Where to Wait for a Taxi?

Xudong Zheng
State Key Laboratory of Software Development Environment
Beihang University
Beijing, China
zhengxudong@nlsde.buaa.edu.cn

Xiao Liang
State Key Laboratory of Software Development Environment
Beihang University
Beijing, China
liangxiao@nlsde.buaa.edu.cn

Ke Xu[*]
State Key Laboratory of Software Development Environment
Beihang University
Beijing, China
kexu@nlsde.buaa.edu.cn

ABSTRACT

People often have the demand to decide where to wait for a taxi in order to save their time. In this paper, to address this problem, we employ the non-homogeneous Poisson process (NHPP) to model the behavior of vacant taxis. According to the statistics of the parking time of vacant taxis on the roads and the number of the vacant taxis leaving the roads in history, we can estimate the waiting time at different times on road segments. We also propose an approach to make recommendations for potential passengers on where to wait for a taxi based on our estimated waiting time. Then we evaluate our approach through the experiments on simulated passengers and actual trajectories of 12,000 taxis in Beijing. The results show that our estimation is relatively accurate and could be regarded as a reliable upper bound of the waiting time in probability. And our recommendation is a trade-off between the waiting time and walking distance, which would bring practical assistance to potential passengers. In addition, we develop a mobile application *TaxiWaiter* on Android OS to help the users wait for taxis based on our approach and historical data.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: data mining, spatial databases and GIS

General Terms

Modeling, Statistics, Experimentation

Keywords

Vacant taxi, waiting time, poisson distribution

1. INTRODUCTION

Taxis play an important role in the transportation of cities. For example, Beijing has more than 60,000 taxis to provide services for about 2,000,000 passengers every day. However, there are still many people often annoyed with waiting for taxis. It is not only because of the imbalance between supply and demand, but also due to the lack of the vacant taxis information provided to passengers. For example, if a person does not know that there is another road nearby with more vacant taxis passing by, he/she would spend much more time on waiting for a taxi in current location. Experienced passengers could choose a better road to wait for taxis based on historical experiences. But more people have little knowledge about like how long they would take to wait for taxis here or where is better to wait for taxis, especially in some strange places for them, which may affect the their travels and schedules very much.

In this paper, we propose a method to estimate the waiting time for a vacant taxi at a given time and place, and then provide an approach to make recommendations for potential passengers on where to wait for a taxi. To make this estimation, we establish a model to describe the behavior of vacant taxis. Our model is based on the following observations:

- The higher proportion of time with vacant taxis parking beside the road, the more chances you can take a taxi immediately here. This situation usually occurs during the idle time around some popular places.
- The more vacant taxis leaving a road, the more chances you can take a taxi quickly. The waiting time is affected not only by the number of vacant taxis entering a road, but also by how many people want to take taxis here at that time. Because we do not have the data to directly show the demand for taxis, we think the number of vacant taxis leaving a road approximately reveals the remaining chance for a passenger to take a taxi here after the demand on the road is all met.

Motivated by the two observations above, we adopt the non-homo-geneous Poisson process (NHPP) [11] to model the events of vacant taxis' leaving and derive the probability distribution of the waiting time. Then we could perform estimations and recommendations based on the distribution. We also do some experiments to demonstrate that our approach is practicable and then develop a mobile application to help people wait for taxis.

Our study is built upon the GPS trajectories of taxis in Beijing, China. This data is collected from more than 12,000

[*]Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
UrbComp '12, August 12, 2012. Beijing, China.
Copyright 2012 ACM 978-1-4503-1542-5/08/2012 ...$15.00.

taxis, which account for about one-fifth of total ones in the city. We select the data between Oct. 2010 and Jan. 2011 to study. Each GPS record contains the identifier of a taxi, current position, timestamp, service status, and some other information. The data sampling interval of each taxi is about 60 seconds.

The major contributions of our work include:

- We employ the NHPP to model the behavior of vacant taxis, which could approximate the real situation well and have a simple form to derive the probability distribution of waiting time.
- We estimate the waiting time for vacant taxis at different time on road segments, analyze the confidence of our estimation, and design a recommender system for the people who want to take taxis.
- We conduct a lot of experiments to evaluate our approach on simulated passengers and actual trajectories of taxis. The results show that our estimation is relatively accurate and our recommendation would be helpful to potential passengers.
- We put our approach into practice by developing a mobile application which could help the user find an appropriate place to wait for a taxi.

The rest of this paper is organized as follows. In Section 2, we give an overview of the related work. Section 3 introduces our model used to estimate the waiting time for a vacant taxi. Section 4 describes the data processing of our work. In Section 5, we analyze the results of our estimated waiting time. Then, we discuss the recommendation for passengers in Section 6. Section 7 shows the experiments and evaluations on our approach. Section 8 introduces an application we developed to help people wait for taxis. Finally, we make a conclusion and propose some future work in Section 9.

2. RELATED WORK

2.1 Recommendations about Taxicabs

Recent years have witnessed the explosive research interest on taxi trajectories [1, 6, 7, 14, 18]. Moreover, many works have also been done to investigate the recommendations for taxi passengers or drivers [2, 5, 9, 13, 17].

Phithakkitnukoon et al. [9] study the prediction of vacant taxis number to provide the information for tourists or taxi service providers. They employ the method based on the naïve Bayesian classifier and obtain the prior probability distribution from the historical data. However their method divides the region of the city into one-kilometer square grids which are too rough to provide practical information for passengers. In addition, their data is only from the traces of 150 taxis in Lisbon, which might not be enough to reveal laws of vacant taxis for the reason of weak statistical significance.

Ge et al. [2] develop a recommender system for taxi drivers which has the ability in recommending a sequence of pick-up points or parking positions so as to maximize a taxi driver's profit. They estimate the probability of pick-up events for each candidate point, and then propose an algorithm to discover a route with minimal potential travel distance before having customer. Li et al. [5] study the strategies for taxi drivers as well. They use L1-Norm SVM to discover the most discriminative features to distinguish the performance of the taxis, and then extract some driving patterns to improve the performance of the taxis. However, all these studies do not concern about the recommendation for passengers.

Yuan et al. [17] propose an approach to make recommendations for both taxi drivers and passengers. They establish a probabilistic model to describe the probability to pick-up passengers, the duration before the next trip, and the distance of the next trip for a vacant taxi. Then they provide some different strategies for taxi drivers, each of which is based on the optimization of one aspect (probability of pick-up, cruising time, or profit). They further extend their work in [16]. Although their methods could also provide recommendations for passengers, their research mainly focuses on drivers. Comparing with them, our research stands on the view of passengers and pays more attentions to estimating the waiting time for vacant taxis.

Yang et al. [13] study the equilibrium of taxi market from the standpoint of economics. They use a bilateral searching and meeting function to characterize the search frictions between vacant taxis and unserved customers. They build a model to describe the relationship between the supply and demand for taxis, and analyze some influencing factors on customer waiting time. But in our study, because of lacking of explicit data for the demand of taxis, we actually estimate the gap between supply and demand through the number of vacant taxis. Moreover, the object of their study is an aggregate taxi market, but our target is to estimate the waiting time for a vacant taxi at a given time and position in a microscopic view.

2.2 Map Matching

Map matching is a main step of data preprocessing in our work. It refers to the process of mapping the GPS points to the road segments to recover a complete path of a trajectory. Quddus makes a survey of map matching algorithms in [10], including geometric, topological, probabilistic, and other advanced algorithms. He also discusses the performances and limitations of them. Lou et al. [8] propose a new algorithm for low-sampling-rate GPS data, which considers the temporal and spatial constraints on the trajectories, then constructs a weighted candidate graph to choose the most appropriate path. Yuan et al. further improve Lou's method in [15] later.

2.3 Non-homogeneous Poisson Process

Poisson process is a stochastic process that is often used to study the occurrence of events. It assumes the arriving rate of events λ is always stable, and has the Poisson distribution of counting and exponential distribution of inter-event time. Non-homogeneous Poisson process [11] is a Poisson process with a time-dependent arriving rate function $\lambda(t)$. This model is more flexible and appropriate to depict the human-related activities because these activities often vary over time and have periodicity. [3] studies the NHPP having cyclic behavior, and [4] introduces a method to estimate the $\lambda(t)$ in NHPP using a piecewise linear function.

3. MODEL

Here we propose a model to describe the waiting time for a vacant taxi, and derive the probability distribution of it. Then we estimate the waiting time using the expectation of the distribution. Finally, we analyze the confidence level of our estimation.

3.1 Motivation

The time to wait for a taxi reflects the availability of taxis on a road. Waiting for a taxi on a road could be divided into the two situations: 1) there are some vacant taxis just stopping beside the road, then you could take the taxi immediately; 2) there are no vacant taxis at hand, you should wait for the coming of next vacant taxi.

We could denote the probability of the first situation as p_{imm}, and the waiting time in the second situation as t_{next}. Then the random variable of actual waiting time t_{wait} could be represented as:

$$t_{wait} = (1 - p_{imm}) \cdot t_{next}$$

Then, we will discuss how to estimate p_{imm} and t_{next}.

3.2 Estimation of Waiting Time

Let's consider p_{imm} at first. We could approximate it by the proportion of time when there are some vacant taxis parking beside the road. We define the *parking time of vacant taxis*, i.e. t_{park}, as the duration with at least one vacant taxi parking on the road to wait for passengers. Therefore, $\hat{p}_{imm}$ for a road r during a timeslot T could be represented as:

$$\hat{p}^{r,T}_{imm} = \frac{t^{r,T}_{park}}{\Delta T}$$

Here ΔT denotes the span of timeslot T. By identifying of some appropriate stops of taxis, we can calculate $\hat{p}_{imm}$ for each road during each timeslot.

For t_{next}, intuitively, the number of vacant taxis leaving a road during a timeslot influences how long you probably spend on waiting for a taxi here. Because vacant taxis departing a road often means that they do not find any passengers on the road and then you have a great chance to take it if you are there. We denote the number of vacant taxis which have left a road as N_{vacant}, and define the *leaving frequency of vacant taxis*, i.e. λ, for a road segment r during a timeslot T as:

$$\lambda^{r,T} = \frac{N^{r,T}_{vacant}}{\Delta T}$$

Human-related activities vary over time, so do taxis. Therefore, we employ the NHPP to model the events of vacant taxis leaving roads. The rate parameter of NHPP is a time-dependent function $\lambda(t)$, and we further assume the rate function has a cycle of 24 hours. For simplicity, we adopt the piecewise linear function as the rate function of NHPP and regard each hour as a timeslot. This model assumes λ for a road is stable during a timeslot and the same timeslot in different days.

To validate our assumptions, we have done the KS-Tests for Poisson distribution on the data of each same timeslot in different days. To avoid the effect of the sparseness, we select the roads with enough data. We conduct the tests on the top 10,000 and top 30,000 roads for comparison[1]. Figure 1 shows the proportion of successful KS-Tests at 95% confidence level. We could see that the proportion for top 10,000 roads is larger than that for top 30,000 roads, and the proportion in the wee hours is rather small. These are because the data in the wee hours and unpopular roads are sparse and more fluctuant. Considering that our approach is mainly related to most of the passengers on popular roads in active time, so our hypotheses basically hold for most cases.

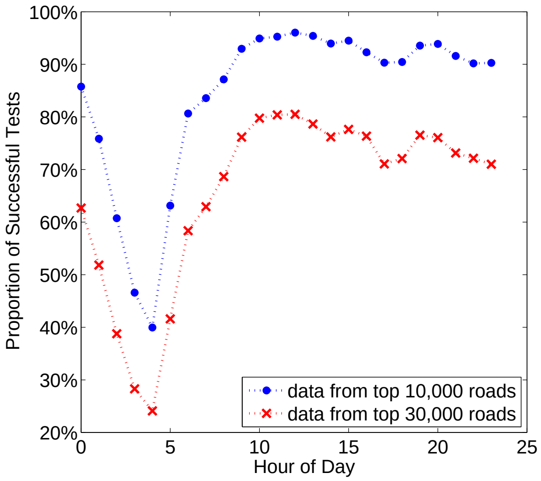


Figure 1: Proportion of successful KS-tests for the hypothesis of Poisson distribution

Under the Poisson hypothesis within a timeslot, we could derive the probability distribution of the waiting time for the next vacant taxi during the timeslot[2]. According to the Poisson process, the probability of the next event occurring within t is [12]:

$$\begin{aligned} P\{t_{next} \le t\} &= 1 - P\{t_{next} > t\} \\ &= 1 - P\{N(t) = 0\} \\ &= 1 - e^{-\lambda \cdot t} \end{aligned}$$

Here $N(t)$ represents the count of the events occurring within t, and $P\{N(t) = k\} = e^{-\lambda \cdot t} \cdot \frac{(\lambda \cdot t)^k}{k!}$. Then the probability density function of t_{next} is:

$$p(t) = \lambda \cdot e^{-\lambda \cdot t}$$

Thus, we can deduce the expectation of t_{next}:

$$\begin{aligned} E[t_{next}] &= \int_0^{\infty} t \cdot \lambda \cdot e^{-\lambda \cdot t} \cdot dt \\ &= \frac{1}{\lambda} \end{aligned}$$

Notice λ in our model denotes the *leaving frequency of vacant taxis*. Therefore with this conclusion, you could realize why the more vacant taxis leaving means the shorter waiting time for taxis. And we could regard the expectation as the estimation of t_{next}.

[1] We select the top roads by the number of pick-up events on it.

[2] For simplicity, we omit the superscript of parameters in the following derivation. It must be noted that the value of the parameter is different for various roads and timeslots.

Then we also need to estimate the λ. Here we employ the maximum likelihood estimation (MLE). If we observe the number of the vacant taxis leaving from a road at the same timeslot T for k days, we denote the count of the ith day is N_i, then the likelihood function is:

$$\mathcal{L}(\lambda) = \prod_{i=1}^{k} \frac{(\lambda \cdot \Delta T)^{N_i}}{N_i!} e^{-\lambda \cdot \Delta T}$$

Setting $\frac{d \ln(\mathcal{L}(\lambda))}{d\lambda} = 0$ and solving λ, we obtain the MLE:

$$\hat{\lambda} = \frac{\sum_{i=1}^{k} N_i}{k \cdot \Delta T} = \frac{\bar{N}}{\Delta T}$$

This conclusion means that we could estimate λ just by counting the leaving of vacant taxis in history.

As a consequence, we could estimate the actual waiting time for vacant taxis as:

$$\begin{aligned} \hat{t}_{wait} &= (1 - \hat{p}_{imm}) \cdot \hat{t}_{next} \\ &= (1 - \hat{p}_{imm}) \cdot \frac{1}{\hat{\lambda}} \end{aligned}$$

3.3 Confidence of Estimation

Now we will analyze the confidence level of our estimation. Let's consider the lower-sided confidence interval of t_{next}. We denote $1 - \alpha$ quantile of the distribution of t_{next} as $t_{next_{1-\alpha}}$, then we could get:

$$\int_0^{t_{next_{1-\alpha}}} t \cdot \lambda e^{-\lambda t} \cdot dt = 1 - \alpha$$

The quantile could be solved as:

$$t_{next_{1-\alpha}} = \frac{\ln(\alpha^{-1})}{\lambda}$$

This result shows that, we have $1 - \alpha$ confidence level of which the waiting time would be no longer than $\ln(\alpha^{-1})$ times of the $\hat{t}_{next}$ we estimated. If we set the upper bound of confidence interval equals to $\hat{t}_{next}$, namely:

$$\frac{\ln(\alpha^{-1})}{\lambda} = \frac{1}{\lambda}$$

We can get $\alpha = \frac{1}{e}$, which means the probability of the waiting time less than our estimation should be $1 - \frac{1}{e}$, which is about 63.21%. These conclusions imply that our estimation could be regarded as a reliable upper bound the possible waiting time in probability.

4. DATA PROCESSING

Our data processing starts with map matching. We have to map the trajectories of taxis to the roads and calculate the entering and leaving time of taxis to the roads. We employ the map matching algorithm proposed by Lou et al. [8]. In addition, we filter some trajectories which seem unusual, such as keeping vacant status too long (5 hours), or staying on the same road too long (2 hours).

Then, according to the model we have established, the processing is divided into two parts. The first part is the calculation of *parking time of vacant taxis*. The key step is to identify the stopping taxis that are waiting for passengers. We should eliminate the situations of waiting traffic lights or other purpose stops. We regard the taxi staying on a road with moderate duration (between 5 minutes and 2 hours) and rather low speed (less than 3.6km/h) as valid. Because too short time of stopping may be caused by traffic lights and too long time of stopping means no desire to take passengers or some unexpected situations.

The second part is to calculate the estimation of *leaving frequency of vacant taxis* λ for each road during each hour. Because the MLE of λ is $\frac{\bar{N}}{\Delta T}$, our task is just to count the number of vacant taxis leaving each road in each timeslot of one hour. And we also filter some outliers before making the average.

We process the trajectories happened during about three month, and calculate the averages $\bar{t}_{park}^{r,T}$ and $\bar{N}_{vacant}^{r,T}$, then the estimated waiting time could be represented as:

$$\begin{aligned} \hat{t}_{wait}^{r,T} &= (1 - \frac{\bar{t}_{park}^{r,T}}{\Delta T}) \cdot \frac{\Delta T}{\bar{N}_{vacant}^{r,T}} \\ &= \frac{\Delta T - \bar{t}_{park}^{r,T}}{\bar{N}_{vacant}^{r,T}} \end{aligned}$$

Here ΔT is the span of a timeslot, i.e. one hour.

However, our estimation of waiting time could not be applied to all roads, because there are some roads forbidding taxis to pick-up passengers. For these roads, there may be many taxis leaving from but few passengers getting on. Due to lack of the data indicating which road forbids the pick-up of passengers, we develop a method to detect these roads through analyzing the trajectories. We define the *pick-up rate*, denoted as θ^r for each road segment r:

$$\theta^r = \frac{\text{number of pick-up on the road segment } r}{\text{number of vacant taxis entering the road segment } r}$$

If there is a road with enough samples (more than 100 vacant taxis entering) and very low pick-up rate (less than 0.03), we will regard it as invalid to wait for taxis. For these roads, we do not make estimations of waiting time.

It is also worthy to be noted that our data is from the taxis which account for 1/5 of the total ones in Beijing. If we assume these 1/5 taxis are randomly distributed in the city, the waiting time would approximately be shortened to 1/5 of our estimation. We also could measure the actual scale factor by in-the-field study. But regardless of what the accurate factor is, the relative order of the waiting time we estimated will be basically kept under the random distribution assumption.

5. ANALYSIS OF THE RESULTS

We apply our approach to the data between Oct. 2010 and Dec. 2010, and then calculate the estimated waiting time for each road and timeslot. Because the data of some roads is very sparse, we only take the top 30,000 road segments with most frequent pick-up events into account[3]. And we also make the estimation of weekday and weekend separately.

Figure 2 gives an overview of the waiting time for vacant

[3] There is only fewer than 1 pick-up event per day in average on each of the remaining road segments.

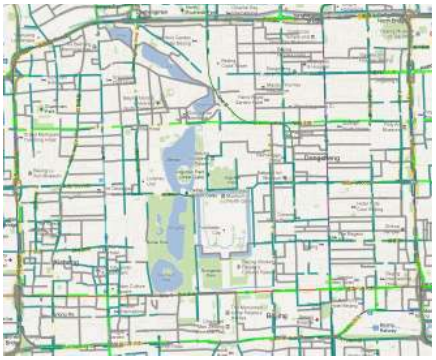


Figure 2: The map of taxi waiting time in Beijing.

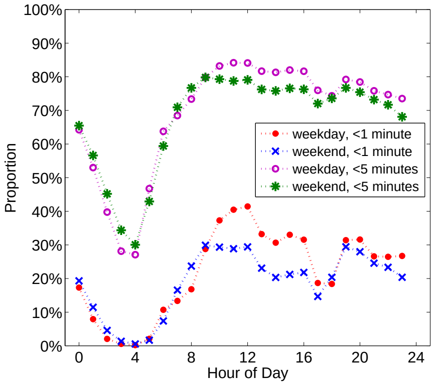


Figure 3: Proportion of roads with estimated waiting time less than 1 minute and 5 minutes.

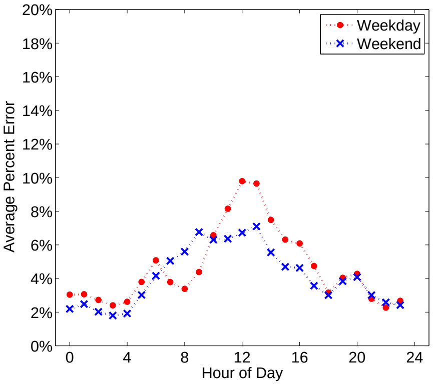


Figure 4: The percent error of the estimation of t_{park}.

taxis at 5 p.m. in a region of Beijing. The light green color denotes the estimated waiting time of less than 1 minute, the dark green color denotes the waiting time between 1 minutes and 5 minutes, and the gray color denotes the waiting time of larger than 5 minutes[4]. As shown in the map, the waiting time may be very different in some roads close to each other, so such information would help people find the appropriate position to wait for a taxi without walking too much.

Then let's analyze the varying of the waiting time in one day. Figure 3 shows the proportions of roads with estimated waiting time less than 1 minute and 5 minutes. From the figure we can see the proportion changes obviously with the time, which indicates the waiting time for taxis varies greatly during a day. The proportion of the roads with short waiting time is really low in the wee hours, because there are only a few taxis providing services. And the proportion reaches the top at noon, which implies that it would be easiest to take a taxi at that time. This is because that the demand of travel is relatively low but most taxis are in the service at noon. We also find that there are some differences between the weekday and weekend. The proportion of roads with short waiting time on weekend is not as high as on weekday in the daytime, the reason of which might be that there are more commercial and entertainment activities during that time on weekend.

6. RECOMMENDATION

With the knowledge we mined from the taxi trajectories, we could provide meaningful information to the people needing to take a taxi. With awareness of the possible waiting time on each road, people could make their schedule better, and avoid wasting time to wait for a taxi on a road with very long possible waiting time.

Furthermore, we also could provide a direct recommendation on where to take a taxi for the person who wants to take a taxi at somewhere and sometime. Considering the speed of pedestrian is slow, we limit the candidate roads to be recommended within a small distance. We denote the candidate roads set as:

$$R_{cand} = \{r : distance(P, r) < d_{max}\}$$

Here P is the position of the person now, r is a candidate road, and d_{max} is the maximal distance people want to walk. Then in the timeslot T, for each road $r \in R_{cand}$, we estimate the total time duration before taking a taxi as:

$$\begin{aligned} \hat{t}^{r,T}_{total} &= \hat{t}_{walk} + \hat{t}^{r,T}_{wait} \\ &= \frac{distance(P, r)}{\hat{v}} + \hat{t}^{r,T}_{wait} \end{aligned}$$

Here $\hat{v}$ is the common speed of the pedestrian. Then we choose the road r in candidates with minimal $\hat{t}^{r,T}_{total}$ as recommendation:

$$r_{best} = \arg\min_{r \in R_{cand}} \hat{t}^{r,T}_{total}$$

In addition, through adjusting the parameters such as d_{max} and $\hat{v}$, we could even control the preference for short waiting time or short walking distance in recommendation.

7. EXPERIMENTS AND EVALUATION

We have conducted comprehensive experiments to evaluate our model. Here we regard the data from Oct. 2010 to Dec. 2010 as the training, and choose three week between Jan. 5th 2011 and Jan. 25th 2011 for testing. We conduct our experiments on the top 30,000 road segments with most frequent pick-up events .

7.1 Validation of Statistics

We first validate two important statistical quantities in our model: *parking time of vacant taxis* t_{park} and *leaving frequency of vacant taxis* λ. Here we use percent error to evaluate the relative accuracy of our estimation. The percent error is defined as:

$$\text{percent error} = \frac{|\text{real value} - \text{estimate value}|}{\text{real value}} \times 100\%$$

Figure 4 shows the average percent error of t_{park}. The total average percent error is 4.52%. The reason of the small average error is that there are a large number of roads rarely having vacant taxis parking on. This result demonstrates the situations of vacant taxis parking beside the road have

[4] This waiting time has already been multiplied by the scale factor 1/5, the same below.

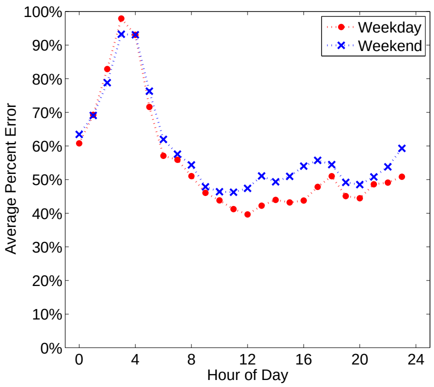


Figure 5: The percent error of the estimation of λ.

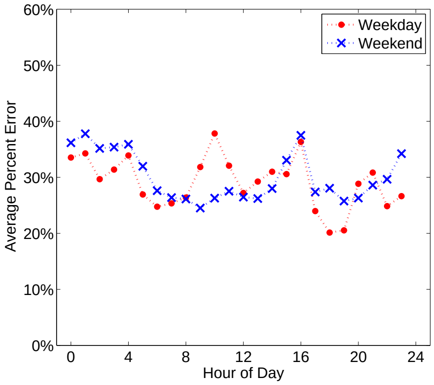


Figure 6: The percent error of the waiting time by simulation.

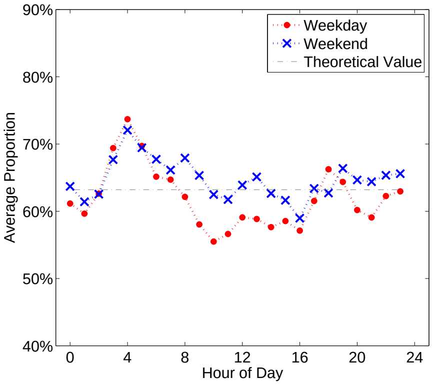


Figure 7: The proportion of tests with simulated waiting time less than the estimation.

little impact on our estimation, but we still keep it to remain the completeness of our model.

Figure 5 shows the average percent error of λ. The total average percent error is 56.12%. We could see the errors are rather larger in the wee hours due to the sparsity and fluctuation of data during that time. But for most time of a day, the percent errors are around 50%. And the weekday has smaller errors, which implies the higher regularity of human-related activities during weekdays.

7.2 Simulation

We also evaluate the estimated waiting time by simulation. For each road, we generate a passenger with a random timestamp, and then calculate how long the passenger should spend on waiting for the next vacant taxi just standing on this road according to the actual testing taxi trajectories. We repeat the simulation 100 times for each timeslot, and compare the average simulated waiting time to our estimated waiting time.

Figure 6 shows the average percent error of the waiting time on all roads at different time in simulation. The total average percent error is 29.37%. This result shows that the estimated waiting time for vacant taxis is relatively accurate and the error of our estimation is acceptable in general. Because the variance of the exponential distribution is relatively large when the λ is small, we could not avoid the errors on the roads with rare vacant taxis passing by.

Figure 7 shows the proportion of tests whose simulated waiting time is less than the estimation. The proportion in all tests is 62.73%. This is very close to the theoretical value 63.21% we have derived from our model (the straight line in the figure), which reflects that our model agrees well with reality from another side. The result also confirms that our estimation could be regarded as a reliable upper bound of the waiting time in probability.

We further evaluate our recommendation about where to wait for a taxi. We randomly generate passengers in a range of the city (no need to be on a road), as well as a timestamp. Then we choose the recommended road according the approach we proposed in section 6[5]. We compare our recommendation with three strategies: 1) *Best strategy* always chooses the road with the best t_{total} according to the actual testing data. This is a virtual strategy because it is based on posterior knowledge of taxi trajectories. It always leads to the best total waiting time, and we regard it as a baseline for comparison of time. 2) *Nearest strategy* always chooses the nearest road nearby, and then stops on it to wait for a vacant taxi. It is a common strategy in reality because people often are reluctant to walk too long. It always leads to the shortest distance to walk, and we also regard it as a baseline to compare walking distance. 3) *Random strategy* just randomly selects a road within the range. It is a possible strategy for the passenger who has no knowledge about the surroundings.

Figure 8 shows the difference of total waiting time compared with the *best strategy*. Our recommendation is obvious better than the *nearest strategy* and *random strategy* in terms of time. And our recommendation is relatively close to the *best strategy*, the total average difference is about 10 minutes. Figure 9 shows the difference of walking distance compared with the *nearest strategy*. Our recommendation is similar to the *best strategy* in terms of distance, and not much different from the *nearest strategy*. The total average difference between our recommendation and the *nearest strategy* is about 100 meters. These results show that, the recommendation made by our approach is a trade off between the waiting time and walking distance, which make the two aspects are all not much different from the best situations.

8. APPLICATION

Based on our approach and actual historical data, we develop a mobile application *TaxiWaiter* on Android OS, which could visualize the waiting time for vacant taxis on roads and also could provide a suggestion on where to wait for a taxi. Figure 10 demonstrates the user interface of the application. The roads are painted with different colors demonstrating the different waiting time on them, which could make the users intuitively understand the availability of taxis on these roads at some time. If the user clicks the *recommend* but-

[5] Here we set d_{max} to 1 km, and $\hat{v}$ to 3.6 km/h in the simulation.

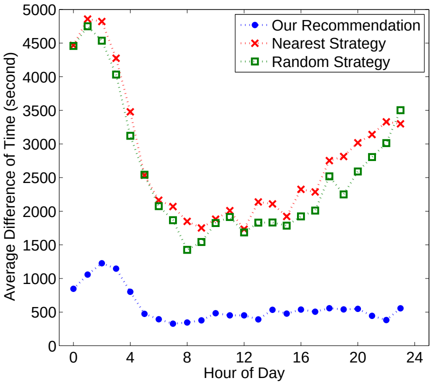


Figure 8: Difference of total waiting time compared with the best strategy.

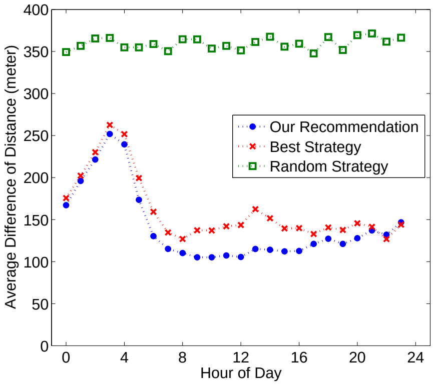
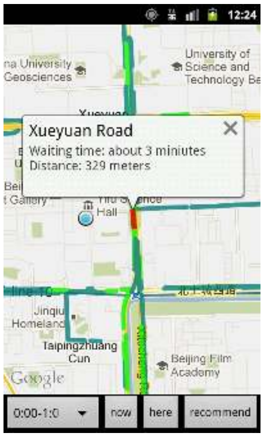


Figure 9: Difference of walking distance compared with the nearest strategy.


Xueyuan Road

Waiting time: about 3 miniutes

Distance: 329 meters

Figure 10: The mobile application *TaxiWaiter* we developed to provide taxi waiting information.

ton, the application could provide the recommended road for the user according to his/her current position (the blue point) and time. The red road in the figure is our recommendation, and the application also shows the possible waiting time and the distance to the recommended road. The user also could adjust the timeslot, walking speed and some other parameters in the application.

With *TaxiWaiter*, the user obtains more information about vacant taxis at different times on road segments and also could get a direct recommendation. These would help he/she make a better decision on where to wait for a taxi.

9. CONCLUSION AND FUTURE WORK

With the model of NHPP to describe the appearance of remaining vacant taxis, we could estimate the waiting time for the next vacant taxis on a road, and then make a recommendation on where to wait for a taxi for potential passengers. The model we established has a concise form and would lead to some meaningful conclusions in theory. The parameters in our model could be estimated from the statistics of historical data directly, which makes our approach practicable.

Through extensive experiments, we could validate that our estimations of taxi waiting time have relatively acceptable errors. The average percent error of the taxi waiting time is about 30%. The result of simulations also shows recommendations made by us would be helpful to the passengers. When the passengers following our recommendations, the total waiting time is just 10 minutes more than the *best strategy* in average, and the walking distance is only about 100 meters farther than the *nearest strategy*. This indicates that our recommendations balance the time and distance, and the two aspects are both close to the best situations.

However, there are still some limitations of our study, which would be the focuses of our future work:

- The rate function of piecewise linearity in NHPP is too simple for practical situations. We will try to use a continuous function to estimate the *leaving frequency of vacant taxis* λ which could change smoothly at any moment. This would make our model more flexible.
- The method we used to estimate the parameters such as $\lambda^{r,T}$ and $p_{imm}^{r,T}$ is just to make averaging on the historical data, and regard them as constants during different days. However, these parameters would also be changing slowly as time goes by. We consider weighing them differently based on period from that time to now, and then our method could adapt to the changes in the overall trend.
- The estimation and recommendation are not accurate enough. We will attempt to use or combine some other methods such as machine learning to improve the results, and we also plan to do some comparisons with different methods.
- We have not yet conducted in-the-field experiments to validate our approach. This type of experiments may be hard to do comprehensively. However, with the application *TaxiWaiter* we developed, we could receive feedbacks from users, which would give us a chance to validate and refine our approach.

10. ACKNOWLEDGMENTS

The research was supported by the National High Technology Research and Development Program of China (863 Program) (No. 2012AA011005), the fund of the State Key Laboratory of Software Development Environment (SKLSDE-2011ZX-02), and the Innovation Foundation of BUAA for PhD Graduates(YWF-12-RBYJ-036).

11. REFERENCES

[1] C. De Fabritiis, R. Ragona, and G. Valenti. Traffic estimation and prediction based on real time floating car data. In *Intelligent Transportation Systems*, pages 197–203, 2008.

[2] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani. An energy-efficient mobile recommender system. In *Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 899–908, 2010.

[3] S. Lee, J. Wilson, and M. Crawford. Modeling and simulation of a nonhomogeneous poisson process

having cyclic behavior. *Communications in Statistics-Simulation and Computation*, 20(2-3):777–809, 1991.

[4] L. Leemis. Estimating and simulating nonhomogeneous poisson processes. 2003.

[5] 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 *Pervasive Computing and Communications Workshops*, pages 63–68, 2011.

[6] X. Liang, X. Zheng, W. Lv, T. Zhu, and K. Xu. The scaling of human mobility by taxis is exponential. *Physica A: Statistical Mechanics and its Applications*, 2011.

[7] S. Liu, Y. Liu, L. Ni, J. Fan, and M. Li. Towards mobility-based clustering. In *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 919–928, 2010.

[8] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate gps trajectories. In *Proc. of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 352–361, 2009.

[9] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti. Taxi-aware map: Identifying and predicting vacant taxis in the city. *Ambient Intelligence*, pages 86–95, 2010.

[10] M. Quddus, W. Ochieng, and R. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.

[11] S. Ross. *Simulation*. Academic Press, 2006.

[12] S. Ross. *A first course in probability*. Prentice Hall, 2010.

[13] H. Yang and T. Yang. Equilibrium properties of taxi markets with search frictions. *Transportation Research Part B: Methodological*, 2011.

[14] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *Proc. of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 99–108, 2010.

[15] J. Yuan, Y. Zheng, C. Zhang, X. Xie, and G. Sun. An interactive-voting based map matching algorithm. In *Mobile Data Management (MDM)*, pages 43–52, 2010.

[16] J. Yuan, Y. Zheng, L. Zhang, and X. Xie. T-finder: A recommender system for finding passengers and vacant taxis. Submitted to TKDE, under second round review, 2013.

[17] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun. Where to find my next passenger? In *Proc. of the 13th ACM International Conference on Ubiquitous Computing*, 2011.

[18] D. Zhang, N. Li, Z. Zhou, C. Chen, L. Sun, and S. Li. ibat: detecting anomalous taxi trajectories from gps traces. In *Proc. of the 13th international conference on Ubiquitous computing*, pages 99–108, 2011.