2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing

How Long a Passenger Waits for a Vacant Taxi?

—Large-scale Taxi Trace Mining for Smart Cities

Guande Qi, Gang Pan*, Shijian Li, Zhaohui Wu
College of Computer Science
Zhejiang University
Hangzhou 310027, China
{qiguande, gpan, shijianli, wzh}@zju.edu.cn

Daqing Zhang, Lin Sun
HANDICOM Lab
Institut TELECOM & Maga.
SudParis, 91011 Evry Cedex, France
{daqing.zhang, lin.sun}@it-sudparis.eu

Laurence Tianruo Yang
Department of Computer Science
St. Francis Xavier University
Antigonish, NS, Canada
ltyang@stfx.ca

***Abstract*—To achieve smart cities, real-world trace data sensed from the GPS-enabled taxi system, which conveys underlying dynamics of people movements, could be used to make urban transportation services smarter. As an example, it will be very helpful for passengers to know how long it will take to find a taxi at a spot, since they can plan their schedule and choose the best spot to wait. In this paper, we present a method to predict the waiting time for a passenger at a given time and spot from historical taxi trajectories. The arrival model of passengers and that of vacant taxis are built from the events that taxis arrive at and leave a spot. With the models, we could simulate the passenger waiting queue for a spot and infer the waiting time. The experiment with a large-scale real taxi GPS trace dataset is carried out to verify the proposed method.**

***Index Terms*—Smart city; taxi trace data; arriving model; passenger's waiting time;**

I. Introduction

Smart city is featured with the utilization of information and communication technology (ICT) to achieve sustainable development of cities. Within these years, the development of pervasive sensing, communication, and computing technology makes it possible to collect, share, and understand the dynamics of a city. These techniques could also benefit the smart city applications in transportation, urban planning, public health, public security, and commerce [1].

Currently, there are tremendous amount of sensors scattered pervasively around our physical world. These sensors can record the states and behaviors of physical objects, individuals and the environment. This big data presents an opportunity and also a challenge, which is to extract semantics and knowledge from it, and to model and re-recognize the original physical world. Based on the understandings of data in cyber world, the real world could be impacted with applications in everyday life, industry and commerce.

With GPS devices installed, taxi, as an important urban vehicle, is generating massive trace data every day. For example, In Hangzhou, an 8-million-population city in China, 9,000 taxis take more than 0.8 million passengers and generate ten millions GPS records per day. In the Shanghai City, with a population of 23 millions, there are 50,000 taxis carrying three millions passengers while leaving sixty millions of records every day.

Such traces may be helpful for improving taxi services, which are provided usually in two manners: 1) passengers make an appointment for a taxi in advance; 2) passengers wait on sides of a street for a vacant taxi passing by. The first case is less frequently used; it usually occurs when people have a fixed schedule. This paper focuses on the second one and predicts how long a passenger will wait for a vacant taxi at a spot from trace data.

Waiting is common in the main type of taxi service. The reason that passengers have to wait for public transportation is that (1) public transportation is limited and cannot be available in any time; (2) mobility of passengers is not organized and may happen at any time. Unlike public vehicles, which have a regular route and schedule, taxi, as a kind of particular public vehicles, runs a customized route to meet personal transportation requirement. The personalization and customization of taxis results in uncertainty of taxi service and waiting time.

From a psychological view, waiting is a problem because waiting time could be long and uncertain. People dislike uncertainty; moreover, the longer they wait, the more anxious they become. Waiting may annoy passengers and impede customers to choose taxi as way of going out in (1) it brings uncertainty to the customers' travelling schedule; (2) waiting itself disturbs people while long time waiting debases service quality [2]; (3) waiting time may be treated as money in decision making [3].

The problem of waiting for taxis is more and more severe in China, led by the increasing population density and limited traffic facilities. Firstly, waiting time could be long. For example, in Beijing, during the rush hour, people have to wait for 20 minutes to take a taxi. Moreover, there can be tens of passengers waiting in a queue in places with high demand of transportation, such as coach/train stations.

Predicting wait time for vacant taxis could help passengers to reduce passengers' dissatisfaction. With waiting time prediction, (1) people would know how long they wait and won't be anxious when waiting; (2) people can decide not to waste their time if need long waiting; (3) people can rearrange their schedules to be punctual; (4) people could know where they can take taxis easily and find the best spot to wait.

It is not extensively investigated to predict the waiting time to find the next vacant taxi for a passenger. Traditional

*Corresponding author

978-0-7695-5046-6/13 $26.00 © 2013 IEEE
DOI 10.1109/GreenCom-iThings-CPSCom.2013.175

researches in the transportation community have ever exploited waiting time, along with taxi fare, disposable income, and occupied taxi journey time, to measure the quality of taxi service [4], [5]. However, traditional prediction is coarse-grained in that they calculate waiting time for large zones and considered no variation in a day. For example, in [5], they divide Hong Kong, whose area is 1,100 square kilometers into only 15 zones, and calculate a constant waiting time for each zone. Such calculation may not be appropriate since, according to our observation, actual waiting time varies with time and spots.

With increasing number of taxis in large cities equipped with GPS sensors, the emergence of taxi GPS data, which largely promotes researches on taxi services, may help prediction of waiting time. Previous work has shown taxi traces have promising applications in guiding drivers to find passengers [6], providing professional route navigation [7], and detecting abnormal taxi traces [8]. The data could provide information about traces of vacant taxis, which is very helpful for predicting waiting time of a passenger for a vacant taxi.

Given a spot and the time, the problem of predicting waiting time is non-trivial. If there is only a single passenger at the spot, the waiting time is just the time between his arrival and the arrival of the next vacant taxi. In such case, waiting time can be calculated simply with prediction of the arriving time of the next vacant taxi. However, as we can see in real world, there may be many passengers waiting and competing for taxis at the same spot. The challenge of the problem is to model the competitive behavior for predicting waiting time.

In this paper, we solve the problem by assuming that the competing strategy follows a first-come-first-served way; that is, passengers wait in a queue for vacant taxis. Contributions of this paper are as follows:

1) We investigate the problem of predicting waiting time for a passenger to take a taxi at a spot in a competitive environment, which has never been addressed before. The prediction considers not only the spatial and temporal variation, i.e., it is closely related to where and when the passenger wants to take a taxi; but also considers the competitions of passengers when taking taxis.
2) We build models for the arriving of passengers and vacant taxis. Taxi traces contain direct information about arrival of vacant taxis, and also pick-up events that relate to passengers' arrival.
3) We present a prediction algorithm to calculate waiting time with taxi traces, given the spots, taxi traces and road map. Firstly, we extract sequence of events (arrival and departure) for each spot by the intersections of traces and spots. Then, historical events are used for training arrival model of passengers and vacant taxis. Finally, the arrival models are used for predicting waiting time with parametric and nonparametric method.

II. Problem Description

Waiting time of passengers for vacant taxis could be much different when varying time and spot. Human activity exhibits patterns that vary with time and place [9]. Such variation leads to difference of human mobility demands in different spots. For example, more passengers move from suburban district to downtown in the morning while they leave at night. It finally causes the temporal and spatial variance of waiting time. Thus, we describe the problem as follows.

Problem: *given taxi trace data Tr from history and a spot r, predict the expectation of waiting time wt a passenger needs to take a vacant taxi in the spot at time t. From the mathematical view, the problem is to solve the function $wt = f(t, r)$, given history data Tr.*

If there is only one passenger requiring a taxi at the spot r, he/she can be served immediately after the next vacant taxi arrives; the waiting time equals the serving time $wt = t_s$. However, a spot usually has a few passengers waiting for a taxi, especially in rush hours. We assume that the passengers obey the first-come-first-served rule. For this case, the waiting time equals the time in a queue t_q plus the serving time t_s: $wt = t_q + t_s$.

Therefore, waiting time could be calculated by modeling a queue for each spot r. The state of a queue is represented with its length, which increases/decreases with the arrival of a new passenger/vacant taxi. The expectation of queue length l can be modeled with arriving rate of passengers μ and vacant taxis λ. Then, since the arriving rate for vacant taxis is λ, waiting time is calculated as:

$$wt = l/\lambda = \frac{g(\mu, \lambda)}{\lambda}.$$

The form of $g(\mu, \lambda)$ depends on how we model the arrival of passengers and vacant taxis. When they are modeled as Poisson processes,

$$g(\mu, \lambda) = \frac{\lambda}{\mu - \lambda}.$$

The challenge for the problem turns to the estimation of arriving rate of passengers and vacant taxis, which is an optimization problem:

$$(\lambda, \mu) = argmax_{\lambda,\mu} P(Tr|\lambda, \mu),$$

where Tr is the set of traces. In detail, the two parameters are optimized separately. We use non-homogeneous Poisson process to model arriving processes of passengers and vacant taxis. Then, the model is solved by maximizing the probability of generating the arriving processes in history trace data.

It is not difficult to calculate λ because the arriving of vacant taxis could be extracted from the taxi trace data. However, it is non-trivial to calculate μ, since there is no information on passengers' arriving in the taxi trace data. In this paper, we solve the problem through estimating passenger arrivals with pick-ups in the sequence of events in each spot. We define the concept **event sequence**, which depicts the sequential events of taxi arrival and departure in each spot:

$$Seq_e = e_1 \rightarrow \cdots \rightarrow e_i \rightarrow \cdots \rightarrow e_n;$$

where each event $e_i = (t_a^i, s_a^i, t_l^i, s_l^i)$ denotes that a taxi with status s_a (occupied: 1, vacant: 0) arrives at t_a and leaves at time t_l with status s_l.

The event sequence contains information about the arriving of vacant taxis (a vacant taxi comes and leaves with vacant state) at a spot, which is used for vacant taxis' arriving model; and about the pick-up event (a vacant taxi comes and leaves with occupied state) at a spot, which is used for modeling passengers' arriving.

The event sequence can be extracted from taxi traces. Movement of a taxi forms a trace, which is recorded in trace data by sampling in discrete time series. Each trace of a single taxi is represented as a time-ordered sequence of GPS samplings:

$$tr : p_1 \to \cdots \to p_i \to \cdots \to p_n;$$

where $p_i \triangleq (t, long., lat., s)$. Trace data is a set of taxi traces: $Tr = \{tr_i\}$.

III. Predicting Waiting Time

Waiting time prediction is to calculate the time a passenger would wait to take a taxi after he arrives at a spot at a time. To solve the problem, we use event sequences to learn the arrival model of vacant taxis and passengers. Based on the arrival model, we develop a parametric method to calculate waiting time from queuing theory, and propose an improvement of the method by exploiting nonparametric model of vacant taxi arrival.

A. Arrival Model: Non-homogeneous Poisson Process

The statistics of arriving number $N(t)$ of vacant taxis or passengers until time stamp t could be denoted as a counting process$\{N(t),\, t \geq 0\}$, which counts the number of events $N(t)$ that occur in a given time interval $[0, t]$. In this paper, for each spot r, such counting process is modeled with Non-homogeneous Poisson process (NHPP):

Counting process $\{N(t), t \geq 0\}$ is called a non-homogeneous Poisson process (NHPP) when the arriving rate changes with time $\lambda(t)$. $\forall s, t,\ N(t+s) - N(s) \sim Pois(\mu(t)\,t)$:

$$P\{N(t+s) - N(t) = n\} = e^{-\mu(t)t}\frac{(\mu(t)\,t)^n}{n!},$$

$n = 0, 1, \cdots;\ \mu(t) = \frac{1}{t}\int_s^{t+s}\lambda(y)\,dy$ [10].

Moreover, we assume arriving rate in NHPP to be a piecewise constant function in our model: $\mu(t) = \lambda_i,\ t_i \leq t < t_{i+1},\ t_1 = 0$. Thus, the counting process $\{N(t),\, t_i \leq t < t_{i+1}\}$ of arriving passengers and vacant taxis for each spot r could be assumed as a general Poisson process with arriving rate $\lambda_{i,r}$:

$$P\{N(t_i + \triangle t) - N(t_i) = n\} = e^{-\lambda_{i,r}\triangle t}\frac{(\lambda_{i,r}\triangle t)^n}{n!},$$

$n = 0, 1, \cdots$.

B. Model Solution: Mining Traces for Arrival Rate

This section uses event sequences to estimate arriving rates of vacant taxis and passengers. The event sequence for a spot depicts the arriving and leaving of taxis including their status, it also contains information of pick-up event (a vacant taxi comes but leaves with occupied status). The relation between such event sequences and a model for arriving of vacant taxis is obvious. Arriving of passengers is not directly known, but can be inferred from pick-up events in these sequences.

For each spot r and each time slot $[t_i, t_{i+1})$, we estimate the parameter $\lambda_{i,r}$ in our model, given the event sequence for the spot: $Seq_e = e_1 \to \cdots \to e_i \to \cdots \to e_n$.

With the event sequence, we can estimate the arriving rate $\mu_{i,r}$ in NHPP for the model of arriving vacant taxis directly. Firstly, for each spot r, the sequence of arriving vacant taxis is extracted from the event sequence Seq_e by retaining elements with $s_a = 0$ (a taxi arrives with empty state); the new sequence could be expressed as $Seq_v = av_1 \to \cdots \to av_j \to \cdots \to av_n$; $av_j = (t_a^j, 0, t_l^j, s_l^j)$. Then, the unbiased estimation of the arriving rate of vacant taxis can be calculated given each time slot $[t_i, t_{i+1})$:

$$\mu_i = \frac{k-j}{t_a^k - t_a^j} \tag{1}$$

where the jth (kth) event av_j (av_k) is the arrival of the first (last) vacant taxi in the time slot; $t_a^{j-1} < t_i \leq t_a^j \leq t_{i+1}$; $t_i \leq t_a^k \leq t_{i+1} < t_a^{k+1}$.

We estimate the arriving rate in NHPP for the model of passengers' arriving by firstly calculating the occurrence rate of pick-up events $\lambda_{i,r}$. A pick-up event is that a vacant taxi comes but leaves with passenger: $s_a = 0, s_l = 1$. For each spot, we extract the sequence of pick-up events Seq_p from the event sequence Seq_e of the spot: $Seq_p = pe_1 \to \cdots \to pe_j \to \cdots \to pe_n$, $pe_j = (t_a^j, 0, t_l^j, 1)$. With the pick-up event sequence, the pick-up time is calculated as $pt_j = \left(t_a^j + t_l^j\right)/2$; the pick-up interval between the jth and kth pick-up event is $pt_k - pt_j = \frac{\left(t_a^k + t_l^k\right) - \left(t_a^j + t_l^j\right)}{2}$. Then, the rate of pick-ups is estimated with:

$$\lambda_i = \frac{2\,(k-j)}{\left(t_a^k + t_l^k\right) - \left(t_a^j + t_l^j\right)} \tag{2}$$

where $t_a^{j-1} < t_i \leq t_a^j \leq t_{i+1}$; $t_i \leq t_a^k \leq t_{i+1} < t_a^{k+1}$.

Finally, we could prove that our unbiased estimation $\lambda_{i,r}$ for the rate of pick-ups is also an unbiased estimation for the arriving rate of passengers. This is proved with the equality of their expectation. As illustrated in Fig. 1, the jth passenger comes at t_j; the kth passenger comes at t_k; which are both unknown in trace data. However, we can observe the jth pick-up in pt_j and the kth pick-up in pt_j. Moreover, note that the waiting time is just the time interval between a passenger arrives and takes a taxi; thus we have $wt_j = pt_j - t_j$ and $wt_k = pt_k - t_k$, which are assumed to obey a same distribution. Expectation of $wt_j - wt_k$ is 0; thus the interval between two pick-ups and the arriving interval of these two passengers has the same expectation. Therefore, their reciprocals, the arriving rate of passengers and the occurrence rate of pick-ups also have the same expectation and our estimation is also unbiased for the arriving rate of passengers.

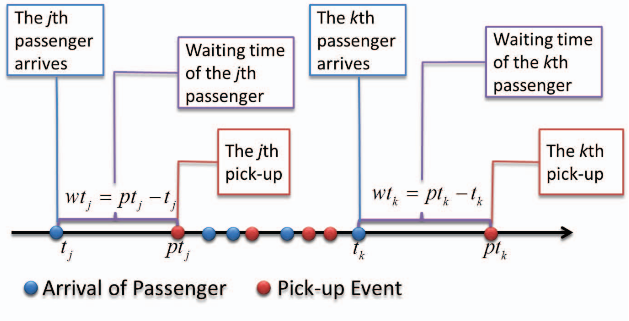


Figure 1: Illustration of passenger's arriving and pick-up during two adjacent vacant events.

C. Prediction with Arrival Model

Based on the model for arrival of passengers and vacant taxis, two methods are used to predict waiting time: parametric method and nonparametric method. For the parametric method, we calculate waiting time with queuing theory, given arriving rate of passengers and vacant taxis. For the nonparametric method, we need the model of passengers' arrival and historical data of taxis' arriving, which includes more information than the theoretic model. Then we generate a waiting queue and calculate the waiting time.

1) Predicting with Parametric Method: Based on the Poisson processes of arrivals of passengers and vacant taxis, our parametric method uses queuing theory to predict the waiting time. Assume that the mean for arriving rate of passengers is λ, the mean for arriving rate of vacant taxis is μ, and the probability for the state of the queue, namely that n passengers wait for taxis at a spot, is P_n. The state of the queue will be changed when a passenger arrives ($n+1$) or a vacant taxi arrives ($n-1$). After a long time, the waiting queue will become stable; such that the rate the queue leaves a state equals the rate the queue turns into the state:

$$\begin{cases} \lambda P_0 = \mu P_1 \\ (\lambda + \mu) P_n = \lambda P_{n-1} + \mu P_{n+1}, \; n > 0 \end{cases} \tag{3}$$

P_n could be solved with Eq. 3 by appending the basic assumption that $\sum_i P_i = 1$:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \tag{4}$$

The expectation of the queue length is:

$$L = \sum nP_n = \frac{\lambda}{\mu - \lambda} \tag{5}$$

The expectation of waiting time is:

$$wt = L/\lambda = \frac{1}{\mu - \lambda} \tag{6}$$

2) Predicting with Non-parametric Taxi Arrival: Since arrivals of vacant taxis could be extracted from trace data, which are closer to reality than those modeled with Poisson process, we promote a nonparametric method that improves parametric method with actual data of vacant taxi arrivals in history. In this method, we estimate the parameter of Poisson process for passengers' arrival (λ) as above and generate the arrival of passengers. Then we calculate the waiting time as the time between arrival and pick-up of each passenger.

How to generate the arrival passengers? Firstly, the time interval between events in Poisson process follows exponential distribution; the time t between the arrivals of two passengers follows an exponential distribution: $t \sim exp(\lambda)$:

$$f(t) = \lambda e^{-\lambda t}, \; t \geq 0.$$

Secondly, we must ensure that each passenger have arrived before he is picked up. The pick-up event is $Seq_p = pe_1 \rightarrow \cdots \rightarrow pe_j \rightarrow \cdots \rightarrow pe_n$, $pe_j = (t_a^j, 0, t_l^j, 1)$. Denote the arrival time of passengers (tp) as a sequence: $tp_1 \rightarrow \cdots \rightarrow tp_j \rightarrow \cdots \rightarrow tp_n$, then $tp_{j-1} \leq tp_j \leq pe_j.t_a$. Thus, arrival of the first passenger is generated randomly before the first pick-up. Arrival of the jth ($j > 1$) passenger follows arrival of the $j-1$th passenger; the interval $t = tp_j - tp_{j-1}$ is drawn randomly from the truncated exponential distribution:

$$f(t) = \frac{\lambda e^{-\lambda t}}{\int_0^{pe_j.t_a - tp_{j-1}} \lambda e^{-\lambda t}}, \; 0 \leq t \leq pe_j.t_a - tp_{j-1}.$$

Then, waiting time is calculated as:

$$wt = \frac{1}{n} \sum (pe_j.t_a - tp_j).$$

D. Algorithm Overview: Predicting Waiting Time from Traces

```
Prediction Algorithm
Input: ${Tr_d}$, $Tr_d : p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$,
$G = \langle V, E \rangle$, $V = \{v_i\}$, $v_i = (x_i, y_i)$, $E = \{\{v_i, v_j\}\}$;
spots $\{r\}$, each spot is a polygon $r = (v_1, v_2, \cdots, v_n)$.
1: Mapping Traces: $\{Tr_d\} \Rightarrow \{Tr_c\}$;
   $Tr_d : p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$; $Tr_c : s \rightarrow v_1 \rightarrow \cdots \rightarrow v_n$
   $\rightarrow d$, s.t. $v_i \in V$ and $\{v_i, v_{i+1}\} \in E$
2: Extract Event Sequence for $r$: $\{Tr_c\} \Rightarrow \{Seq_e\}$
   $Seq_e = e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow e_n$, $e_i = (t_a^i, s_a^i, t_l^i, s_l^i)$, $t_a^i$ and
   $t_l^i$ is the time a trace intersects the spot $r$.
3: Estimate Vacant Taxis'Arrival Rate: $\{Tr_c\} \Rightarrow \{\mu_i\}$
   given $Seq_e$, retains the event of vacant taxi's arrival as $Seq_v$
   $= av_1 \rightarrow av_2 \rightarrow \cdots \rightarrow av_n$, $av_j = (t_a^j, 0, t_l^j, s_l^j)$; for the $i$th
   time slot $[t_i, t_{i+1})$, find the first ($j$th in $Seq_v$) and last ($k$th
   in $Seq_v$) arrival of vacant taxi; arrival rate of vacant taxis is:
   $\mu_i = \frac{k-j}{t_a^k - t_a^j}$.
4: Estimate Passengers' Arrival Rate: $\{Tr_c\} \Rightarrow \{\lambda_i\}$
   given $Seq_e$, retains the pick-up event as $Seq_p = pe_1 \rightarrow pe_2$
   $\rightarrow \cdots \rightarrow pe_n$, $pe_j = (t_a^j, 0, t_l^j, 1)$; for the $i$th time slot
   $[t_i, t_{i+1})$, find the first ($j$th in $Seq_p$) and last ($k$th in $Seq_p$)
   pick-up; arrival rate of passengers is: $\lambda_i = \frac{2(k-j)}{(t_a^k + t_l^k) - (t_a^j + t_l^j)}$.
5: Calculate Waiting time:
   Parametric method: for the $i$th time slot, $wt_i = \frac{1}{\mu_i - \lambda_i}$;
   Nonparametric method: for the $i$th time slot, generate arrival
   of passengers as the sequence: $tp_1 \rightarrow tp_2 \rightarrow \cdots \rightarrow tp_n$;
   $wt_i = \frac{1}{n} \sum (pe_j.t_a - tp_j)$.
Output: waiting time $wt_{i,r}$ for time slot $[t_i, t_{i+1})$, and spot $r$.
```

Based on the model and methods above, we propose the whole algorithm for predicting waiting time, given taxi traces, road map and spots. In this algorithm, for each spot, the event sequence (taxi arrives at and leaves a spot) is extracted for estimating arrival models and calculate the waiting time.

Firstly, to calculate when a taxi arrives at or leaves a spot, we need to calculate its actual traces. Trace data, which is discrete samplings of traces, could be used for inferring actual traces by mapping to road network. Road network is represented as a graph ($G = \langle V, E \rangle$). $V = \{v_i\}$, $v_i = (x_i, y_i)$ represents an end point of a road segment. $E = \{e_{i,j}\}$, $e_{i,j} = \{v_i, v_j\}$ represents the road segment with v_i and v_j as end points.

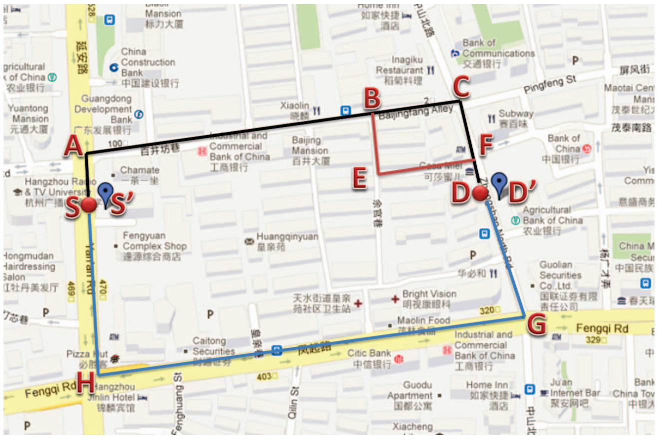


Figure 2: Illustration of trace mapping.

Figure 2 illustrates the process of trace mapping, in which we find the shortest path on the map as the continuous trajectory (algorithm 1 in [11]). Firstly, points in the trace $Tr_d : p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_i \rightarrow \cdots \rightarrow p_n$ are mapped to the nearest points on road $p'_1 \rightarrow p'_2 \rightarrow \cdots \rightarrow p'_i \rightarrow \cdots \rightarrow p_n$' . Then, the nearest route between each two points are found and expressed as $p'_i \rightarrow v_1 \rightarrow \cdots \rightarrow v_i \rightarrow \cdots \rightarrow v_n \rightarrow p'_{i+1}$, where $v_i \in V$ and $\{v_i, v_{i+1}\} \in E$. Finally, by connecting all segments between p'_i and p'_{i+1}, the trace is retrieved: $Tr_c : p'_1 \rightarrow v_1 \rightarrow \cdots \rightarrow v_i \rightarrow \cdots \rightarrow v_n \rightarrow p'_n$.

In the step of extraction of event sequence, we calculate the time a taxi arrives at and leaves a spot with the intersections of the continuous trajectory and a spot. The intersection is got by assuming that taxi moves uniformly during the trace between each two adjacent points. The status of a taxi when it arrives at (leaves) a spot equals to the status of the nearest record (in time) among its trace data.

An ideal output for prediction is a function of waiting time value wt which varies with passenger's arriving time t for each spot r: $wt = f_r(t)$. In this paper, we consider a simpler version; we divide a whole day into several time segments and assume that waiting time changes little in each segment. Thus this function is piecewise constant: $wt_i = c_{r,i}$, where r is the spot and i is the ID of time slot which arriving time t lies in: $t \in [t_i, t_{i+1})$.

IV. Experiment

A. Dataset

Taxi trace data is from more than 7,000 taxis in the Hangzhou City (of 8 million residents), which is generated by GPS devices installed on taxis with a sampling time of nearly 1 minute. The dataset has 200 billion records from April 1st 2009 to April 20th 2010. Each record contains the following fields:

- VEHICLE_ID: unique ID of the taxi in the dataset;
- LONGITUDE: current longitude of the taxi;
- LATITUDE: current latitude of the taxi;
- STATE: current status (occupied/vacant) of the taxi;
- TIME: sampling timestamp in the format of "YYYY-MM-DD HH:MM:SS".

B. Setup

Our experiments are carried separately for test cases, namely each time slot of the test day in each spot. Each test case includes a predicted value and the ground truth. Prediction error is the absolute difference of predicted value and ground truth.

We predict and test for each spot and each time slot separately. The spots are defined by clustering pick-up positions. Road segments were used previously as spots [12]. But for segments longer than 1 kilometer, passengers at two ends of the segment have different waiting times. Considering such disparity of waiting time, a revised DBSCAN method [13] is used to cluster the pick-up positions; each cluster's convex hull then defines a spot. 913 spots are retrieved. The time slot is defined by dividing a whole day into slots with same length.

For each spot, the predicted value is calculated by training our methods with trace data from time slots in days before the test day. The length of the time period between the training data and prediction data is prediction length. The number of days used for training is the length of training data. For example, given 10 days as the length of training data and the prediction length, when we predict the 33th day of the whole 385 days, we use data from the 14th day to the 23th day.

The ground truth is calculated by simulation. The simulation method is similar to the non-parametric method. For each test case, we simulate and generate passengers with the Poisson distribution learned from test data. With the data of arriving taxis in the test day, we then calculate the waiting time as test value.

With the prediction error, the performance of our methods is analyzed by: (1) optimizing values of parameters, including length of training data, and length of time slot; (2) comparing the performance of the two methods, namely parametric and nonparametric method; (3) going deep into the performance of the method with the distribution of the prediction error and the long-term prediction (predicting more than one day after training data) error; (4) showing how time and region ID affects the prediction error.

C. Results and Analysis

1) Performance Analysis: Firstly, two important parameters, namely the length of training data, and the length of time slot, are adjusted. The length of training data is important since the fewer time length of data for training, the less cost to build a prediction model, while possibly worse prediction result. Length of time slot is respected with the basic assumption that the waiting time is a piecewise function, constant in a time slot.

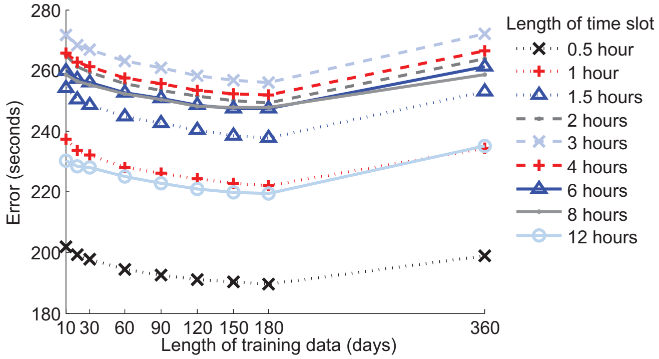
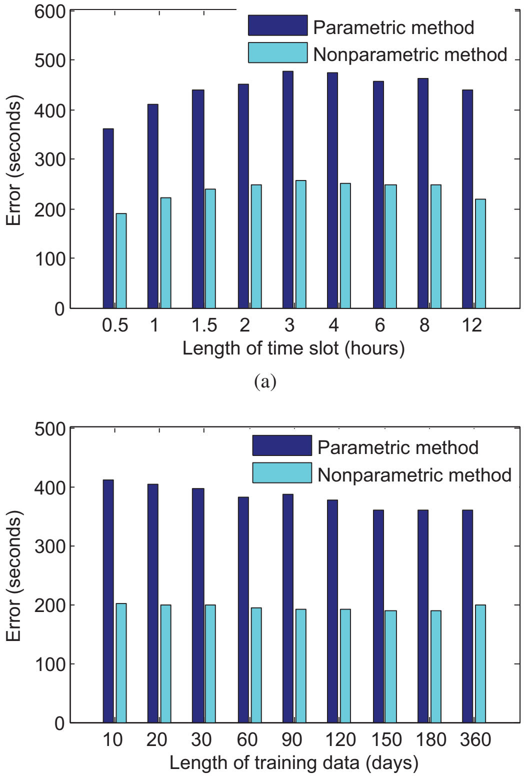


Figure 3: Prediction error with each parameter combination.



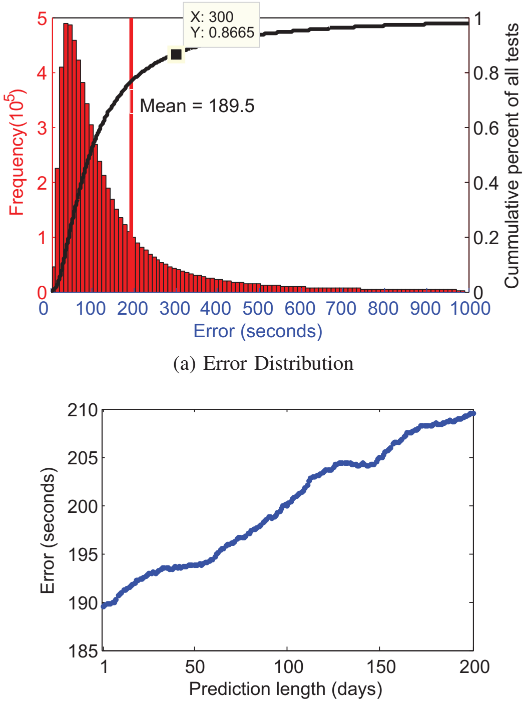
(a)

(b)

Figure 4: Comparison of the two methods under (a) different lengths of time slot; (b) different lengths of training data.

For both of the two methods (parametric and nonparametric method), we vary the length of training data from ten days to nearly one year, and the length of time slot from half an hour to half a day. For the nonparametric method, the prediction error with each parameter combination is shown in Fig. 3. Within these combinations, the prediction error is minimized when the length of time slot is half an hour and 180 days' data is used for training. This parameter combination also optimizes the parametric method.

The two methods, used for prediction of waiting time, are compared by performance under varied parameter values. Firstly, we fix the length of training data to the best parameter value (180 days) and change the length of time slot. The prediction error of the two methods is shown in Fig. 4a. The result indicates that the nonparametric method reduces the error of parametric method to half for different lengths of time slot. Secondly, the length of time slot is fixed to be the best (half an hour) and the length of training data is varied. As shown in Fig. 4b, the performance of nonparametric method is always better than the parametric method for the varied lengths of training data. The nonparametric method is better since we use arrivals of vacant taxis extracted from trace data, which are closer to reality than those modeled with Poisson process in parametric method.



(a) Error Distribution

(b) Long-term prediction error

Figure 5: (a) Distribution of the error with the nonparametric method under the best parameter combination; (b) illustration of long-term prediction error.

2) Error Distribution and Long-term Performance: This experiment is to further evaluate performance of our methods by error distribution and long-term prediction error. We use the nonparametric method under the best parameter combination to calculate the error distribution. The result in Fig. 5a shows that the mean error is around 3 minutes and the standard deviation of the error is about 10 minutes. Moreover, 86.65% of the test cases have error less than 5 minutes. For the long-term prediction, the best parameter combination is also used in nonparametric method; the prediction length is varied from one day to 200 days. As Fig. 5b shows, the error of prediction changes from 189.5 seconds to 209.5 seconds when we increase the prediction length. Thus our method is rather stable even when predicting waiting time of 200 days later.

3) Error Analysis: We analyze how time and space affects prediction error in this section. Firstly, we use 180-days data for training nonparametric method with different lengths of time slot, and calculate prediction error of each test day (from the 181st day to 385th day in data). As shown in Fig. 6, our result shows that prediction error changes with time and follows a certain variation pattern that is similar under different parameter combinations (length of slot is half an hour and two hours).

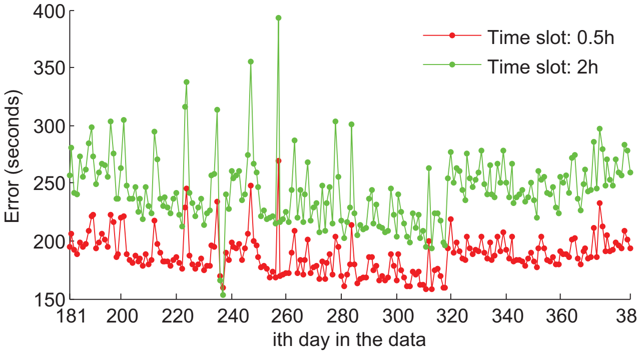


Figure 6: Variation of prediction error with time.

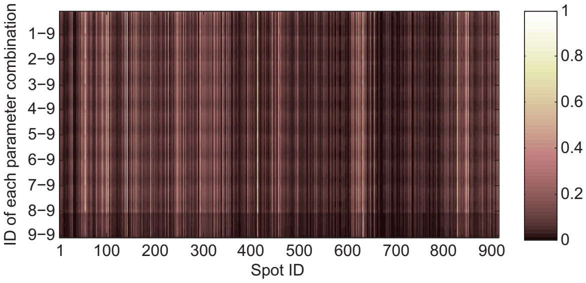


Figure 7: Grey-scale image of error for each spot under different parameter combination. Each row of pixels is the prediction error for all spots under a parameter combination. Label of the Y-axis is the parameter combination, which is written as 'length of training data'-'length of time slot'.

Secondly, we use nonparametric method under different parameter setting to calculate prediction error for each spot, which is then normalized in each parameter setting. The grey-scale map of the normalized error is shown in Fig. 7, with every row of pixels representing normalized prediction error for all spots with corresponding parameter combinations. This figure shows in each row the grey-scale of pixels change according to a pattern. Moreover, such pattern is similar in different rows. This means that error varies with spots and such variation pattern is similar for all parameter setting, which means that the prediction error is related to where passengers wait.

V. Related Work

We summarize previous works related to this paper as the following three classes:

A. Improving Taxi Service

With taxi GPS traces, more and more researchers are devoted to helping taxi service system by (1) recommending how to find passengers: [14] consider the strategy of finding passengers with hunting or waiting, they mine the strategy for regions and rank the strategy for recommendation. (2) recommending pick-up position: several works [15] try to improve efficiency of taxi service by reducing the time a taxi spends on finding passengers. They find the spatial-temporal pattern of pick-ups; detect the hotspots where passengers are picked up frequently, and recommend some hotspots to vacant taxi drivers or passengers. (3) recommending waiting position: [12] recommends the place to wait for a taxi based on waiting time prediction. [16] presents a recommender system for both taxi drivers and people expecting to take a taxi with by recommending both the place to pick up a passenger and to wait for a taxi. (4) monitoring service quality: [8] proposes a method to detect anomalous taxi trajectories; such abnormal traces could be used to monitor the misbehavior of taxi drivers. [1], [17] reviewed the research issues and potential applications of traces for smart cities.

B. Model for Arriving Processes

The statistics of arriving number of vacant taxis or passengers during a given time interval could be denoted as a counting process. Traditionally, counting process for such arrivals was often assumed to be Poisson process [18], [19] with constant arriving rate. However, with more and more traffic data collected, researchers observe that human activity varies with time and arrivals of passengers or transportation tools are actually non-homogeneous in time. The arriving of vacant taxis is verified to obey non-homogeneous Poisson process (NHPP) in previous work [12], namely the arriving rate is changes with time. Arriving of passengers for taxi service has similar behavior with passengers of other traffic tools, which is also modeled with NHPP [20].

C. Predicting Waiting Time

The work about prediction of waiting time is also important and attracts researchers. Traditional researches can only calculate a general value of waiting time for a large area constant at every day and waiting time is then treated as a measure of service quality of taxi service under different regulations [21], [22]. As an example, [5] builds macroscopic models to investigate how regulation affects the demand-supply equilibrium of taxi service. In their research, the city Hong Kong is divided into 15 districts and the origin-destination demand is considered based on these districts. These researches calculate a coarse-grained waiting time and neglect the variation of it with time and spots because the lack of real world data.

Nowadays, with the emergence of taxi trace data, the arriving number of vacant taxis and passengers is modeled and predicted; while the arriving process is unexplored. [23], [24] predict the number of vacant taxis in a given area based on time of the day, day of the week, and weather condition. [25] investigates human mobility patterns in an urban taxi transportation system and predict the number of passengers that would arrive at a spot in a given time, with historical pick-up number. [26] also presents online predictions regarding the spatial distribution of passenger demand throughout taxi stand networks.

Recently, [12] mined taxi GPS traces to predict the waiting time under the condition that only a single passenger waits on road. They calculated the waiting time with the time between arrival of the passenger and the next vacant taxi. However, there may be more than one passengers waiting on road; which is investigated in details in our paper.

VI. Conclusion

Smart city is featured with using ICT infrastructure to achieve sustainable urban development. Pervasive sensing technology gathers massive data about dynamics of a city. Mining from these sensor data to extract semantics about city dynamics is important both for understanding a city and for developing smart city applications.

As an example of mining trace data for improving traffic service, this paper proposes an approach to solve the prediction problem of waiting time for a passenger at a spot. The approach models the competition of passengers when waiting for taxis and considers the variance of waiting time with time and spots. We develop an algorithm to mine taxi traces, build the arrival model of passengers and vacant taxis, and predict waiting time. The performance is validated by a large-scale real-world taxi trace dataset.

Acknowledgment

This work was partly supported by Qianjiang Talent Program of Zhejiang (2011R10078), National Key Basic Research Program of China (2013CB329504), National Key Technology R&D Program (2012BAH94F01), the Fundamental Research Funds for the Central Universities, and Microsoft Research Asia Academic Program.

References

[1] G. Pan, G. Qi, W. Zhang, S. Li, L. Yang, and Z. Wu, "Trace analysis and mining for smart cities: issues, methods, and applications," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 120–126, 2013.

[2] L. Dube-Rioux, B. Schmitt, and F. Leclerc, "Consumers' reactions to waiting: when delays affect the perception of service quality," *Advances in Consumer Research*, vol. 16, no. 1, pp. 59–63, 1989.

[3] F. Leclerc, B. Schmitt, and L. Dube, "Waiting time and decision making: Is time like money?" *Journal of Consumer Research*, vol. 22, no. 1, pp. 110–119, 1995.

[4] H. Yang, Y. Lau, S. Wong, and H. Lo, "A macroscopic taxi model for passenger demand, taxi utilization and level of services," *Transportation*, vol. 27, no. 3, pp. 317–340, 2000.

[5] H. Yang, S. Wong, and K. Wong, "Demand–supply equilibrium of taxi services in a network under competition and regulation," *Transportation Research Part B: Methodological*, vol. 36, no. 9, pp. 799–819, 2002.

[6] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in *Proceedings of the 13th International Conference on Ubiquitous computing*, Beijing, China, September 2011, pp. 109–118.

[7] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: enhancing driving directions with taxi drivers' intelligence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 220–232, 2013.

[8] D. Zhang, N. Li, Z. Zhou, C. Chen, L. Sun, and S. Li, "ibat: detecting anomalous taxi trajectories from GPS traces," in *Proceedings of the 13th International Conference on Ubiquitous computing*, Beijing, China, September 2011, pp. 99–108.

[9] L. Liao, "Location-based activity recognition," Ph.D. dissertation, University of Washington, 2006.

[10] S. Ross, *Introduction to probability models*. Salt Lake City, Utah: Academic press, 2009.

[11] C. White, D. Bernstein, and A. Kornhauser, "Some map matching algorithms for personal navigation assistants," *Transportation Research Part C: Emerging Technologies*, vol. 8, no. 1, pp. 91–108, 2000.

[12] X. Zheng, X. Liang, and K. Xu, "Where to wait for a taxi?" in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, Beijing, China, August 2012, pp. 149–156.

[13] G. Pan, G. Qi, Z. Wu, D. Zhang, S. Li *et al.*, "Land-use classification using taxi GPS traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 113–123, 2013.

[14] B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi, and Q. Yang, "Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset," in *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Seattle, WA, USA, March 2011.

[15] J. Lee, I. Shin, and G. Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *Fourth International Conference on Networked Computing and Advanced Information Management, 2008. NCM'08.*, Gyeongju, Korea, September 2008, pp. 199–204.

[16] N. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: a recommender system for finding passengers and vacant taxis," 2012.

[17] P. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi GPS traces to social and community dynamics: a survey," *ACM Computing Surveys*, 2013.

[18] P. Marguier and A. Ceder, "Passenger waiting strategies for overlapping bus routes," *Transportation Science*, vol. 18, no. 3, pp. 207–230, 1984.

[19] M. Siikonen, "Elevator traffic simulation," *Simulation*, vol. 61, no. 4, pp. 257–267, 1993.

[20] L. Matias, J. Gama, J. Mendes-Moreira, and J. Freire de Sousa, "Validation of both number and coverage of bus schedules using AVL data," in *13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Madeira Island, Portugal, September 2010, pp. 131–136.

[21] A. De Vany, "Capacity utilization under alternative regulatory restraints: an analysis of taxi markets," *The Journal of Political Economy*, vol. 83, no. 1, pp. 83–94, 1975.

[22] G. Douglas, "Price regulation and optimal service standards: The taxicab industry," *Journal of Transport Economics and Policy*, vol. 6, no. 2, pp. 116–127, 1972.

[23] H. Chang, Y. Tai, and J. Hsu, "Context-aware taxi demand hotspots prediction," *International Journal of Business Intelligence and Data Mining*, vol. 5, no. 1, pp. 3–18, 2010.

[24] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti, "Taxi-aware map: identifying and predicting vacant taxis in the city," in *Proceedings of the First International Joint Conference on Ambient Intelligence*, Malaga, Spain, November 2010, pp. 86–95.

[25] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang, "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers of Computer Science in China*, vol. 6, no. 1, pp. 111–121, 2012.

[26] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Online predictive model for taxi services," in *7th International Symposium on Intelligent Data Analysis*, Helsinki, Finland, October 2012, pp. 230–240.