose a simple yet effective method to capture and identify traffic status without any aid of costly traffic monitoring systems. We leverage vehicular mobile networks such as GM's OnStar [5]. OnStar is an optional system with GPS, two-way radio, and sensors built in GM vehicles mainly for emergency and vehicle maintenance. Such a service could allow vehicles to transmit GPS data to service providers and receive useful traffic information from them in return. Moreover, many cellular phones, PDAs, and vehicle navigation systems already have GPS plus two-way data communication capabilities, including mobile WiMAX [6], HSDPA/UPA, or EV-DO/DV. In this paper we show how GPS data collected from these vehicles and devices can be used to provide personalized road traffic information in real time. We present an algorithm that can extract useful traffic information from such collective GPS traces, and show that even a modest fraction of vehicles so equipped can enable real-time traffic monitoring. Our real-time traffic monitoring system is shown in Figure 1.

Such traffic estimation is relatively straightforward on freeways; they have limited access, and traffic interruptions are the exception rather than the rule. Contrast this with surface streets, with traffic lights, stop signs, crosswalks, and similar traffic obstacles. In this setting, identifying troublesome traffic states is more complex. For instance, a low speed does not necessarily mean a bad traffic jam, because it may be caused by traffic signals, whereas the actual traffic flow is fine. Furthermore, all roads are not the same; they have different speed limits, numbers of lanes, road lengths, or road curvatures. We cope with such diversity
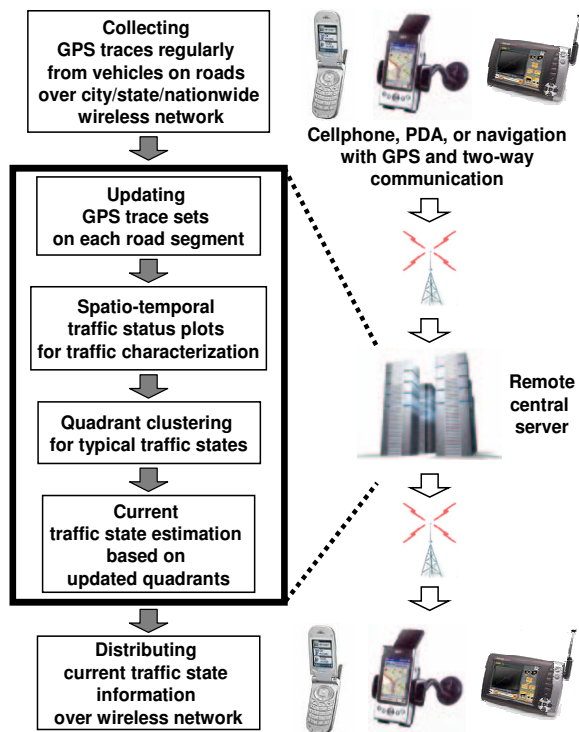
Figure 1: A block diagram of our real-time traffic monitoring system.

by partitioning roads into segments, and then characterize unique traffic patterns on each segment, considering vehicle speeds, traffic light durations, segment lengths, and traversal times—all extracted from GPS data. Based on the traffic characteristics on each road segment, we can identify current traffic states on the road segment using GPS data transmitted from vehicles in real time.

For evaluation, we applied our method to two different trace sets collected in the Ann Arbor area. Although we had to collect GPS data during a wide time window because of a limited number of GPS-equipped vehicles, evaluation results show that we can achieve higher than 90% accuracy after ten traversals on a road segment.

The remainder of this paper is organized as follows: Section 2 covers background and related work. Section 3 provides a brief overview of our system. Section 4 shows how to extract traffic information from GPS data, while Section 5 describes how to classify traffic states and identify current traffic status based on the traffic features extracted in Section 4. Section 6 evaluates our method by applying it to two different GPS trace sets, and Section 7 concludes this paper.

2. RELATED WORK

Vehicle traffic has been an important research issue ever since the invention of the automobile. Greenshields conducted empirical studies on the traffic flow theory in the 1930s, and a number of work has been done to analyze traffic in terms of speed, flow rate, density, and many other factors [11, 12, 15]. However, most work was focused on finding the empirical relationships between many traffic factors like speed and flow rate.

Since the idea of intelligent transportation system (ITS) was introduced a couple of decades ago, research focus started to move on to how to measure traffic remotely and detect incidents or delays automatically by using sensors built on roads. For example, Lin and Daganzo [16] and Coifman [10] proposed simple schemes to detect freeway incidents using traffic detectors on roads. But using detectors on roads has a limited coverage problem due to high installation and maintenance costs. So some started to pay attention to using vehicles equipped with GPS and/or a cellular positioning system as probes [18, 22, 17]. However, they just proposed how to measure vehicle speeds usually on simple freeways, whereas we focus on how to identify traffic states in complex urban areas rather than just how to estimate speeds directly from GPS traces.

Recently, some real-time traffic monitoring products have entered the marketplace. For example, IntelliOne [3] introduced a new real-time traffic monitoring service using mobile phone positioning systems. Unlike our approach, it simply reports vehicle speed on road segments—users are left to convert this to overall traffic conditions. On limited-access roads, this may be reasonable, but it is not effective for surface streets. As another example, Inrix [1], a Microsoft's spin-off, also introduced their new real-time traffic monitoring system using statistical analysis of a large variety of data inputs. These include sensor data from Department of Transportation and probe data from commercial fleets. However, they also just show vehicle speeds on roads using probe data, and require outside information—e.g., weather, incident, or construction reports—to identify traffic states. Instead, we capture and identify traffic states with only GPS traces in a much simpler way.

3. OVERVIEW

The goal of our work is to identify traffic states from GPS data. To achieve this goal, we proceed to address a number of fundamental questions.

(1) What kind of traffic information can be extracted from GPS data?

The information we can obtain directly from GPS data is very limited: user's coordinates, timestamp, and optional speed and heading. Although many traffic features may be embedded in the raw data, it is not clear which features are suitable for capturing and identifying traffic status. For this reason we studied several candidate features and analyzed them to determine which had explanatory power.

(2) What is a good or bad traffic state?

It is not straightforward to define a good or bad traffic state, because there are no obvious objective criteria. But based on past data, we developed a method to make decisions on traffic states in comparison to what is *expected*. This method was then evaluated by comparing the estimated states to surveys of real drivers.

(3) What is a good way to classify the current traffic state, given past data?

When past data is given, there are several possible methods to classify an observation into one of several traffic states, for example: maximum a-posteriori (MAP), maximum likelihood (ML), or simple threshold-based classification. Although MAP would be the best in the ideal case, it is not applicable in our situation as its assumptions do not hold. Consequently we compared and evaluated the remaining methods to find the right one for our setting.
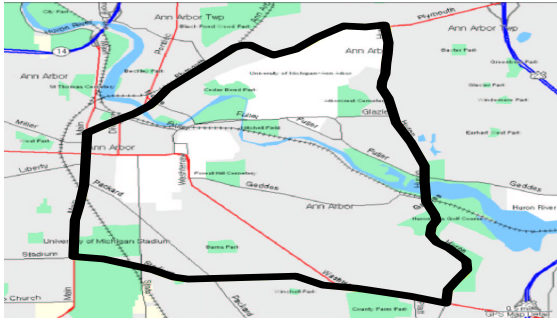
Figure 2: The pre-specified route where GPS data was collected for the preliminary study in Ann Arbor, Michigan.

(4) How can a traffic state be quantified for comparison?

Even if we can identify all the bad traffic states correctly, all of them may not be the same bad states; one can be worse than the other. Thus we also developed a metric to quantify traffic states. We will show that this metric can be useful for the detection of unusual traffic states such as traffic backup due to construction.

(5) How accurate is our system?

For evaluation, we used real GPS traces collected from several vehicles over ten months. Furthermore, we used forward visual data recorded in the vehicles and took visual surveys to obtain the ground truth on traffic states for further analysis.

4. EXTRACTING TRAFFIC INFORMATION

### 4.1 Preliminary data collection

For a preliminary study, we collected GPS traces by driving a GPS-equipped vehicle in Ann Arbor, Michigan. Although it would be ideal to deploy a number of GPS-equipped vehicles simultaneously for GPS data collection, for cost reasons we drove a single vehicle along a pre-specified route repeatedly.

We selected a route passing through the Ann Arbor downtown and suburban areas so that we can build a system working in both areas. Specifically, the route starts from north campus of the University of Michigan along Plymouth Road, Main street through downtown, Stadium Boulevard, Washtenaw Avenue, Huron Parkway, and Plymouth Road back to north campus as shown in Figure 2. The total distance is about 10.5 miles and there are 30 traffic lights along the route. All roads have two lanes per direction except the downtown area with a single lane per direction. Speed limits are 25 mph on Main, 40 mph on Plymouth, Huron Parkway, and Stadium, and 45 mph on Washtenaw.

We used a Garmin cf que 1620 compact flash memory type GPS receiver, installed in a Dell Axim x50v PDA. The GPS receiver records timestamp, altitude, speed, distance, heading, and coordinates once every 4 to 10 seconds, irregularly. Between 2 PM and 7 PM from April 1 to April 20, 2006 and from May 25 to June 1 in the same year, we have collected GPS data of 50 laps on weekdays from a personal vehicle driven by four different drivers to reduce any bias caused by each driver. We will also show in Section 6 that traffic data is generally driver-independent in these scenarios.
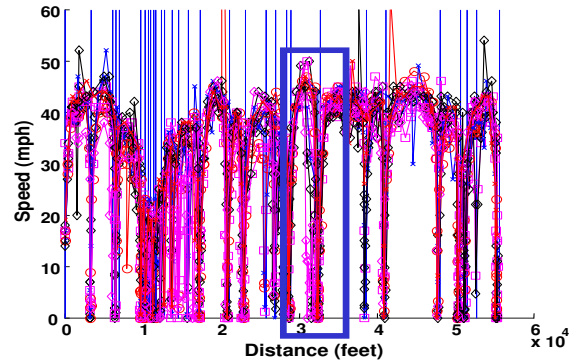
(a) Speed changes along the entire path.
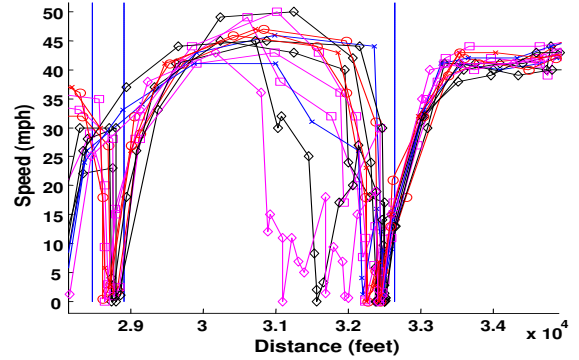
(b) A magnified graph of the rectangular area in (a).

Figure 3: Sample trace data along the pre-specified path. Vertical lines indicate the locations of intersections.

During data collection, there was only a single one-day road construction, blocking a right lane and causing an unusual traffic backup near the intersection of Stadium Boulevard and Washtenaw Avenue on June 1. There was no severe weather except for light rain on April 2 and 3, and no car accident throughout the entire period of data collection. Mild to moderate traffic jams occurred regularly during rush hours depending on the road.

4.2 Road segmentation

Figure 3 shows vehicle speed changes along the entire path specified in Section 4.1. Vertical lines in the plot indicate the locations of intersections on the path. Speed tends to vary with location as shown. However, there are two observations to make. First, speed changes appear to repeat in a similar and consistent fashion whenever the vehicle traverses the same road segment between two neighboring intersections, as shown in Figure 3(b). This is because vehicles traversing a road segment are affected by the same fundamental road conditions such as traffic light, road length, width, and the number of lanes. On the other hand, each road segment has its own unique shape of traffic pattern, because different segments have different conditions, as shown in Figure 3(a). Thus, traffic patterns must be characterized on a segment-by-segment basis. Figure 3(a) also shows that roads should be divided into segments between two neighboring intersections, as traffic lights are natural traffic delimiters. Likewise, stop signs and crosswalks can also be used as delimiters for road segmentation. Thus there will
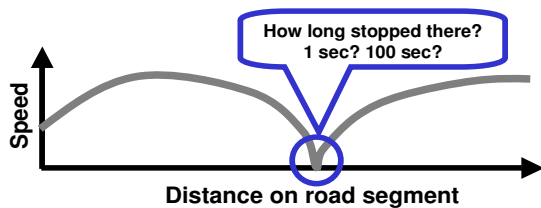
(a) A speed-location plot that loses temporal information.
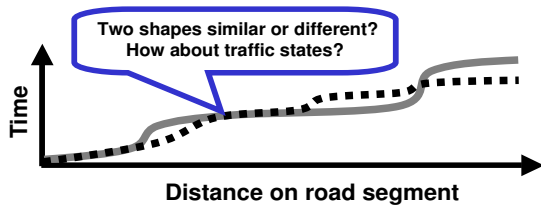
(b) A time-location plot which is hard to represent distinctive traffic states.

Figure 4: Problems of traffic feature candidates.

be more segments in downtown areas and fewer in suburban areas. Locations of traffic delimiters can be easily found from GPS traces by looking for frequent stopping positions. The coordinates of these locations can be acquired from a digital map. Throughout the paper, our traffic analysis and characterization are on a segment-by-segment basis. Thus, there will be no significant trade-offs between too many and too few segments in a region.

4.3 Traffic features and problems

GPS data contains the receiver's latitude, longitude, and timestamp, which are not directly related to vehicular traffic. Optionally, recent GPS receivers also provide vehicle speed and heading information. However, it is not clear how traffic state may be described with such information. For example, how low must speed be before it is interpreted as bad traffic? What role is played by road characteristics such as road length, road width, or traffic light timing? How about traversal time or stopping distances? To address these questions, we extract several potential traffic features and check if they are suitable for capturing and describing traffic states.

1) **Speed feature - Average of speed samples on a road segment during some time window**: Consider the average of all speed samples from vehicles on a road segment during some time window. Presumably, a low average speed indicates worse traffic conditions and a high average implies better conditions. This may be true on highways, but it may not be the case on urban roads with a lot of traffic lights, because vehicles may have to stop even if there are no other vehicles at all. Because an average speed on a road segment during some time window does not consider spatial information—e.g., where vehicles stopped—it may not be able to clearly tell a traffic jam apart from a red light in urban areas.

2) **Speed feature - Speed distribution on a road segment during a time window**: Instead of average, a speed distribution may be used to represent overall speed information from vehicles on a road segment. However, it is not ideal because of the lack of spatial information. For example, if vehicles in good overall conditions are decelerat-

ing, stopping, and accelerating due to a traffic signal change, their speed distribution is likely to look similar to that of poor conditions where vehicles are frequently stopping and starting.

3) **Spatial speed feature - Speed changes along a road segment**: If vehicles send their speed and location data regularly, one may plot their speed changes along a road segment (a speed-location plot) and use them to maintain spatial information. However, one loses temporal information in the process. Suppose two vehicles happened to generate exactly the same speed-location plots, having stopped once at the same location as shown in Figure 4(a). But suppose that one has stopped there for just one second while the other has stopped for ten minutes due to some serious road-blocking accident. Although the two traffic situations are quite different, their speed-location plots turn out to be identical.

4) **Temporal speed feature - Total time spent in traversing a road segment**: From the timestamps in the GPS data, one can compute traversal times of vehicles on a road segment, expecting that poor conditions may cause a longer traversal time. In fact, this feature is directly related to each vehicle's average speed over time on the road segment, because the time average of vehicle speed is the road segment length—a constant that we can figure out from the GPS coordinates—divided by the total time spent passing through the road segment. So traversal time is another form of speed-related feature. However, it is not straightforward to find a perfect threshold to distinguish a good traversal time from a bad one, because traversal times are impacted significantly by traffic lights on urban roads.

5) **Spatial and temporal speed feature - Cumulative time along a road segment**: To overcome the problem of the speed-location plot and also to consider both spatial and temporal information, one may want to use a cumulative time plot along a road segment instead of using speed directly. Even if a vehicle stops for a long time, the time-location plot is able to catch it. In addition, it still contains speed information, because the inverse of slope at a certain location represents the vehicle's instantaneous speed there. However, the problem of this feature, as shown in Figure 4(b), is that there could be many different shapes of plots and there is no clear way to compare and categorize them. Suppose one choose to apply spatial correlation, spatial difference, or some other statistical measures (e.g., chi-square statistics) to categorize them into several different classes of shapes. However, it is still not easy to tell which shape of plot belongs to a good or bad condition, what threshold should be used to distinguish them, or which shape implies a better traffic condition.

6) **Other types of features**: GPS data can also provide other types of data in addition to speed. For example, when a traffic slowdown or backup occurs, drivers often have to decelerate and accelerate repeatedly. In the GPS trace data, this is shown as a local minimum in the speed-location plot. One may choose to count the number of local minima, expecting that a bad traffic is likely to have many such points. Likewise, one may choose to count the number of stopping points for the same reason. However, these features may or may not show real traffic status, because a traffic jam does not necessarily produce a large number of local minima of speed or stopping points. For instance, some road-blocking accident may cause serious traffic backup and make drivers

completely stop just once or twice, while drivers in good traffic conditions also may stop once or twice on a red light. On the other hand, if one uses the locations of local minimum speeds or stopping points to consider spatial information, it has to be noted that stopping far away from a traffic light does not necessarily mean a traffic backup either.

With all these problems considered, it is clear that (1) we should use both spatial and temporal speed information to capture traffic status and (2) we should also quantify it for comparison and classification. Moreover, (3) we want to avoid any manual setup in doing so. To achieve all of the above, we propose a spatio-temporal traffic status plot with dynamically adjusted quadrants and a universal traffic status metric.

4.4 Spatio-temporal traffic status plot

Among traffic feature candidates in Section 4.3, one that contains both spatial and temporal information was a cumulative time-location plot. The problem in using this was that it is hard to quantify the shapes of plots and compare them for identification and classification of traffic states. To overcome these drawbacks, we convert each observation to a single point in the spatio-temporal space, minimizing loss of spatial and temporal traffic information.

Figure 5(a) shows cumulative time-location plots of 50 traversals on a road segment on Washtenaw Avenue in the city of Ann Arbor, Michigan. Among the sample traces here, we pick one, denoted by an arrow, and illustrate how to convert it to a point in the spatio-temporal traffic status plot.

A point in the spatio-temporal traffic status plot has two values as coordinates: a temporal mean speed and a spatial mean speed. A temporal mean speed is the average speed over time, i.e., the length of a road segment divided by the traversal time. In a cumulative time-location plot, it is straightforward to compute a temporal mean speed, because the length of a road segment and the traversal time are X and Y coordinates of the end point of the time-location plot as shown in Figure 5(b).

On the other hand, a spatial mean speed is the average speed over location, i.e., the arithmetic mean of instantaneous vehicle speeds at the evenly spaced locations (e.g., every 50 feet). An instantaneous vehicle speed is equivalent to the inverse of tangential slope of time-location plot at each location, as illustrated in Figure 5(b). Intuitively, a lower spatial average speed given a particular temporal average indicates more "stop and go" driving.

If GPS data was not collected regularly, each gap between the data points can be interpolated. In fact, time-location plots in Figure 5(a) were interpolated so that the instantaneous speeds in every sampling distance could be computed. We have set the sampling distance to 50 feet, considering the accuracy (less than 25 feet) of GPS receiver that we used to collect the data.
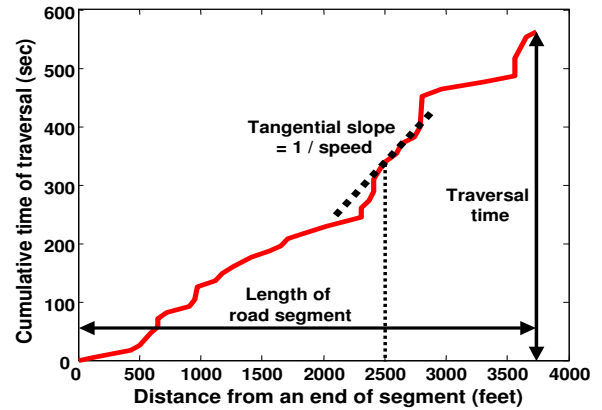
Figure 5(c) shows the spatio-temporal traffic status plot on the road segment converted from the time-location plots in Figure 5(a). The point denoted by an arrow in Figure 5(c) corresponds to the similarly-marked trace in Figure 5(a).
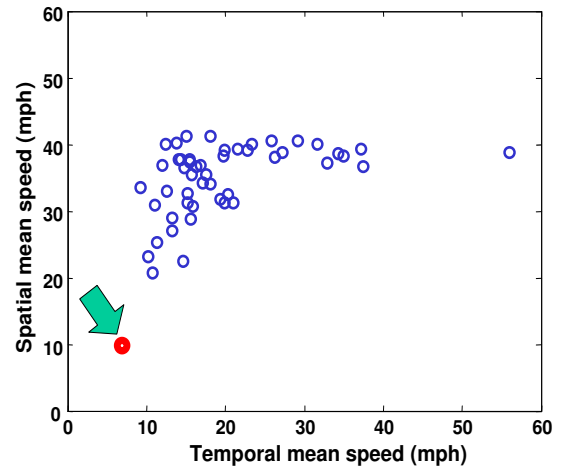
5. CLASSIFYING TRAFFIC STATES

Given a representation of individual vehicle/segment traversals, we next turn our attention to classifying these traversals into qualitative bins.
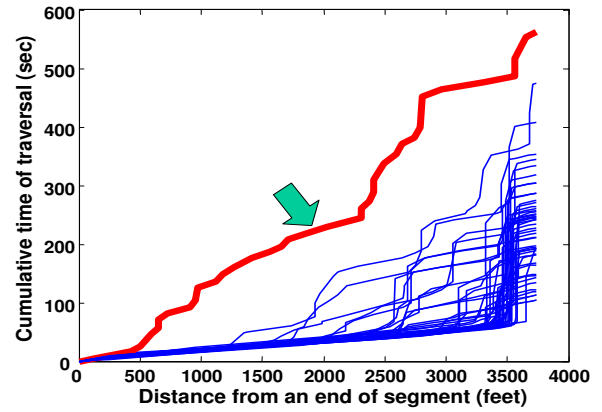
(a) Cumulative time-location plots of 50 sample traces.

(b) A sample trace in some heavy traffic jam during rush hours.

(c) A spatio-temporal traffic plot converted from (a).

Figure 5: Converting cumulative time-location plots on a road segment (on Washtenaw avenue in Ann Arbor) to points in the spatio-temporal traffic status plot

5.1 Clustering unlabeled past data

Although the spatio-temporal traffic status plot shows how traffic data on a road segment is distributed spatially and temporally, it still does not show which points in the plot belong to good or bad traffic conditions. To resolve this

issue, one could apply basic clustering algorithms to classify the data, because GPS trace data is unlabeled. Classifying unlabeled data is a well-known problem [21, 9] in the area of data mining and can be solved in general by clustering. However, most clustering algorithms are usually based on certain assumptions: for example, the number of clusters is known in advance, clusters are not completely overlapped in the scatter plot to make a distinction, or the characteristics of data do not vary with time. If these assumptions do not hold, classifying unlabeled data becomes more complicated. If clusters are exactly overlapped in the scatter plot or data points are almost uniformly distributed without distinct gaps, general clustering algorithms do not work well. Furthermore, if the total number of clusters is unknown, it complicates the problem even more. Unfortunately, GPS trace data suffers from all of these problems.

To address such shortcomings, Lapidot et al. [13] developed a method that sets the total number of clusters to a very large number (i.e., over-clustering) and then continues to merge two nearest clusters until a metric indicates that the best number of clusters has been reached. Unfortunately, this approach cannot help us process GPS trace data, because it is extremely inefficient to execute the over-clustering process given the fact that trace data points are collected at regular, frequent time intervals. Moreover, when we applied the over-clustering to the trace data with a voting strategy of clustering validation index [8] as a metric, we found that it was not helpful because GPS trace data points overlap without much in the way of distinct gaps as shown in Figure 5(c). The typical result was a single, large cluster containing almost all data points, plus a handful of outliers including only one or two data points. Since general clustering algorithms do not work with unlabeled GPS data, we developed a threshold-based clustering algorithm customized for our data, called *threshold-based quadrant clustering*. We describe this method below.

A temporal mean speed of X in the spatio-temporal traffic status plot is the road segment length divided by the traversal time. Because the road segment length is constant, the temporal mean speed is inversely proportional to the traversal time. So the rightmost point in Figure 5(c) represents the shortest traversal time on the road segment, which happens when a vehicle passes through the segment under the best traffic condition ever observed in the trace data. Generally, the vehicle does not stop during such traversals. However, not all vehicles are expected to traverse at this speed. For example, consider a vehicle that does not get caught on a red light in the best traffic condition and so it had the fastest traversal time on the road segment. Compare this to a vehicle which does stop on red for a full cycle, but otherwise has identical traffic conditions. The second vehicle has a traversal time close to the fastest total time plus a single red light cycle. In other words, the shortest traversal time plus a red light duration on the road segment implies that the traffic condition was still close to the best, but that the vehicle had to stop. Thus, if we can figure out the red light duration from GPS trace data, we can consider the minimum traversal time plus the red light duration as a reasonable threshold for *good* traffic conditions.

Figure 6 is a histogram of stopping durations of vehicles observed in the trace data on the same road segment as in Figure 5. If a vehicle stopped more than once, each stop is plotted separately. In Figure 6, stopping durations ranged
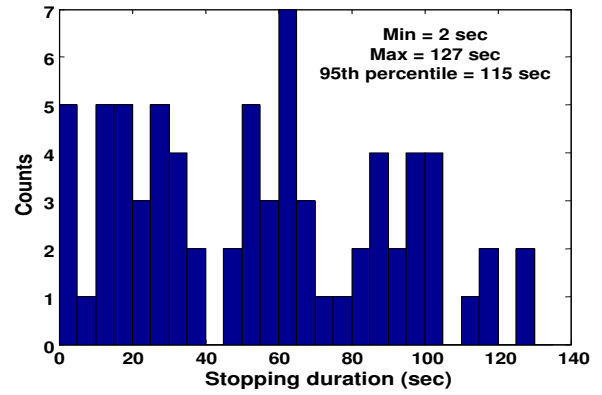

Figure 6: A histogram of stopping durations on the road segment in Figure 5. Minimum = 2 second, maximum = 127 second, and the 95th percentile = 115 second.

Intersection	95th percentile of stopping duration	Average red light duration (σ)
1	115	128.4 (2.5)
2	43	47 (0)
3	51	48 (0)
4	50	45 (0)
5	48	56.8 (0.4)
6	53	48 (0)

Table 1: The 95th percentiles of stopping durations observed in the trace data and the average red light durations measured 10 times during rush hours. σ is a standard deviation.

from 2 to 127 seconds. On the other hand, the red light duration measured ten times in a row during rush hours was 128.4 seconds on average with the standard deviation of 2.5. Although our GPS receiver has a low sampling rate of much less than 1 Hz, the maximum stopping duration observed in the trace data is very close to the actual red light duration. This suggests that the maximum stopping duration comes from the red light duration, as it agrees with intuition. Table 1 shows the comparison of the 95th percentile of stopping durations and the red light durations measured at six different intersections. Instead of maximum values, the 95th percentile of stopping durations was used to avoid extremely unusual data samples or glitches. The red light durations were measured ten times in a row during rush hours in the afternoon on May 5 and June 9, 2006. Errors in Table 1 are mostly due to the low sampling rate of the GPS receiver.

Once a red light duration is estimated from the trace data, we can split a spatio-temporal traffic status plot into two temporally different subspaces by setting a temporal threshold to

$$\frac{\text{road segment length}}{\text{5th percentile of traversal time} + \text{red light duration}}$$

as shown in Figure 7(a). The right-sided subspace indicates traffic conditions with "good" traversal times, while the left-sided one shows less desirable conditions.

Recall that low spatial averages generally indicate stop-and-go or slow-and-go conditions. An interesting observa-
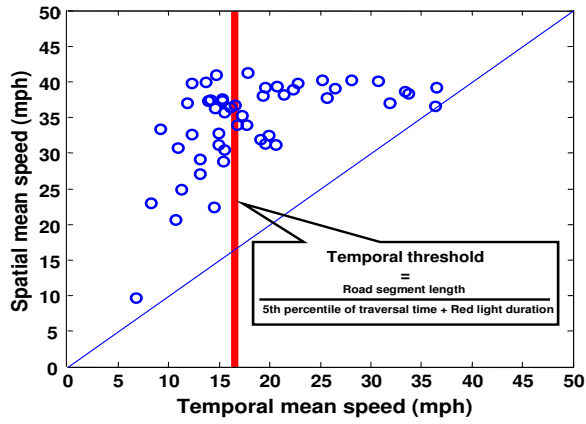
(a) A temporal threshold splitting vertically.
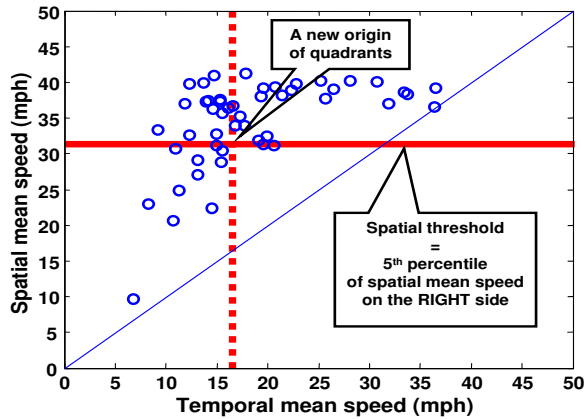
(b) A spatial threshold splitting horizontally.

Figure 7: Quadrants splitting the spatio-temporal traffic status plot into four spatially and temporally different subspaces on the same road segment as in Figure 5.

tion in Figure 7(a) is that a vertical (i.e., spatial) traffic breakdown or deterioration does not occur noticeably in the right-sided subspace, while trace data spreads out vertically in the left-sided one. In other words, trace data is distributed like a logarithmic curve which drops quickly as the data is getting close to the origin. This observation can be made very generally when we look at the spatio-temporal traffic status plots of all trace data.

Consequently we can define a spatial threshold as the 5th percentile of spatial mean speed in the right-sided subspace, which divides a spatio-temporal traffic status plot horizontally as shown in Figure 7(b). We used the 5th percentile of spatial mean speed instead of a minimum for the same reason as before.

Among four quadrants in Figure 7(b), the first (upper-right) quadrant indicates a both spatially and temporally good traffic—steady travel at good speed. The second (upper-left) quadrant represents a spatially good but temporally bad one. Similarly, the fourth (lower-right) quadrant depicts good average speed, but with some slow-and-go periods. The last (lower-left) quadrant shows poor traffic conditions.

Put another way, spatially bad traffic means that there is some serious traffic delay, slowdown, or backup all over the

entire road segment. For example, if a vehicle passes through a road construction area at a reduced speed continuously, it suffers from a spatially bad traffic. On the other hand, a low average speed but high spatial speed indicates a sudden stopping or slowdown for a while at a certain location on the road segment, but no where else. For instance, if a traveling vehicle suddenly has to stop for a while due to some road blocking accident, the trace data of the vehicle shows that its traversal time was unusually long (i.e., temporally bad) although a traffic was mostly good over the entire road segment (i.e., spatially good).

Traffic can be both spatially and temporally "bad"; the former is likely to give rise to the latter. For example, passing through a road construction area naturally results in a long traversal time. But the converse does not necessarily hold. In other words, a long traversal time with good speed "in most places" happens rather frequently, while the opposite is rare. This also explains why a spatial traffic breakdown usually does not occur in the right-sided subspace in Figure 7.

5.2 Identifying a current traffic state

Once traffic states are defined from past trace data on a road segment, we must classify new data points into one of them to identify a current traffic state. In general, one may think of three options in making a decision, depending on the given conditions. If no prior information is given, one can use one of the simplest methods such as the nearest neighbor (e.g., [7]) or threshold-based classification. If more information is given, one may use a maximum likelihood (ML) with likelihood functions or a maximum a posteriori (MAP) classification with both likelihood functions and a priori probabilities. If the given condition is ideally true, MAP would be the best choice to classify a current data point. However, this is not the case in our work, because all the prior information is empirically obtained from past trace data by using kernel density estimation [20]. Thus, to identify a current traffic state based on the past data, we used as candidates three different methods: ML, MAP, and our threshold-based quadrants defined in the previous subsection. When ML and MAP are used, traffic state quadrants previously defined from past data are slightly adjusted according to the computations of ML and MAP. On the other hand, the same thresholds are used for the threshold-based quadrants classification. We will show the evaluation results in Section 6.

5.3 Quantifying traffic states

Even if many data points belong to the same qualitative state, all of them do not necessarily reflect the same traffic quality. In particular, in the third quadrant, traffic status becomes worse as trace data points are getting closer to the origin. Thus it is necessary to quantify traffic status of trace data points in the plot to differentiate them. More specifically, we need to define a metric to quantify bad traffic. All the data points in the first quadrant can be considered as simply good. Since the second quadrant indicates a spatially good but temporally bad traffic state, it should be quantified based on how slow travel is in the quadrant. Likewise, the data points in the fourth quadrant mean a spatially worse traffic as they go down. Those in the third quadrant become worse as they are getting closer to the origin. In other words, traffic status becomes worse as the data points in the second,
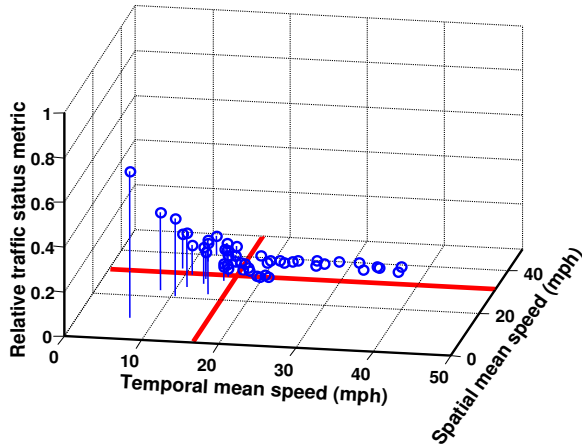
Figure 8: Relative traffic status metrics on the same road segment as in Figure 5.

third, or fourth quadrants are getting far away—leftwards, downwards, or both—from the first quadrant. Therefore we can define a distance-based metric from a new origin that the quadrants of spatio-temporal traffic status plot make as shown in Figure 7(b).

The distance-based traffic metric, denoted by $T(x, y)$, is defined as

$$T(x, y) = \begin{cases} 0, & (x, y) \in \text{1st quadrant} \\ x_o - x, & (x, y) \in \text{2nd} \\ (x_o - x) + (y_o - y), & (x, y) \in \text{3rd} \\ y_o - y, & (x, y) \in \text{4th} \end{cases}$$

where (x, y) is the coordinates of a data point in the spatio-temporal traffic status plot (i.e., $x =$ a temporal mean speed, $y =$ a spatial mean speed), and (x_o, y_o) indicates the coordinates of the new origin made by the quadrants. Manhattan distance was used instead of Euclidean distance, because it can show the metric difference more clearly in the third quadrant. Although the metric unit looks the same as spatial and temporal mean speeds, we do not use any metric unit, because it is different from a speed unit in the third quadrant and thus may be misleading.

Since traffic status metrics on different road segments are not comparable due to different quadrants, it may not be intuitively clear or informative how bad traffic really is given an estimate. We therefore convert T to a relative traffic status metric, T_r between 0 and 1, compared to the worst possible traffic status on each road segment as follows.

$$T_r(x, y) = \frac{T(x, y)}{T_w} = \frac{T(x, y)}{T(0, 0)} = \frac{T(x, y)}{x_o + y_o}$$

where T_w means the worst traffic status case (i.e., completely stopped vehicles) on the road segment. We used the worst traffic status instead of the best, because there is no best traffic state in a spatio-temporal traffic status plot whereas the worst state exists at the origin of the plot. Figure 8 shows the relative traffic status metric, T_r, on the same road segment as in Figure 7. By definition, $T(x, y)$ and $T_r(x, y)$ are all continuous over the entire spatio-temporal traffic status plot regardless of quadrants. Intuitively, this measurement is appealing, as drivers have some notion of "how bad" a particular segment is.

6. EVALUATION

In this section, we evaluate our scheme against two sets of traces. The first were taken by the University of Michigan Transportation Research Institute—a broad but irregular set of naturalistic driving behaviors. The second is a more regular set of traces described in Section 4.1.

6.1 Evaluation: UMTRI traces

The University of Michigan Transportation Research Institute (UMTRI) has performed a field operational test of their road departure crash warning (RDCW) system by deploying 11 fully-instrumented vehicles driven by 78 drivers for ten months beginning in May, 2004 [14]. Each vehicle has been equipped with the RDCW system, a forward video recorder, and a GPS logger, collecting the data of about 400 channels at 10 Hz. The total size of the data set is roughly 200 GB excluding video data. For evaluating our scheme, we obtained GPS traces and forward video data from UMTRI. GPS data contains vehicle's coordinates, speed, and heading with timestamps. Any private information that may reveal personal identities was not included in the data.

Although 78 drivers participated in the test for about 3 weeks each, GPS traces were very sparse, because test areas were not limited. Moreover, the traces showed that test vehicles were not frequently used. Thus it was difficult to find road segments with enough traversals for our evaluation. For this reason we chose one—Plymouth road between Huron parkway and Green road in Ann Arbor—near UMTRI having 80 traversals for ten months. The road segment length is about 950 meters and a single traversal GPS trace contains roughly 500–700 data samples depending on traffic conditions.

For evaluation, we had to know the ground truth of GPS traces on traffic states. However, it is practically impossible to know the perfect ground truth of traffic states, because there is no objective and absolute definition of good or bad traffic that everyone agrees on. Thus we gathered a representative estimate of traffic states by conducting surveys with forward video data. Since people are likely to lose their concentration if they have to look at the low-resolution video clips repeatedly, we randomly chose 40 traversals out of 80 on the road segment, and divided them into four subsets of 10 traversals each to maintain subject focus. We collected five surveys for each subset, a total of 20 surveys from participants.

For simplicity, we had participants select one of the only two traffic states: (1) good or normal and (2) bad. Then we collected a ground truth of traffic states, only if four or more answers out of five agreed on each traversal video clip. Among a total of 40 traversals, 36 of them—34 as good or normal and 2 as bad—had four or more common answers out of five. Throughout the following evaluation, we used these 36 traversals.

6.1.1 Results

Since survey answers are either of two traffic states, we categorized the first, second, and fourth quadrants as a good or normal traffic state, and the third quadrant as a bad one. Then we computed the following three metrics throughout the evaluation.

(1) **Accuracy**: A ratio of correct results over all results. A result is considered correct when it shows the same traffic state as the ground truth obtained from surveys.
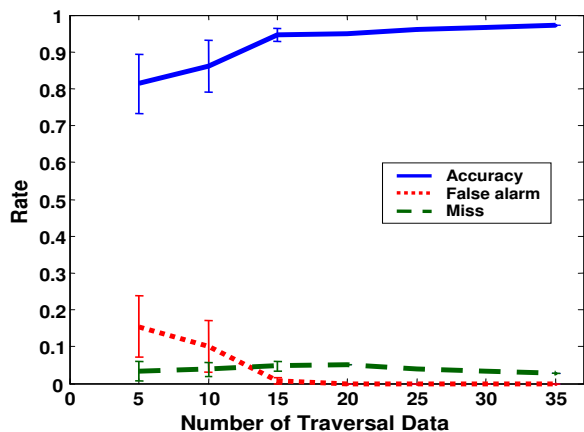
Figure 9: Accuracy, false alarm, and miss rates of the threshold-based quadrant clustering, varying with the number of traversal data used. Error bars indicate the 95% confidence interval.

(2) **False alarm rate**: A result is considered as a false alarm when it says a bad traffic state although the ground truth is a good one.

(3) **Miss rate**: A result is considered as a miss when it says a good traffic state although the ground truth is a bad one. Accuracy, false alarm rate, and miss rate sum up to 1 by definition.

Evaluation results consist of two parts. First, we evaluate the performance of threshold-based quadrant clustering that we developed to classify unlabeled GPS data into different traffic states in Section 5.1. Then we measure the performance of three methods—MAP, ML, and the same threshold-based quadrants as used in clustering—in identifying a current traffic state based on the quadrants obtained from past data. Throughout the evaluation, we will refer to a single GPS data point as a single traversal. The GPS trace of a single traversal contains more than 500 GPS samples along the road segment.

Figure 9 shows accuracy, false alarm, and miss rates of the quadrant clustering by varying the number of unlabeled GPS data. For example, the results with five traversal data can be computed when the quadrant clustering is applied to all subsets of five consecutive traversals among 36 of them. It is shown that accuracy of 90% or higher can be achieved when more than 10 traversals are used in applying the quadrant clustering to classify unlabeled GPS data. With 15 or more traversal data, accuracy becomes very consistent without much variation. So once more than 10 traversal data are collected on a road segment, the quadrant clustering method can characterize traffic patterns on the road segment and distinguish traffic states with a fairly high accuracy.

Table 2 shows the effect of weekday rush hours on the performance of the quadrant clustering method using all traversal data. It shows that the performance during rush hours is slightly better than that of others by about 10%. Although we may need more data to make sure, this suggests that the traffic patterns during rush hours are more consistent; we will show evidence for this in Section 6.2. But in general, it seems that accuracy is independent of time-of-day. Likewise, Table 3 shows a high accuracy overall in spite of seasonal differences.
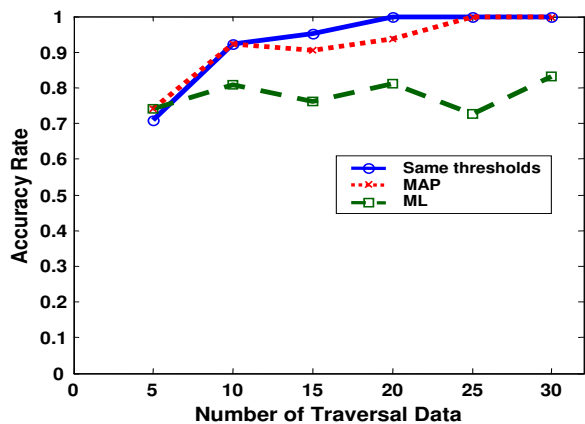
(a) Accuracy.
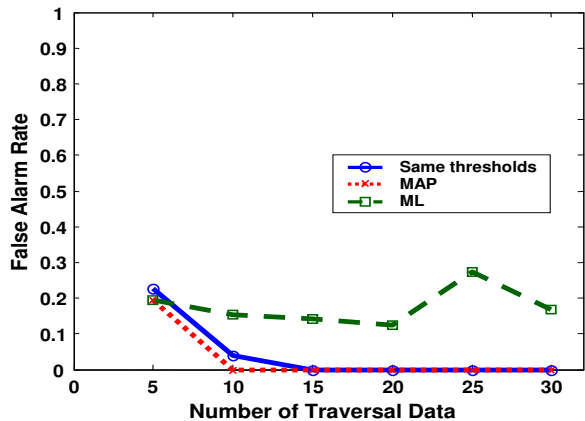
(b) False alarm rate.
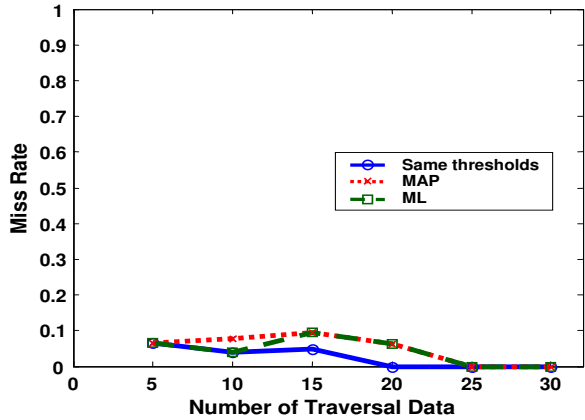
(c) Miss rate.

Figure 10: Accuracy, false alarm, and miss rates of three classification methods, varying with the number of past data.

Now that we have verified the performance of quadrant clustering, we evaluate the classification methods in identifying current traffic states based on the results of the quadrant clustering. In Figure 10, we compared the three methods in terms of accuracy, false alarm, and miss rates. Recall that the three methods are MAP, ML, and the quadrants with the same thresholds used in clustering. Although MAP and the quadrants method clearly outperform ML overall, the
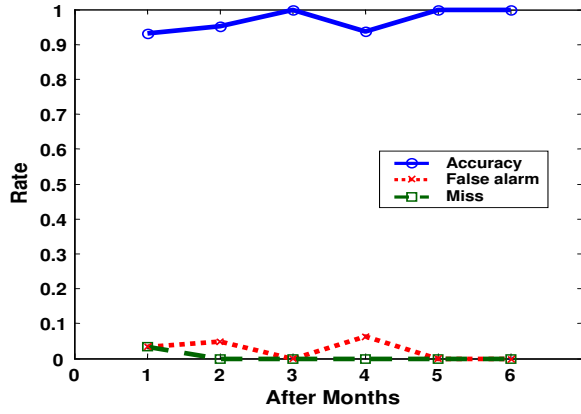
Figure 11: Accuracy, false alarm, and miss rates of the same threshold-based classification, varying with the freshness of past data. Previous two-month past data was used for each month.

	Weekday rush hours (7-9, 16-18)	Others (10-15, 19-23)
Accuracy	1	0.8947
False alarm	0	0.0526
Miss	0	0.0526

Table 2: Clustering accuracy, false alarm, and miss rates in weekday rush hours and other hours.

	Summer (Jun-Aug)	Fall (Sep-Nov)	Winter (Dec-Feb)
Accuracy	0.9333	0.8889	1
False alarm	0	0.1111	0
Miss	0.0667	0	0

Table 3: Clustering accuracy, false alarm, and miss rates in weekday rush hours and other hours.

quadrants method using the same thresholds is slightly better than MAP when the number of past data is less than 25. But both of them can identify a current traffic state with higher than 90% of accuracy by looking at just 10 past traversal data. In terms of false alarm and miss rates, the same quadrants method also slightly outperforms MAP.

In addition to the number of past data, we also evaluated the effect of freshness of past data on the performance of our system. Starting with a two-month window, we classified traffic states after a various number of months. A two-month window was used to have a sufficient basis on which to cluster. Although it was expected that the freshness of past data may affect the performance, Figure 11 shows that the effect of data freshness is almost negligible. While we need more data and analysis to make a conclusive statement, this may imply that traffic patterns on a road segment are very consistent over a significantly long period of time.

Table 4 shows the seasonal effect on the performance of the quadrants method in identifying a current traffic state based on the two-month window as before. Like Tables 3 in clustering, it is shown than we can achieve higher than 90% of accuracy regardless of seasons.

	Summer (Aug-Sep)	Fall (Oct-Nov)	Winter (Dec-Feb)
Accuracy	0.9091	1	0.9167
False alarm	0	0	0.0833
Miss	0.0909	0	0

Table 4: Classification accuracy, false alarm, and miss rates of the threshold-based method with previous two-month past data, varying with season.
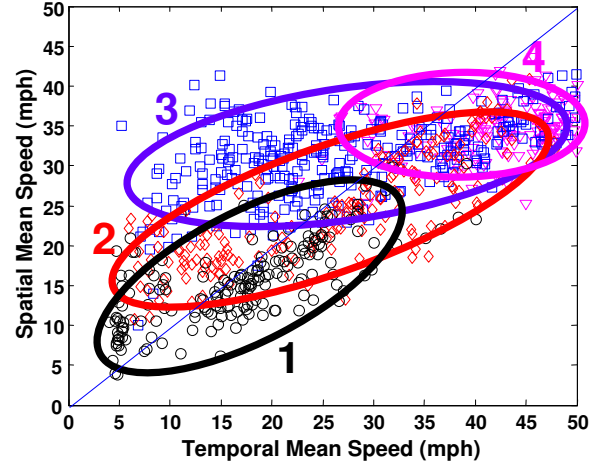

Figure 12: A spatio-temporal scatter plot of trace data on all road segments showing roughly four different distributions.

6.2 Case study: Traces for preliminary study

Now that the performance of our system has been evaluated, we will apply it to another trace data for further study. In Section 4.1, we have collected GPS traces for preliminary study and now will use them to show the consistency of traffic patterns on roads and the detection of unusual traffic status from GPS traces.

In collecting the trace data, four drivers drove the same vehicle along the same route within nearly the same time window for about four weeks. But, it was possible that there might be some bias caused by the different drivers and driving times. However, it turns out that traffic data is very consistent, independent of driver, except on the suburban roads with a relatively long segment length and a low density of vehicles. This also lends weight to the claim that traffic patterns during rush hours with a high density of vehicles are likely to be more consistent, made in Section 6.1.1. To show this, we categorized road segments into four different groups depending on the distributions on the spatio-temporal traffic status plot as shown in Figure 12.

Among four categories, roads in Category 1 generally have short segment lengths and high densities of vehicles, whereas those in Category 4 tend to have long segments and low densities. Usually, roads in the center of downtown belong to Category 1 and suburban roads to Category 4. Categories 2 and 3 are in between them in terms of segment length and vehicle density. Then, we plotted the GPS data collected by different drivers on each road segment in all four categories to see if the distributions of GPS data are independent of drivers. For simplicity, we compared the traffic data of two
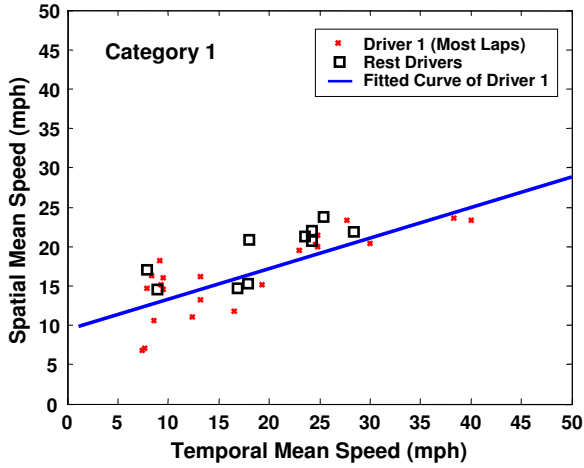
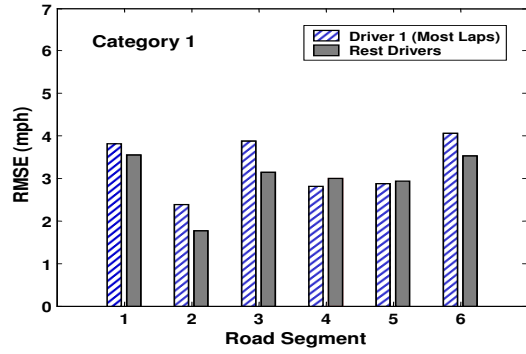**Figure 13: A least-squares fit to Driver 1's data points on one of the road segments in Category 1**

(a) Category 1 (e.g., roads in the center of downtown)

(b) Category 2.

(c) Category 3.

(d) Category 4 (e.g., suburban roads).

Figure 14: RMSEs of Driver 1's data and the rest drivers' data on each road segment, compared to the line fitted to Driver 1's data points in all four categories.

groups of drivers: a driver having driven most, denoted by Driver 1, and the rest of the drivers. There is not a single clear method to see whether the two different spatial distributions of data points are the same. But if they are visually close, we may verify the eye-ball test by comparing their errors to the least-squares fit of the data points as follows. First, we plot all the traversal data of Driver 1 and the rest of the drivers on each road segment. For example, Figure 13 shows the plot on one of the road segments in Category 1. Then we can plot a fitted line of Driver 1's data on a least squares basis. From the fitted line, we can compute each driver group's root mean squared error (RMSE) of traffic data. If we repeat this comparison on all road segments in each category, we can plot the graphs of RMSEs as in Figure 14.

If we compare the two distributions of RMSEs in each category with a chi-square statistic test, those in Categories 1, 2, and 3 are identical with probability of 99% or higher. In other words, the traffic data on the road segments in those categories is consistent regardless of drivers. However, on the relatively long and sparse roads in Category 4, drivers may have more space so that they can drive on their own by speeding or passing. So the traffic data in Category 4 may be more affected by driver behavior and thus less consistent.

If we collect enough traffic data on a road segment, we are able to not only identify a current traffic state but also detect unusual traffic conditions based on the past data. We will show how our system does this by using the trace data collected on a road segment on Stadium Boulevard in Ann Arbor where there was a one-day road construction event, blocking one of the lanes on June 1, 2006.

Suppose that today is the construction day and we have collected trace data on the road segment until yesterday. In Figure 15(a), all the traces except Traces 1, 2, and 3 show the past traversal data collected on this segment. Trace 4 indicates some unusual traffic jam at about 3:30 PM a couple of days ago, which was the worst traffic state ever observed in the trace data to this point. Then we can build a spatio-temporal traffic status plot of the road segment from the past trace data as shown in Figure 15(b).
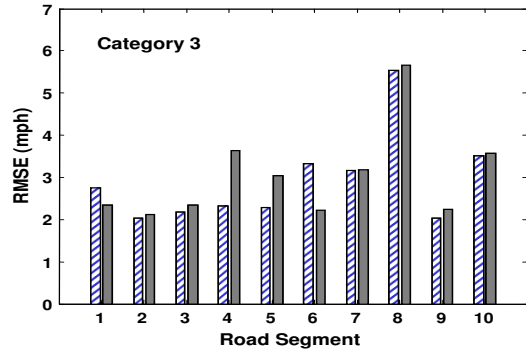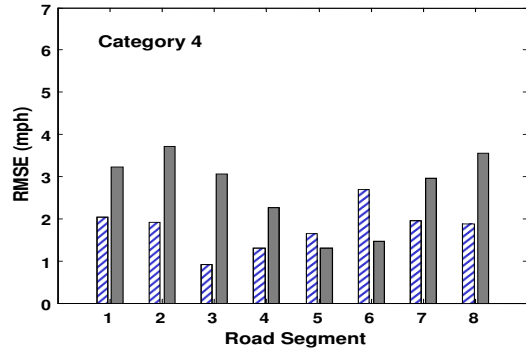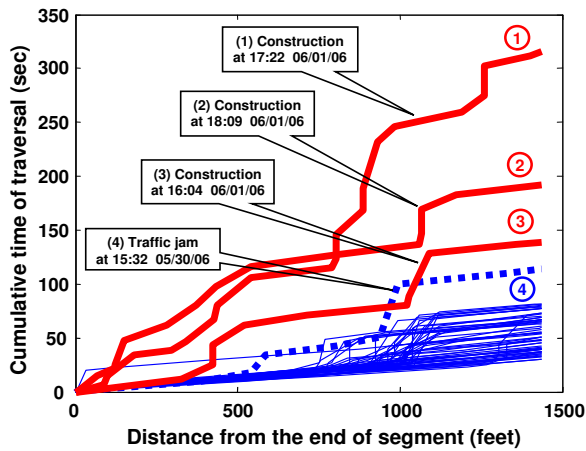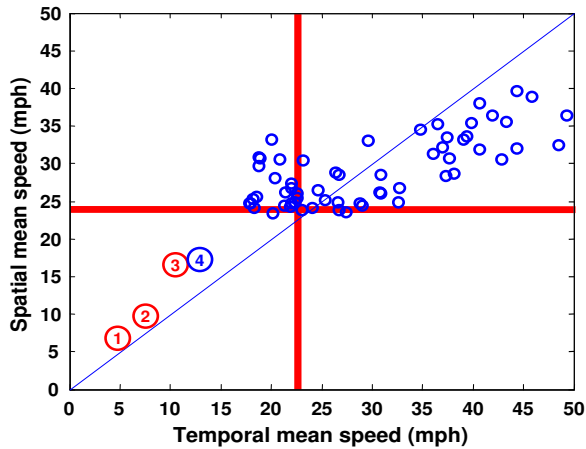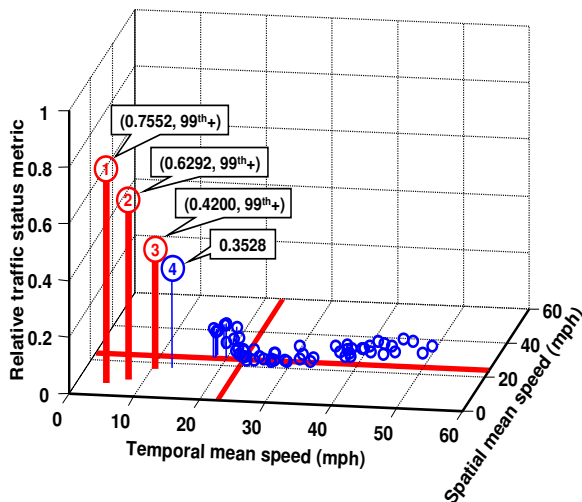
On the construction day, a vehicle passes through the road segment at about 4 PM, resulting in Trace 3 in Figure 15(a).
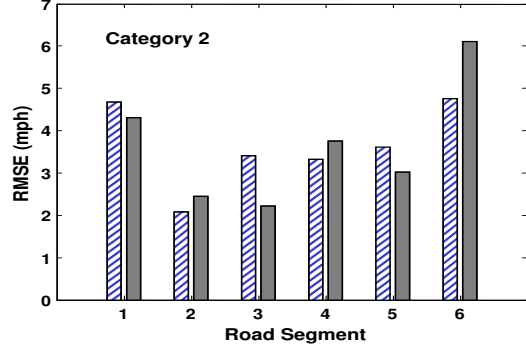
(a) Cumulative time-location plot.

(b) Spatio-temporal traffic status plot.

(c) Relative traffic status metrics and the percentile from past data.

Figure 15: Detection of unusual traffic status on a road segment (on Stadium boulevard in Ann Arbor) where there was one-lane blocking road construction on June 1, 2006.

According to our system, the current traffic status at that time shows a traffic metric of less than 0.5 but greater than the 99th percentile of the past data on the road segment. So a traffic service collecting and analyzing GPS traces can interpret the current traffic state as an unusual but not serious traffic delay.

At about 5:30 PM, Trace 1 is collected and its metric increases from 0.42 to 0.76. Clearly this is an extremely unusual traffic status during rush hours at this location. In about half an hour, the metric is reduced to 0.63, which means the traffic status on the road segment is still seriously bad but it is getting a little better. If the next metric gets lower, it will imply that traffic is going back to normal.

7. CONCLUSION

Traffic problems such as a traffic jam or delay are some of the most critical issues in our life. If correct traffic information can be given to drivers in advance, these problems may be significantly alleviated. However, collecting traffic information in the traditional way requires enormous costs for sensor installation and maintenance on the roads. As GPS gains in popularity and mobile data communication systems—e.g., mobile WiMAX, HSDPA/UPA, or EV-DO/DV—are available, a new solution is possible.

In this paper we proposed a simple yet very effective method that can capture traffic states in complex urban areas. Unlike existing systems which simply report vehicle speeds on the roads, our system can characterize unique traffic patterns on each road and identify traffic states in a road-specific manner. For evaluation, we applied our system to two different GPS trace data sets collected in the Ann Arbor area in Michigan. The results show that accuracy of higher than 90% can be achieved if ten or more traversal traces are collected on each road. Moreover, traffic patterns turned out to be fairly consistent over time, which allows us to use a larger history in classifying current traffic conditions.

8. REFERENCES

[1] Inrix. http://www.inrix.com/, 2006.

[2] Intelligent transportation systems - incident management. http://www.calccit.org/itsdecision/serv_and_tech/Incident_management/, 2006.

[3] Intellione. http://www.intellione.com/, 2006.

[4] New york city department of transportation - real-time traffic cameras. http:

//www.nyc.gov/html/dot/html/motorist/atis.html, 2006.

[5] Onstar. http://www.onstar.com/, 2006.

[6] Wimax. http://www.wimaxforum.org/, 2006.

[7] A. Ault, X. Zhong, and E. J. Coyle. K-nearest-neighbor analysis of received signal strength distance estimation across environments. In *Proceedings of 1st workshop on Wireless Network Measurements (WiNMee)*, Trentino, Italy, April 2005.

[8] N. Bolshakova and F. Azuaje. Improving expression data mining through cluster validation. In *Proceedings of the 4th Annual IEEE Conference on Information Technology Applications in Biomedicine*, pages 19–22, Birmingham, UK, April 2003.

[9] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. In *Data Mining and Knowledge Discovery*, volume 2, pages 121–167. Kluwer Academic Publishers, 1998.

[10] B. Coifman. Identifying the onset of congestion rapidly with existing traffic detectors. In *Transportation Research*, volume 37 of *Part A*, pages 277–291. 2003.

[11] N. H. Gartner, C. Messer, A. K. Rathi, F. L. Hall, R. J. Koppa, R. W. Rothery, R. Kuhne, P. Michalopoulos, J. C. Williams, S. Ardekani, E. Hauer, B. Jamei, R. J. Troutbeck, W. Brilon, N. Rouphail, A. Tarko, J. Li, and E. Lieberman. Traffic flow theory. Revised special report, Transportation Research Board (TRB), 2005.

[12] B. S. Kerner. *The Physics of Traffic*. Springer-Verlag, 2004.

[13] I. Lapidot, J. Ajmera, H. Bourlard, and I. McCowan. Unknown-multiple speaker clustering using hmm. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, September 2002.

[14] D. LeBlanc, J. Sayer, C. Winkler, R. Ervin, S. Bogard, J. Devonshire, M. Mefford, M. Hagan, Z. Bareket, R. Goodsell, and T. Gordon. Road departure crash warning system field operational test: Methodology and results. Technical report, The University of Michigan Transportation Research Institute, April 2006.

[15] W. Leutzbach. *Introduction to the Theory of Traffic Flow*. Springer-Verlag, 1988.

[16] W. Lin and C. Daganzo. A simple detection scheme for delay-inducing freeway incidents. In *Transportation Research*, volume 31A of *Part A*, pages 141–155. 1997.

[17] M. McNally, J. Marca, C. Rindt, and A. Koos. Tracer: In-vehicle, gps-based, wireless technology for traffic surveillance and management. Technical Report UCB-ITS-PRR-2003-23, California Partners for Advanced Transit and Highways (PATH), July 2003.

[18] K. Sanwal and J. Walrand. Vehicles as probes. Technical Report UCB-ITS-PWP-95-11, California Partners for Advanced Transit and Highways (PATH), January 1995.

[19] D. Schrank and T. Lomax. The 2005 urban mobility report. Annual report, Texas Transportation Institute, The Texas A&M University System, May 2005.

[20] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[21] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. In *Journal of Machine Learning Research*, volume 6, pages 211–232. 2005.

[22] J. Ygnace, C. Drane, Y. Yim, and R. Lacvivier. Travel time estimation on the san francisco bay area network using cellular phones as probes. Technical Report UCB-ITS-PWP-2000-18, California Partners for Advanced Transit and Highways (PATH), September 2000.