Detecting Crowdedness Spot in City Transportation

Siyuan Liu, Member, IEEE, Yunhuai Liu, Member, IEEE, Lionel Ni, Fellow, IEEE, Minglu Li, Member, IEEE, and Jianping Fan, Member, IEEE

Abstract—Crowdedness spot is a crowded area with an abnormal number of objects. Detecting crowdedness spots of moving vehicles in an urban area is essential to many applications. An intuitive method is to cluster the objects in areas to get the density information. Unfortunately, the data capturing vehicle mobility possesses some new features, such as highly mobile environments, supremely limited size of sample objects, and non-uniform, biased samples and all these features have raised new challenges that make traditional density-based clustering algorithms fail to retrieve the real clustering property of objects, making the results less meaningful. In this paper we propose a novel, non-density-based approach called mobility-based clustering. The key idea is that sample objects are employed as “sensors” to perceive the vehicle crowdedness in nearby areas using their instant mobility, rather than the “object representatives”. As such the mobility of samples is naturally incorporated. Several key factors beyond the vehicle crowdedness have been identified and techniques to compensate these effects are accordingly proposed. Furthermore, taking the detected crowdedness spots as a label of the taxi, we can identify one particular taxi to be a crowdedness taxi which crosses a number of different crowdedness spots. We evaluate the performance of our methods and baseline approaches based on real traffic situations (to retrieve the real traffic crowdedness) and real life datasets. Finally, the interesting findings are provided for further discussions.

I. INTRODUCTION

As more and more people are immigrating to urban cities, many metropolitan cities, especially in China, are facing a number of serious problems such as frequent traffic jams, unexpected emergency events, and even disasters. Many of these problems are relative to crowded moving objects such as vehicles, trains and pedestrians, etc. The research on Smart City Project, which is recently attracting growing attention [1], aims to address these problems.

One of the major tasks in the Smart City study is identifying crowdedness spots of moving vehicles in an urban area [1]. Informally, crowdedness spots of vehicles can be described as areas of high crowdedness of vehicles. The crowdedness spots of extremely high crowdedness are usually the sites of traffic congestion. An immediate application of crowdedness spot study is that we can predict vehicle speeds based on the crowdedness distribution. Indeed, crowdedness spots are often the potential sites of interests due to the higher likelihood of the events and opportunities (e.g., traffic jam, exhibitions, and commercial promotions). But it is hard to collect the location information of all the vehicles in the city because of the privacy issues or localization equipment limitations [2], [3]. In this work, we study crowdedness spot related issues utilizing taxi’s statistics as the sample data. The raw data set is from City Traffic Bureau of a major city in China [2], [3]. The ultimate goal of this research is to have a better understanding of city traffic via quantitative research on crowdedness spots. To do so we have several key tasks to accomplish: 1) define and quantify the vehicle crowdedness of an area; 2) picture the crowdedness distribution and identify the crowdedness spots; and 3) investigate the evolution of crowdedness spots.

Given the dynamic temporal and spatial information of moving vehicles, crowdedness spots can be considered as a general case of object clustering in mobile environments. Clustering for
static objects is a well-studied topic (e.g., DBSCAN [4], BIRCH [5], R-trees, CLARANS [6]). Many interesting algorithms have been proposed and exciting achievements have been made. In recent years, web related clustering, evolutionary clustering in low mobility environments and uncertain data streams have also drawn a lot of attention (e.g., micro-clustering [7], [8], evolutionary clustering [9], UMicro [10]–[12]). In our application scenarios, however, some new unique features make previous algorithms fail to capture the real clustering property of moving vehicles.

The first major challenge is incomplete information. Existing algorithms (for static or mobile) are all density-based approaches that use inter-node distances as a critical measure. They depend on the location information of target objects to exploit the clustering property. However, in many practical applications, it is unlikely to obtain such information from the total population of vehicles. It greatly degrades the effectiveness of density-based algorithms. The second major challenge is the extremely limited samples. The sample object set is a specific type of vehicles. It has very limited generality to represent general vehicles. Besides these, there are also some practical challenges and concerns such as drastic dynamism, and high mobility of the object.

To deal with these challenges, we propose a novel, non-density-based approach called mobility-based clustering. Mobility-based clustering is based on a simple observation that usually vehicles are intended to have high mobility (speed). Based on the history data learning, a vehicle of high mobility can largely indicate a low crowdedness and vice versa. By this, the sample vehicles are not simply used as objects but employed as “sensors” to perceive the vehicle crowdedness in nearby areas. The main advantages of mobility-based clustering are several folds. First, mobility-based clustering is less sensitive to the size of the sample object set, though a larger sample set can produce more precise readings of the crowdedness sensing. Second, mobility-based clustering does not require accurate location information and thus is robust to the location inaccuracy. Third, mobility-based clustering naturally incorporates the mobility of vehicles. It is therefore particularly suitable for high mobility environments.

Because of these advantages, mobility-based clustering greatly outperforms the existing density-based clustering algorithm in terms of prediction accuracy of vehicle density. Figure 1 illustrates a snapshot of the vehicle crowdedness in the city at 14:00PM on 12th Dec., 2006. The crowdedness situations are represented by color density. The dots indicate the instant locations of the sample taxis. If we run the density-based clustering using taxis as samples, Region 2 is considered as a crowdedness spot, and Region 1 is not considered as a crowdedness spot. But actually, Region 1 is a crowdedness spot, and Region 2 is not a crowdedness spot. Hence it is clear that the density-based clustering using taxis as samples will generate a quite deviated result. Such a deviation, which is mainly due to the inherent limitation of density-based approaches, is our main motivation for this work.

In this paper, we make contributions as following aspects. First, we propose a novel mobility-based model to quantify the crowdedness of certain areas, fully taking the mobility and object dynamism as an advantage. This is, to the best of our knowledge, the first attempt of non-density-based object clustering in literature. Second, several key factors, which have the great impact on the accuracy of the vehicle crowdedness measurements, are identified and investigated. Effective techniques to compensate the negative effects have been developed. Third, we find that different spots can be categorized using the presented spot mobility (it is actually the mobility of vehicles...
at the spot) and the crowdedness dynamism. This result provides usefully insight to the city planners for future city developments. Fourth, we study the crowdedness spots of top vehicle crowdedness values and investigation results show that several top crowdedness spots are quite locality consistent over time, while more crowdedness spots present more locality variations. Furthermore, base on the detected crowdedness spots, we can classify taxis. Specifically, we identify one particular taxi to be a crowdedness taxi, which crosses a number of crowdedness spots.

We validate results of mobility-based clustering through three ways. One is to use a subset of the sample set as the test set. We then compare the predicted speed with the actual speed of the test set. Our prediction error is less than 10 kmph. The second is to use traces of taxies and buses. We compare the results from these two traces. Comparisons show that the two traces produce quite consistent outcomes with the correlation coefficient up to 63% in average. The third is through field study. Referring the real traffic situation in Shanghai, mobility-based clustering can accurately measure the spot crowdedness. The accuracy is 61.3% at a certain spot using 0.3% of total vehicle population as the samples, compared with the accuracy of 3.3% of UMicro under the same scenarios [12].

The remainder of this paper is organized as follows. In Section II, we provide some preliminary information. Section III will be the foundation of vehicle crowdedness, focusing on an appropriate function of spot crowdedness. In Section IV, we discuss the crowdedness spot issues in more practical environments, including spot characterization, vehicle profiling, spatial-temporal crowdedness spots and crowdedness regions. In Section V, the crowdedness spot acquisition is also discussed in this section. We validate the mobility-based clustering through field study which is presented in Section VI. In Section VII we provide interesting findings via our methods running on real life datasets and the further discussions. In Section VIII, we give an overview of the related work. At last, we conclude the paper and outline the directions for future work.

II. PRELIMINARIES

In this section, we first present characteristics of the raw data set used in our work. Moreover, we introduce road gridding. At last, we present the main observations and design principles of mobility-based clustering.

A. Raw data set characteristics

The data set is originated from the City Traffic Bureau. It contains one-year-long real traces from 5631 taxis that have been equipped with GPS receivers (one for each). The GPS receivers periodically report their current states to a data center via GPRS links. The reports include the instant speed, speed direction, the geographic location and the status of occupied or unoccupied (by guest) of the taxi. In the remainder of this paper, we use the terms “sample” and “taxi” interchangeably, as well as the terms “location” and “spot” if not otherwise stated. The term “vehicle” is used to represent the general vehicles including taxi samples, taxis not sampled, buses, and private cars.

The GPS system that we use is installed for civic applications. Due to the low cost of these applications, the data reports mainly have the following limitations. First, the data set is incomplete. Though the data set contains statistics of over 5631 taxis, it accounts for no more than 0.3% of the two million vehicles in the city. Moreover, the statistics of samples may not be complete. Notable amount of reports were missing due to weak GPRS signals (via which taxis are connected to the system) or limited bandwidth of GPRS wireless channels. The error will be significant if we use this tiny sample to represent the large number of general vehicles. Moreover, all sample objects are taxis which are only one specific type of vehicles. Taxis are highly incentive-oriented that have strong preferences on some desired locations. They would like to aggregate on sites of high customer flows, such as business areas, train stations, and traffic reconnections. Such preferences make it a bad option to employ this one type of vehicles as the representative of others.

Second, it is known that the reported GPS data may not be accurate due to blocked GPS signals (e.g., taxis in tunnel or surrounded by high
buildings). Since GPRS is a paid communication service, it is costly to frequently report their current status information. In the city, taxis are allowed to report their data at an arbitrary time, with a desired 5 second period. In 5 seconds a vehicle can drive more than 80 meters at 60 kmph speed. Concerning all these factors, the location errors of vehicles are on the order of hundreds of meters. It becomes impractical to apply the traditional density-based approaches which are critically relied on the accurate locations.

Third, the data is biased in temporal and spatial spaces. For example, 90% roads have no data for more than 80% of the time in a day, and 50% have no data in 12 continuous hours. To the opposite, 80% of the reports are collected from 20% of roads. How to mine meaningful information from the biased samples is another great challenge. Motivated by these new challenges, we propose a novel, mobility-based clustering method.

B. Road griding

For ease of computation, we discretize the time dimension in the unit of second. The time instance \( t = 0 \) is the initial time. The time instances \( t \) and \( t+1 \) are consecutive time instances. We discretize the physical space by dynamically partition the whole area into a number of rectangle grids. The size of each grid depends on the intensity of the reports at the area, ranging from 10 meters for report rich areas to 90 meters for report scarce areas. We believe this granularity is sufficient for most applications. Each grid is represented by its center location so that all spots in the grid will be treated the same as that of the grid center. A 2-tuple \( \iota = (x, y) \) represents one grid where \( x \) is the index of the grid along the longitude and \( y \) is that along the latitude.

From our raw data, we are able to capture the speed direction. Generally speaking, the road is divided into two directions. Accordingly, we divide the speed mainly into two different sets, road direction set and reverse direction set.

We especially introduce domain knowledge to amend the grids and retrieve much more accurate spot locations. Because the road topology and type will impact the vehicle, not only the speed, but also the drive pattern, hence we study the following problems based on road grid. At the same time, the domain knowledge could help us purify the reports. For example, in practice, there may be some vehicles having low speed, but not indicating crowded spots. Because these spots may be the taxi stops or residential areas. Hence, to achieve better detection accuracy, we preprocess the raw data sets by learning from the history data.

C. Observations and design principles

Different from the traditional density-based approaches, mobility-based approach is based on two simple observations. The first one is that vehicles prefer high mobility in a sparse region. To the opposite, for security concerns vehicles will drive slowly when the nearby area is crowded. Motivated by it, we employ vehicles as sensors using their instant speed to sense the vehicle crowdedness of vicinity. In Figure 2, we conduct three field experiments to illustrate this observation. In Figure 2, x-axis is the speed, and y-axis is the number of vehicles. We select three spots in the city to verify our observation. The result shows that the low speed indicates high vehicle density, while the high speed indicates the low vehicle density. The second one is that the reported locations can be erroneous, while the reported speeds are usually quite accurate because they are directly obtained from the speedometers installed on taxis. In addition, for safety concerns sudden changes of speeds are rare. Therefore the speed errors coming from the unsynchronized reports are also small.

Roughly speaking, in mobility-based clustering we collect statistics of taxi speeds at each spot. The spot crowdedness is then a relative measurement concerning the instant speed, maximum
speed, and minimum speed. Though a higher crowdedness usually leads to a smaller mobility, a smaller mobility is not always caused by high crowdedness. Besides the spot crowdedness, there are many other factors having similar effects on taxi mobility.

First, one fact is that drivers may have various driving styles and habits. In particular, due to incentive-oriented nature, occupied taxis (by guests) often have higher speeds than unoccupied taxis which may be looking for guests (a typical driving behavior in China). Profiling these different drivers will help to interpret taxi motility more precisely.

Second, mobility of vehicles is environment dependent. Some roads are designed for high speed traffic, while others are mainly for connection purposes. Traffic lights clearly slow down vehicles, which is not due to the high crowdedness of the spots. We should characterize spots so that to reduce these negative effects.

Third, spot crowdedness may have spatial and temporal correlations. Adjacent spots may have strong connections in between. A crowded spot is very likely to be crowded again in the next time stamp. Crowdedness spots may evolve over both time and spatial dimensions. To well capture the crowdedness of spots, we should take all these factors into account so that the derived crowdedness values can adequately reflect the real crowdedness of spots.

### III. Spot Crowdedness Fundamental

#### A. Assumptions and notations

Intuitively, the crowdedness of a spot reflects how crowded a spot is. It becomes difficult when the distribution of the vehicle set is unknown. In order to quantify the crowdedness of unknown object set at a spot, we propose a mobility-based model.

We assume a set $N$ of vehicles deployed in a two-dimensional city plane $A$. Among these $N$ vehicles, a small subset $N_s \subseteq N$, $|N_s| \ll |N|$, is the sensor set that we have the completed knowledge. We, however, have no knowledge of the rest part $N \setminus N_s$. Our task is to use reports of $N_s$ to infer the status of objects in the remainder part $N \setminus N_s$, especially their spatial properties.

### TABLE I NOTATIONS USED IN THIS PAPER

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>Set of target vehicles; $N_s \subseteq N$, $</td>
</tr>
<tr>
<td>$\phi_m^{(t)}$</td>
<td>$\phi_m = (x_m, y_m, v_m, \beta_m)$ is a sensor report</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>The set of all reports</td>
</tr>
<tr>
<td>$V^{(t)}_\Phi(\ell)$</td>
<td>The speed spectrum of a location $\ell = (x, y)$ at $t$</td>
</tr>
<tr>
<td>$v^{(t)}_\Phi(\ell)$</td>
<td>The spot speed $\ell$ at $t$ computed based on $V^{(t)}_\Phi(\ell)$</td>
</tr>
<tr>
<td>$H_L$</td>
<td>Linear crowdedness value</td>
</tr>
<tr>
<td>$H_S$</td>
<td>Statistical crowdedness value</td>
</tr>
<tr>
<td>$Ls^{(t)}$</td>
<td>The set of crowdedness spots</td>
</tr>
<tr>
<td>$Rs^{(t)}(\zeta_{th})$</td>
<td>Crowdedness region where $\zeta_{th}$ is the threshold to determine a crowdedness region</td>
</tr>
<tr>
<td>$Rs^{(t)}_{k}$</td>
<td>Top-k crowdedness region that contains crowdedness spots of k largest crowdedness values</td>
</tr>
<tr>
<td>$\theta_{Rs_{1},Rs_{2}}$</td>
<td>Area difference ratio of two crowdedness regions</td>
</tr>
<tr>
<td>$\nu_{Rs_{1},Rs_{2}}$</td>
<td>Crowdedness difference ratio of two crowdedness regions</td>
</tr>
</tbody>
</table>

Assume the city $A$ is a given area. Sensory objects in $N_s$ move arbitrarily. Their mobility provides a means of sensing the crowded conditions of vicinity of $A$ unintentionally. These sensors are supposed to report their current states periodically in every five seconds in an asynchronous manner. The report is in a form of 6-tuple $\phi = (m, x, y, v, \beta, t)$ where $m \in N_s$ is an object identifier to uniquely identify the sensor associated with the report, $(x, y) \in A$ is $m$’s current location, $v$ represents $m$’s instant speed, $\beta$ is a binary value to indicate whether $m$ is occupied by guest or not, and $t$ is the reporting time. It is also written as $\phi_m^{(t)} = (x_m^{(t)}, y_m^{(t)}, v_m^{(t)}, \beta_m^{(t)})$ if there is no confusion. As the reports are sent back every five seconds, every second we will have the reports from one-fifth of the sensors in average. These reports are called sensing reports.

To increase the granularity of our sample data set, besides these real sensing reports we estimate the states of vehicles during the non-reporting time by applying a linear interpolation method.

More specifically, two sensing reports of a sensor are said to be consecutive if there is no sensing report in between along the time dimension. Given two consecutive sensing report $\phi_m^{(t)}$...
and \( \phi_m(t_2), t_2 \geq t_1 \), an interpolation report \( \phi_m(t) \) at time instant \( t \in [t_1, t_2] \) can be calculated by \( x_m(t) = \frac{t-t_1}{t_2-t_1}(x_m(t_2) - x_m(t_1)) + x_m(t_1) \). Those for other variables \( y_m, v_m \) are similar. We assume \( \beta_m^{(t)} = \beta_m^{(t_1)} \).

Such reports generated by applying interpolation method are called interpolation report. In this work we do not differentiate the sensing report and interpolation report, using both as input for our study and simply call them reports. We denote such reports generated by applying interpolation method are called interpolation report. In this work we do not differentiate the sensing report and interpolation report, using both as input for our study and simply call them reports. We denote the set of all reports as \( \Phi = \{\phi_m|t \in [0, +\infty), \forall m \in N_s\} \). Note that interpolation report is an estimation of the vehicle’s mobility and therefore may introduce errors. These errors will finally propagate to the measurement and prediction of the crowdedness. Better estimation schemes are left for future work.

Table I lists some of the notations used in this work. In the next we define the crowdedness functions which are used to quantify the degree of spot crowdedness.

**B. A simple linear crowdedness function**

A simple way to quantify the crowdedness of a spot is applying a linear function of the instant speed, the maximal and minimal speeds of the spot.

**Definition 1:** Given a physical location \( \iota = (x, y) \) and the report set \( \Phi \), the speed spectrum \( V_\Phi^{(t)}(\iota) \) is the set of all the reported speeds (of vehicles) at time \( t \) in location \( \iota \) in \( \Phi \), i.e.,

\[
V_\Phi^{(t)}(\iota) = \{v|\exists \phi_m^{(t)} \in \Phi, x_m^{(t)}, y_m^{(t)}, v_m^{(t)} = (x, y, v)\}
\]

(1)

The speed spectrum at all time instances is also written as \( V_\Phi(t) = \cup_{t \in [0, +\infty]} V_\Phi^{(t)}(\iota) \). Since the time is discrete, \( \Phi \) contains a finite number of reports and thus the speed spectrum is finite as well.

**Definition 2:** The spot speed \( v_\Phi^{(t)}(\iota) \) is defined as the average of the speed spectrum in location \( \iota \) at \( t \),

\[
v_\Phi^{(t)}(\iota) = \frac{1}{|V_\Phi^{(t)}(\iota)|} \sum_{v_m \in V_\Phi^{(t)}(\iota)} v_m
\]

(2)

**Definition 3:** The linear crowdedness \( H_{L}^{(t)}(\iota) \) of a given spot \( \iota \) is defined as an exponential moving average of the complementary relative ratio of the instant speed over the speed spectrum,

\[
H_{L}^{(t)}(\iota) = \alpha_t H_{L}^{(t-\tau_t)}(\iota) + (1-\alpha_t) v_{\text{max}}(\iota) - v_{\Phi}^{(t)}(\iota)
\]

(3)

where \( v_{\text{max}}(\iota) \) and \( v_{\text{min}}(\iota) \) are maximal and minimal speeds of \( \iota \), and \( \alpha_t \) and \( \tau_t \) are two parameters to capture the dynamism of \( \iota \).

**C. A statistical crowdedness function**

The linear crowdedness function \( H_{L}^{(t)}(x, y) \) is based on an implicit assumption that speeds of vehicles are uniformly distributed in the speed spectrum, which is not necessarily accurate in practice.

Figure 3 depicts the speed spectrum of four different spots on 18th Dec., 2006. It is a cumulated distribution function (CDF) of speeds. Apparently the distributions are not uniform and different spots present various distributions. For Spot 3, though the maximal speed is as large as 120 kmph, over 80% of speeds are no more than 25 kmph. In order to well capture the effect of this un-uniform speed distribution on the crowdedness measurement, we propose a statistical crowdedness function.

**Definition 4:** The statistical crowdedness \( H_{S}^{(t)}(\iota) \) of a given spot \( \iota = (x, y) \) is defined as an exponential moving average of the complementary cumulative distribution function (CCDF) of the instant speed over the speed spectrum, i.e.,

\[
H_{S}^{(t)}(\iota) = \alpha_t H_{S}^{(t-\tau_t)}(\iota) + (1-\alpha_t) (1-P(v \leq v_\Phi^{(t)}(\iota)))
\]

(4)

where \( P(v \leq v_\Phi^{(t)}(\iota))) \), \( v \in V_\Phi(\iota) \) represents the probability that the speed at \( \iota \) is less than or equal to the spot instant speed \( v_\Phi^{(t)}(\iota) \). \( \alpha_t \) and \( \tau_t \) are two parameters to capture the dynamism of \( \iota \).

The statistical model is a more general form as the linear model can be considered as a special case with a linear CDF. To calculate the \( H_{S}^{(t)}(\iota) \), we maintain the spot history information for six months. As the computation is carried out at a data center in a centralized manner, the computational and storage cost is acceptable.

We validate the two models in Section VI-A, and conclude that the statistical crowdedness function produces a much better fit crowdedness distribution. Accordingly, in later of this paper
we will use the statistical crowdedness function to measure the spot crowdedness.

In the following section, we study the methods to improve the statistical model to investigate the mobility.

IV. SPOT CROWDEDNESS IN PRACTICE

In practice, many factors influence the mobility of nearby vehicles, making mobility-based model fail to accurately capture the spot crowdedness property. In this section, we investigate the significance of these factors and present techniques to compensate these effects.

A. Characterizing spots

Apparently the vehicle mobility is highly environmental dependent. The mobility (of vehicles) at different spots is not necessary to be identical. In this subsection, we unveil the various characteristics of the spot mobility, focusing on the system parameters \( \alpha \) and \( \tau \), which are used to capture the distinctive property of the spot crowdedness.

The parameter \( \alpha \) is a smoothing factor of exponential moving average in \( H_s \). It is used to capture the degree of dynamism of the spot mobility. In general, a smaller \( \alpha \) indicates a higher dynamism and vice versa. The parameter \( \tau \) is the interval between two moving averages. It reflects the periodic property of the spot crowdedness.

In Section VI-A we set parameters as \( \alpha = 0.5 \) and \( \tau = 1 \) for validation convenience, which may not be adequate in practice. Indeed, different spots will present distinctive mobility behaviors, resulting in various settings. As shown in Figure 3, Spot 1 presents a speed spectrum of nearly a step function. Spot 2 and Spot 3 are similar that the major portion of 60% of speeds are around 20 kmph. Spot 4 has a mimic trend of Spot 2 but the major part is around 40 kmph. Figure 4 gives a more detailed speed distribution of Spot 1 over one day on 16\(^{th}\) Dec., 2006. We hope to extract useful information from this deviated speed series and bring insight into the dynamism of vehicle mobility in the location.

In order to systematically study the speed distribution, we apply Fourier Transformation (FT). FT can transform the function from time domain to frequency domain, revealing inherent periodic property of original corresponding function as well as the amplitude of the corresponding frequency. Specifically, given the speed distribution function over time \( v(t) \) at a spot \( t \), its FT can be calculated by,

\[
\hat{f}(\xi) = \int_{-\infty}^{+\infty} v(t)(l)e^{-2\pi\iota\xi dt}
\]

where \( t \) is the variable. Figure 5 depicts the FT function \( \hat{f}(\xi) \) over the frequency domain of the speed distribution in Figure 4. We can clearly identify several periods of Spot 1. Among them, the dominating one is at the frequency of 0.15 (per minute) with the amplitude of \( \hat{f}(0.15) = 0.86 \). Noticing that in this area the interval between traffic lights are 5 minutes, which is consistent to the period of the second largest amplitude with the frequency of 0.2. We therefore believe this second period does not reflect the innate periodic property of the spot. As no other periods are comparable to the dominating period in terms of the amplitude, we set the two parameters according to the dominating period given by

\[
\begin{align*}
\alpha_t &= \max(\hat{f}(\xi)) = 0.86 \\
\tau_t &= 1/\arg\max_{\xi}(\hat{f}(\xi)) = 6.67\text{minutes}
\end{align*}
\]

The two derived parameters \( \alpha_t \) and \( \tau_t \) are called dominating average weight and dominating period of the spot. They will be computed and associated with each spot.

The evaluation of this approach is reported in Section VI-A. The error of the speed predicting is reduced by 14%.

B. Sensor object profiling

Intuitively, vehicle mobility depends on drivers’ driving styles. A younger driver may drive more aggressively with higher speeds, while an older driver may be more prudent to drive slowly. In particular, taxis are supremely incentive-oriented. They prefer to high mobility with guest taken to deliver the guest as soon as possible, while low mobility when it is empty to look for a new guest.

The data shows that the reality is, however, not always consistent with these intuitions. Figure 6 depicts the CDF of mean speeds of individual taxis at four different spots. Results of other spots are similar. It shows that though different spots present different mobility, the majority of
taxis have a similar behavior at a same spot. Their mean speeds do not differ too much. For instance, at spot NanJing Road, more than 40% have the mean speed between 20 and 30 kmph, and 30% are between 30 and 40 kmph, though the mean speed ranges from 0.5 kmph to 100 kmph. At spot YanAn Road, 50% taxis have the mean speed between 20 and 30 kmph. As other factors have more influences on the spot mobility, we ignore the impacts of various drivers, assuming that different drivers have the similar behavior at the same spot.

We, however, find that the states of occupied/unoccupied have more significant impact on taxi mobility. Figure 7 depicts the differences of mean speeds when taxis are occupied and unoccupied. The CDF of the differences at four different spots are pictured. At NanPu Bridge, 76% taxis have the differences between 4 and 12 kmph. At YanAn Road, 55% taxis have the differences between 4 and 11.5 kmph. As the difference is common for each spot, its impact can hardly ignored. For this we use a simple approach to calibrate the spot mobility according to the state of the reported taxis, whether it is occupied/unoccupied. We use occupied taxis as the base line which are considered as normal. An unoccupied taxi will be calibrated to have more mobility than the reported speed when we according it compute the the spot mobility. The degree of calibration is determined by the spot based on its differences of mean speeds when taxis are occupied and unoccupied. With this state calibration the prediction error is further reduced by about 16% when $N_a > 1200$.

C. Crowdedness spots and crowdedness regions

Intuitively the crowdedness of a spot is correlated along the time dimension. In Section IV-A we studied the general temporal property of spot crowdedness. In this subsection we investigate crowdedness spots of top crowdedness values.

**definition 5:** Crowdedness spot set $L_s(t)$ is defined as the set of spots with the local maximal of the crowdedness value, i.e.,

$$L_s(t) = \{ t | \exists \varepsilon > 0, \forall t' \in A, \text{dist}(t', t) < \varepsilon \Rightarrow H_s(t') < H_s(t) \}$$

where $\text{dist}(t', t)$ denotes the physical distance between the spot $t' = (x', y')$ and $t = (x, y)$,

$$\text{dist}(t', l) = ((x' - x)^2 + (y' - y)^2)^{1/2}.$$  

It is, however, not straightforward to study the temporal property of crowdedness spots directly. One major reason is that crowdedness spots depend on the crowdedness of nearby spots. One crowdedness spot may disappear because of a minor change of the nearby spot crowdedness, and then reappear soon due to the same reason. From the application point of view, these two changes may be too minor to be taken into consideration. To facilitate the investigation on crowdedness spots, we define the crowdedness regions and top-$k$ crowdedness regions.

**definition 6:** Given the crowdedness distribution $H_s(t)$ and a threshold $\zeta_{th}$, the crowdedness regions $R_s(t)_{\zeta_{th}}$ are the set of spots with the crowdedness value being more than $\zeta_{th}$, i.e.,

$$R_s(t)_{\zeta_{th}} = \{ t | \exists \varepsilon > 0, H_s(t) \geq \zeta_{th} \}.$$  

Apparently there is a one-on-one mapping between a threshold $\zeta_{th}$ and a crowdedness region. Given $\zeta_{th}$, the derived crowdedness region can...
be continuous or the union of a set of disjointed sub regions, depending on $\mathcal{H}_k(t)_{(i)}$ and $\zeta_{th}$. Each sub region contains at least one crowdedness spot, though not all crowdedness spots are necessary to be contained in crowdedness regions. We are interested in the top crowdedness spots (e.g., top 10 crowdedness spots) and the crowdedness regions containing these crowdedness spots.

**Definition 7:** Top-$k$ crowdedness region $R_{k}^{(t)}$ is the crowdedness region with $\zeta_{th}$ set as the k-th largest spot crowdedness of all crowdedness spots, i.e.,

$$R_{k}^{(t)} = \{ R_{s}^{(t)}(\zeta_{th}) \}$$

(10)

where $\zeta_{th}$ is the k-th largest crowdedness in $L_{s}^{(t)}$.

**Theorem 1:** $R_{k}^{(t)}$ is the smallest (in area) crowdedness region that contains the top-k crowdedness spots in the area.

**Proof:** We prove the theorem by contradiction. If an area contains the top-k crowdedness spots while is not the smallest crowdedness region, so there will be another crowdedness region in the area to be the smallest one. According to the definitions of crowdedness spot and top-k crowdedness region, it is contradictory. Hence, $R_{k}^{(t)}$ is the smallest (in area) crowdedness region that contains the top-k crowdedness spots in the area.

Throughout the rest of this paper we study the crowdedness spot issues via investigations on the top-k crowdedness regions.

**D. Temporal crowdedness spots**

In this subsection, we study temporal properties of crowdedness spots. At first we define correlation coefficient as,

$$\rho_{H_{1}, H_{2}} = \frac{T \sum h_{1}^{(t)} h_{2}^{(t)} - \sum h_{1}^{(t)} \sum h_{2}^{(t)}}{T \sqrt{ \sum (h_{1}^{(t)})^2 - (\sum h_{1}^{(t)})^2} \sqrt{ \sum (h_{2}^{(t)})^2 - (\sum h_{2}^{(t)})^2}}$$

(11)

where $H_{1} = \{ h_{1}^{(0)}, \ldots, h_{1}^{(T)} \}$ and $H_{2} = \{ h_{2}^{(0)}, \ldots, h_{2}^{(T)} \}$ are two given time series.

A higher correlation indicates more consistency and the upper limit of the correlation is one which happens only when the two time series are linearly dependent.

For each spot, we compute correlation coefficients between its crowdedness of a day and that of the day before. We make the computation for statistics over one month and compute the mean correlation coefficient for each spot. The CDF of mean of correlation coefficients are plotted in Figure 8. There are mainly three types of spots. About 15% spots have low temporal correlation around 0.2. About 45% spots have middle correlation ranging from 0.4 to 0.5. And the other 40% spots presents high temporal properties with coefficients around 0.8, indicating clear self-similarity of spot crowdedness over days. Hence, we can leverage this high temporal correlation to infer unexpected events.

As a case study, Figure 9 pictures the spot crowdedness as a function of time at Jiangwan Stadium on 22nd and 23rd Dec., 2006. Their differences over the time are also plotted. We can find that in most the time the two crowdedness curves are very similar. Dramatic deviation appears from 18:00PM, implying some unexpected events happened. Actually, there was a big show then and audiences were trying to aggregate in the stadium.
E. Evolutionary crowdedness regions

Let $A(Rs)$ denote area of a crowdedness region. We quantify spatial correlations of crowdedness regions by area difference ratio $\theta$. Given two crowdedness regions $R_{s1}$ and $R_{s2}$, their area difference ratio is calculated by

$$\theta_{R_{s1},R_{s2}} = 1 - \frac{A(R_{s1} \cap R_{s2})}{A(R_{s1} \cup R_{s2})}. \quad (12)$$

A small ratio indicates more spatial correlations and a large ratio implies more spatial drift of the region. Notice that $\theta_{R_{s1},R_{s2}}$ takes no spot crowdedness into account. Issues about regions crowdedness are left for future work.

Figure 10 depicts area difference ratio between $R_{s_k}^{(t)}$ and $R_{s_k}^{(t-1)}$ over one day with varying $k=2, 4, 10, 20$. We can find that the top-2 crowdedness region, which mainly characterizes the top-1 crowdedness spot, does not vary dramatically over the whole day. The maximal difference is no more than 1% which happens during noon from 12:00PM to 18:00PM. The top-4 crowdedness region presents more variations during the day time. The difference begins to arise at the morning peak time, maintains stable during working hours, and drops quickly during the afternoon peak time. This is similar for top-10 and top-20 crowdedness region but having even earlier arising time and later dropping time. These results reveal that the single top crowdedness spot is fairly stable, while the several top crowdedness spots have much larger dynamism. For top-20 crowdedness spots, their spatial drifts are up to 80% over a day.

In order to study the vehicle crowdedness in an region, we define region crowdedness as the average of spot crowdedness over all spots in the region. It is calculated by

$$H_{s_k}^{(t)}(Rs) = \sum_{i \in Rs} H_{s_k}^{(t)}(i) \quad (13)$$

Given two regions $R_{s1}$ and $R_{s2}$, their crowdedness difference ratio is defined as

$$\theta_{R_{s1},R_{s2}} = \frac{H_{s_k}^{(t)}(R_{s1}) - H_{s_k}^{(t)}(R_{s2})}{H_{s_k}^{(t)}(R_{s2})} \quad (14)$$

Figure 11 depicts $\theta$ between $R_{s_k}^{(t)}$ and $R_{s_k}^{(t-1)}$ over one day with varying $k=2, 4, 20$. As expected, after the morning traffic peak time, the crowdedness increases sharply. For $k=2$, the crowdedness increases by more than 60% in three hours, maintains stable during the working hour, while drops even sharply from 21:00PM to 24:00PM. Other of more spots present a similar trend of changing.
V. CROWDEDNESS SPOT ACQUISITION

Crowdedness spot can be considered as a higher level of feature retrieved from the taxi. Hence, we can furthermore utilize the crowdedness spot to study the taxi. For example, the taxis always cross crowdedness spots may be have more chances to capture the crowded areas’ information or pick up passengers, at the same time, these taxis’ behavior may help us give more investigation of the city transportation. In this subsection, we build SVM-based intelligent search to classify the taxis. Limited to the page space, we take the crowdedness taxi as a case study to demonstrate the crowdedness spot acquisition.

The flow chart in Figure 12 illustrates the crowdedness taxi intelligent search process. First, a domain expert (in this paper, the data providers) prepares the targeted taxi features, use them to create the learning data sets, and utilize the data sets to train and build the predictive model. Second, the targeted features are published to the users. Third, a user select a feature of interest to retrieve the relevant list of crowdedness taxis from a search engine. Fourth, the retrieved taxis are analyzed and classified by the predictive model. Finally, only the taxis that are scored as relevant are sent back to the user.

Definition 8: A taxi trace is a sequence of location samples from the GPS reports in the time order.

Definition 9: Crowdedness taxi is a taxi whose trace crosses a number of different crowdedness spots (the threshold is \( \lambda \)) in a given time interval \( t \).

If a taxi \( m \) crosses a crowdedness spot \( l_{S_i}^{(t)} \), we assign 1 to the taxi at this spot, while if not cross, 0 is assigned. Thus, we can construct a “0-1” vector for each taxi to record its trace, e.g., \( x_m^t = \{0,1,1,...,1\} \).

To build the crowdedness taxi classifier, we employ Support Vector Machine (SVM) [13]. Given a set of training vectors \( x_i \in R^a, i = 1, 2, ..., d \), where \( d \) is the number of vectors, in two cases, and a target vector \( y \in R^d \) such that \( y_i \in \{0,1\} \), both classifiers solve the following unconstrained optimization function.

\[
\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{d} \varphi(w; x_i, y_i) \tag{15}
\]

where \( C \) is a positive penalty parameter, \( \varphi(w; x_i, y_i) \) is a loss function.

In this paper, we choose \( L_2 \)-SVM to build our predictive model because of the great efficiency [14]. The evaluation of the method is provided in Section VI-C.

VI. FIELD STUDY EVALUATION

In this section, first we evaluate the crowdedness function by the real life data sets (taxi data and bus data), and second, to verify the effectiveness of our mobility-based clustering, we compare it with the current method on a number of field studies (field camera records) and empirical data sets. Last, we evaluate the intelligent SVM-based crowdedness taxi classifier.

A. Crowdedness function validation

Our task of validation is to evaluate which of the linear and statistical crowdedness functions produces a better-fit crowdedness distribution. We carry out the validation through two approaches. One is by conducting the prediction about vehicle speed. By this we look for an optimal configuration for mobility-based clustering. The second validation is to compare the optimal setting mobility-based clustering with traditional algorithms through field studies.

Figure 13 illustrates basic procedures of the first part of the validation. In this validation we randomly split the original data set into two subsets based on vehicle IDs. One sub data set is used as the input set (sensing report set), and the other one is used as the test set (denoted as \( \Phi^c \)). We ensure that input set and test set are disjoint.
We apply the two crowdedness functions on the input set to compute the crowdedness distributions. The parameter $\alpha_i$ is set as 0.5, and $\tau_i$ is set to 1 for validation convenience. The impacts of these two parameters are investigated in Section IV-A. With the crowdedness distribution available, we design a simple mobility prediction algorithm. Given the crowdedness measurement $H_L^{(t)}(i)$, the vehicles at spot $i$ are expected to have the speed $\tilde{v}_i^{(t)} = (H_L^{(t)})^{-1}(i)$ where $(H_L^{(t)})^{-1}$ is the inverse function of $H_L^{(t)}$. That of $H_S^{(t)}(i)$ is similar. Taxi in the test set $\Phi$ are used to be the representatives of the general vehicles. We evaluate the effectiveness of the derived crowdedness distribution by investigating the error between the predicted speeds and the real speeds, denoted as $\varepsilon$. We measure the Mean Absolute Difference (MAD) which is given by

$$MAD(\varepsilon) = \frac{1}{N} \sum_{\forall v_{k}(i) \in V_k^{(t)}} \varepsilon.$$

Figure 14 depicts $MAD(\varepsilon)$ with varying sample object sizes $N_s = 250$ to 2750. We can observe that the linear model $H_L$ is sensitive to the sample object size. The setting of 250 sample objects ($N_s = 250$) produces a mean of error up to 28 kmph. Note that the average speed in the downtown is around 30 kmph (due to space limitation, this is a relative error over 90%). As $N_s$ increases, the accuracy of the linear model is sharply improved. The setting of $N_s = 1200$ has the mean ($\varepsilon$) around 19 kmph with a relative error about 80%. There are, however, no obvious further improvements when $N_s$ continues to increase. Compared with the linear model, the statistical model is much more robust to the size of the sample set. The mean ($\varepsilon$) is 21 kmph, 23% better than that of linear model when $N_s = 250$. The error keeps decreasing when the sample set increases. When $N_s = 2750$, the error mean is only 15 kmph, which is 50% as the relative error. From the results, we can conclude that our method achieves better accuracy when the data scale up.

To test the robustness, we validate the crowdedness functions using not only the traces of taxis, but also traces of 3108 public buses.

It is notable that the bus traces are limited to fixed routes. Buses are a specific type of vehicle with more particular moving patterns. They have regular service routes, periodic moving schedules, and fixed stops. Hence, they are not good enough to help detect crowdedness spots. But they are capable of validating our method. We apply the statistical crowdedness function on both taxi and bus traces, deriving two crowdedness distributions. Results show that these two distributions are fairly consistent. The consistency is measured by correlation coefficient as defined in Section IV-D. A higher correlation indicates more consistency and the upper limit of the correlation is one which happens only when the two time series are linearly dependent. Note that the average correlation coefficient is up to 63%. Over 70% spots have the correlation coefficient $\rho$ between 0.6 and 0.8, and about 10% are more than 0.8. And about 20% spots have very low $\rho$ in between. It is mainly because most of these spots are bus stops and traffic reconnection sites. In these areas bus speeds are mainly determined by the site type and thus provide little information for spot crowdedness.

In Section IV, we study spot crowdedness in practice. As a consequence, we find two approaches to reduce the error of speed predicting. The first one is to characterize spots, and the second one is to profile sensor objects. In Figure 15, we report the evaluation of the first approach, and then we evaluate the impact from the characteristics of sensor objects. Figure 15 plots MAD of $H_S$ with the simple parameter setting, that is, $\alpha_i = 0.5, \tau_i = 1$ and with the advanced settings of dominating average weight and dominating period. We observe that with the improved setting, the error can be further reduced by 11% when $N_s = 600$, and by 15% when $N_s = 2400$. We conclude that the improved setting is more advantaged under low sample set scenarios, while the advantages are largely revoked by the larger sample set. With the state calibration by profiling the sensor objects, the prediction error is further reduced by about 16% when $N_s > 1200$. Hence we can conclude that our method achieves better accuracy when the data scale up.

B. Mobility-based clustering validation

In order to verify the effectiveness of our mobility-based clustering, we conduct a number of field studies. We setup video cameras
at predetermined sites in Shanghai, record the real traffic situations in fields, and then measure the crowdedness of these areas through an offline manner. The results are denoted as “real situation”. Note that in the real situation, we take the crowdedness as the density of vehicle in the area. On selecting the comparison method, we choose UMicro, one of the latest representative method for clustering uncertain data streams [12]. For UMicro, we assume that each taxi in the sample set represents 80,000 general vehicles, because our data set has records of 5631 taxis, accounting 0.3% of the two million of the total vehicle population in the city. For mobility-based clustering, the produced crowdedness is a relative value. It needs an appropriate scalar to generate the absolute vehicle densities. To obtain this scalar, we need to collect the real initial state of the spot for calibration. In practice, this calibration usually incurs relatively high cost. We argue, nevertheless, that this calibration is a single-run operation such that the high cost can be amortized over a long operation time.

Figure 16 depicts the real vehicle crowdedness, the estimated number by mobility-based clustering, and that by UMicro at the spot of Jiangsu Road over two hours (11:00 AM to 13:00 PM) on 15th, Jan, 2009. Recall the definition of correlation coefficient defined in Section IV-D. It can be used to measure the accuracy of the derived spot crowdedness. Moreover, we utilize precision, recall and F-score to give insight into the accuracy. Precision means the percentage of true crowdedness spots in the detected crowdedness spots. Recall means the percentage of detected true crowdedness spots in true crowdedness spots. F-score is the weighted harmonic mean of precision and recall [15].

Table II lists the comparison between mobility-based clustering and UMicro approach (Appr.). We evaluate the results by correlation coefficient (Corr.), precision (Pre.), recall (Rec.), and F-score. The data is from three locations (Loc.). The results in the table are the mean results of different thresholds (we test the methods in top-10, top-20, top-30, top-40 and top-50 crowdedness spots’ detection). It shows that at Jiangsu Road, the results of mobility-based clustering are 10 to 20 times better than UMicro under the same scenario. The interesting thing here is that UMicro performs pretty well at Rail Station. We believe this is because the Rail Station is a taxi aggregated area with high customer flows. Therefore, the samples at this spot are much larger than those at other spots, resulting in a much better accuracy by density-based approaches.

### C. Crowdedness spot acquisition validation

We employ one month data to evaluate our SVM-based intelligent classifying and searching the taxis. A domain expert, such as a transportation officer, can accurately and easily classify each taxi’s traces to either relevant or irrelevant to the taxi of interest (crowdedness taxi in this
TABLE III
EVALUATION RESULTS FOR THE PREDICTIVE CLASSIFICATION MODEL

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_2$-SVM</td>
<td>0.897</td>
<td>0.813</td>
<td>0.853</td>
</tr>
</tbody>
</table>

Fig. 17. Categorizing spots by dominating average weight and period

paper) through a quick inspection of its feature representation. 728 taxis (12.9\% of the total 5631 collected taxis) were classified as relevant to the crowdedness taxis. In the experiment, recall measures the ability of the model to retrieve a good rate of relevant taxis from the total original relevant taxis available, precision measures the rate of really relevant taxis from the total taxis that the model has claimed to be relevant, and F-score represents the overall look to both the precision and recall. In Table III, we report the experiment results. The results show that the precision is over 89\%, and the recall is over 81\%.

VII. FINDINGS AND DISCUSSIONS

In our crowdedness model, we introduce two parameters to describe the dynamics of the spot in a given location (\(\alpha_i\) and \(\tau_i\) in Eq. 4). The parameter setting of dominating average weight and dominating period is unique for each spot and thus can be utilized to characterize spots. A higher \(\alpha_i\) indicates lower dynamism of the spot and larger \(\tau_i\) indicates longer period of the spot crowdedness. We therefore categorize spots by presented dynamism and periods.

Figure 17 depicts \(\alpha_i\) and \(\tau_i\) of different spots in the weight and period plane. We find that the spots are highly aggregated with clustering features. They can be roughly grouped to five groups named from Group I to Group V. Mapping these groups with geographic information we have following findings.

- Group I presents high dynamism and frequent change of crowdedness. About 30\% spots, a major portion of the spots, belong to this group. Among these spots, over 93\% are traffic reconnection sites (e.g., rail station and bus connections);
- Group II also presents high dynamism but the crowdedness changing frequency is much lower than that of Group I. About 10\% spots fall into this category. Over 94\% of them are shopping and entertainment areas.
- Group III is mostly traveler’s interests, with 7% spots.
- Group IV accounts another major portion of the spots with over 40% spots. These spots have steady states and change gently, which are mostly freeways, and expressways.
- Group V presents very low dynamism and middle frequency of changing. About 8\% spots fall into this group which are mostly working, business and living areas.

Besides these five types, there are about 2\% outlier spots which can hardly be clearly categorized. An interesting finding is that spots in one group are not necessary to be the same geographic type. For instance, 7\% of Group I spots are not the traffic reconnection spots but present the same characteristic as the spots in Group I. For this we mainly have two hypotheses. One is that this is due to the inappropriate planning of the city development. In other words, these sites should be constructed as the traffic reconnection spots to feed the traffic requirements. Another is due to the limitation of our model which is unable to fully distinguish the characteristics of different sites. By current data we can hardly state which hypothesis is closer to the reality, while both of them are raised for planning of future city development.

VIII. RELATED WORK

Clustering: Object clustering is a well studied problem with a great deal of research efforts being devoted in. One of the most promising approaches for spatial static clustering can be found in the research work of DBSCAN [4], [16]. Recently, clustering moving objects is becoming a
crowdedness research issue. In the research work [7] Li et al. discussed the clustering of moving objects and extended the concept of micro-cluster. High quality moving micro-clusters are dynamically maintained which leads to fast and competitive clustering results. Chakrabarti et al. [9] proposed evolutionary clustering which is able to well deal with the mobile clusters in low dynamic environments. Chen et al. [16] proposed a framework D-Stream to efficiently identify outliers of clusters. Extensive efforts are also devoted to the clustering with uncertain data [17]–[19], data streams [20], uncertain data streams [8], [12], [18], complex event processing [18], [20], co-clustering on large data sets [18], ranking queries [11], [18], [21], and evolutionary clustering [9], [18]. The above existing works are density-based approaches. Especially in their study scenario, they all consider the density or quantity of the objects is enough to cluster. Thus when the density or quantity of the objects is not that good enough for special application scenarios, they will fail. In our application scenario, these methods do not work due to the new arising features, such as extreme less samples and notable data point location errors.

Traffic data analysis: Gaffney et al. [22] studied the problem of clustering trajectories and proposed to use short data sequences as object movements. Li et al. [23] studied traffic flow patterns in road networks and proposed a density-based algorithm called FlowScan. Kriegel et al. [24] introduced a statistical approach to describe the likelihood of any given individual in road networks to be located at a certain position and time. These works mainly focused on how to accurately measure and predict vehicle speeds while showing very limited insight into mobile vehicle clustering. Other proposals suggested to use dedicated sensors deployed on roads to perceive vehicle crowdedness. Coifman [25] studied how to detect freeway incidents by traffic detectors on roads. Some utilize vehicles equipped with GPS and/or a cellular positioning system as probes. Hollmén et al. [26] studied spatio-temporal road condition forecasting by Markov chains and artificial neural networks. Castro et al. [27] proposed a method to construct a model of traffic density based on taxi traces. Bacon et al. [28] tried to use real-time road traffic data to evaluate congestion. Yuan et al. [29] presented a Cloud-based system computing customized and practically fast driving routes for an end user using (historical and real-time) traffic conditions and driver behavior. Zheng et al. [30] detected flawed urban planning using the GPS trajectories of taxicabs traveling in urban areas. Liu et al. [31] proposed algorithms to construct outlier causality trees based on temporal and spatial properties of detected outliers. Yoon et al. [32] tried to detect traffic conditions on surface streets given location traces collected from on-road vehicles, such as GPS location data, and infrequent low-bandwidth cellular updates. The current work assumed that dedicated sensor devices had been deployed so that the collection of vehicle crowdedness becomes straightforward. In our work, we do not have dedicated sensors but employ mobile objects as “sensors” to perceive the crowdedness. The difference between our approach and the floating car [33] in Intelligent Transportation System (ITS) is that, floating car in ITS focus on the speed of vehicles, and for crowdedness spot issues, they have the common assumptions that each float car represent a number of real vehicles. In our problem however, taxis are not good representatives of other vehicles and therefore such approaches will fail.

IX. CONCLUSION AND FUTURE WORK

In this paper, we proposed mobility-based clustering, a novel approach to identify crowdedness spots in a highly mobile environment with extremely limited and biased object samples. The unique feature of mobility-based clustering is to use speed information to infer the crowdedness of moving objects. Furthermore, we study the crowdedness spot categories and the crowdedness taxi acquisition from the detected crowdedness spots. We evaluated the performance of mobility-based clustering based on real taxi data collected in the city through field studies. The future work can be conducted along following directions. First, in mobility-based clustering the speed information is critical. Due to the small sample data set, we used a simple approach to estimate the mobility of vehicles at the spot of no data. Better mobility estimation can produce better crowdedness values. Second, there are many factors besides the spot crowdedness that will have impact.
on the vehicle mobility, such as traffic lights and car accidents. We leave them for future work. Third, we need more field studies, though labor intensive, to further verify the effectiveness of the mobility-based approach. Fourth, the better road griding method is needed for retrieving much more precious locations. Finally, depending on other characteristics of moving objects, other non-densit-based clustering may be worth further investigations.

X. ACKNOWLEDGMENTS

This research was supported in part by HKUST-Nansha Research Fund NRC06/07.EG01, China NSFC Grants 60933011 and 60933012, and the National Basic Research Program of China (973 Program) under Grant No. 2006CB303000. This research is also supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

REFERENCES