Sensing the Pulse of Urban Refueling Behavior

Fuzheng Zhang[1,2], David Wilkie[2,3], Yu Zheng[2], Xing Xie[2]
[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research Asia, Beijing, China
[3]University of North Carolina at Chapel Hill, USA
zhfzh@mail.ustc.edu.cn, wilkie@cs.unc.edu, {yuzheng, xingx}@microsoft.com

ABSTRACT
Urban transportation is increasingly studied due to its complexity and economic importance. It is also a major component of urban energy use and pollution. The importance of this topic will only increase as urbanization continues around the world. A less researched aspect of transportation is the refueling behavior of drivers. In this paper, we propose a step toward real-time sensing of refueling behavior and citywide petrol consumption. We use reported trajectories from a fleet of GPS-equipped taxicabs to detect gas station visits, measure the time spent, and estimate overall demand. For times and stations with sparse data, we use collaborative filtering to estimate conditions. Our system provides real-time estimates of gas stations' waiting times, from which recommendations could be made, an indicator of overall gas usage, from which macro-scale economic decisions could be made, and a geographic view of the efficiency of gas station placement.

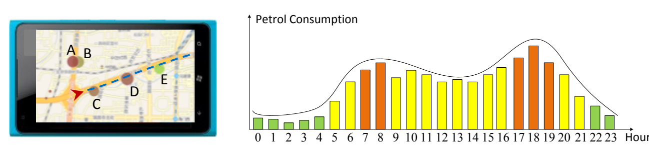
Author Keywords
Refueling Event, Knowledge Cell, Expected Duration, Arrival Rate

ACM Classification Keywords
H.2.8 [Database Management]: data mining, spatial databases and GIS.

INTRODUCTION
Urban transportation is the backbone of city life, but transportation authorities rarely have a real-time view of traffic statuses or patterns. Additionally, due to the heavy and growing reliance on petroleum and the environmental impacts of emissions from fossil fuel consumption, energy consumption for urban transportation represents a pressing challenge. An integral and under-researched component of the transportation system is the refueling behavior of individual cars, which we propose to monitor in real-time using ubiquitous sensing data. We propose a step toward real-time sensing of refueling behavior, overall petrol consumption, and a framework for analyzing gas station efficiency.

In this paper, we propose a system that uses city-wide sensing by human actors to capture both the individual refueling experiences (e.g. time spent at a gas station) and the macroscopic system dynamics (e.g. city-wide petrol consumption, efficiency of gas stations, etc.). We use human-generated trajectory data to identify refueling events, estimate the time spent, and infer other local and global properties.

Energy use in vehicle transportation is difficult to ascertain. This is especially true for real-time estimates. Gas stations are typically owned by an assortment of different, competing organizations which do not want to make data available to competitors. There is also a cost associated with monitoring and publicizing data, from which station owners would derive no benefit. Estimating energy use is also a difficult problem, as it is a function of a car's acceleration, which is highly variable and difficult to estimate.



(a) User recommendation (b) Petrol consumption

Figure 1. Application scenarios of refueling activity understanding

(a)Lack of demand (b) Demand surplus

Figure 2. A local view of gas stations

We propose a complete data-driven framework to understand urban refueling activity. We focus on estimating the time spent and the arrival rate of each knowledge cell (a spatial-temporal unit detailed latter). These two indicators can be applied in the following scenarios:

- *User Refueling Recommendation*: Figure 1(a) shows several gas stations' time spent at a point in time. The redder the color, the more time spent. Assuming a driver is in the position denoted by the arrow, even if station C is the closest, other stations might be rec-

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.

DOI string from ACM form confirmation.

ommended due to their shorter waiting times, i.e., station E is a satisfactory choice.

- *Gas Station Planning*: Figure 2 shows a local view of several gas stations. The size of the stations indicates the drivers' average arrival rate. The larger the size, the more drivers have visited that station. We see a large amount of vehicles have refueled in the area shown in Figure 2(a), thus these stations have long waiting times (colored red). It might be worthwhile to consider building a new gas station nearby to relieve the issue of insufficient supply. On the contrary, Figure 2(b) indicates that the gas stations are very dense in this area even though very few drivers have visited there (colored green). Therefore, the government could consider closing some of them to reduce waste.
- *Energy Consumption Analysis*: In Figure 1(b), the curve gives a direct view of this city's time-varying petrol consumption, based on the drivers' arrival rate during each period. This can be used by station operators to formulate better commercial strategies.

Our approach is a 'human as a sensor' approach that draws inferences from GPS-trajectories passively collected by taxicabs. At first, we take a novel approach to detect refueling events, which are visits by taxis to gas stations. The detection includes the time spent waiting at the gas station, and the time spent refueling the vehicle. For knowledge cells which cover enough detected refueling events, the time spent in each of these cells is estimated directly. For those with few or even without refueling events, we use a context aware collaborative filtering approach to solve the data sparsity problem. Finally, we treat each gas station as a queue system and the time spent in the station is used to calculate drivers' arrival rate, which is the number of customers during this period and can indicate the petrol consumption. Therefore, the output is a global estimate of time spent and fuel use at each gas station in each time period.

Our evaluation consists of multiple parts. First, we conduct several experiments on the refueling event detection algorithm. We analyze the performance on a manually-labelled GPS data set as well as a data set generated by the authors. Next, we show the performance of the time spent estimation and the effectiveness of collaborative filtering. Finally, we evaluate the effectiveness of the arrival rate estimation by comparing the number of customers deduced with the results collected in a case study.

Our work presents a step towards real-time, persistent monitoring of urban transportation energy use and refueling behavior by passive human sensing. Our main contributions include the following:

- We propose a method for the discovery of refueling events from GPS trajectories
- We present a context aware collaborative filtering method to estimate the time spent at gas stations when data is sparse.
- We develop an approach that uses queue systems to calculate the overall arrival rate at gas stations from the inferred time spent during a period.

We evaluate our system using large-scale and real-world datasets, which consists of a trajectory dataset, POI dataset, and road network dataset.

PRELIMINARY

In this section, we will clarify some terms used in this paper and briefly describe our major datasets.

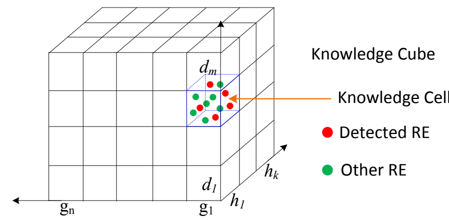


Figure 3. Knowledge cube and knowledge cell

Trajectory: A trajectory is a sequence of GPS points that is composed of a latitude, a longitude and a timestamp.

Point of Interest (POI): A POI refers to a specific point location that someone may find useful or interesting. It is described by a latitude, a longitude, and a category (such as restaurant, gas station, etc.).

Refueling Event (RE): A refueling event describes the phenomenon a vehicle refueling at a gas station. It is composed of the arrival time, departure time and the selected gas station. A refueling event's duration represents the time spent there, which is the difference between the arrival time and departure time.

Knowledge Cell and Knowledge Cube: A knowledge cell is a spatial temporal division for refueling events. A knowledge cell C_{ijk} corresponds to a gas station g_i with the timestamp of hour h_j and the timestamp of day d_k, as shown in Figure 3. Each RE falls under one certain cell (its selected gas station is mapped to g_i, its arrival time is mapped to h_j and d_k), and therefore all cells combine to form a knowledge cube. A knowledge cell is the finest granularity we use for urban refueling behavior analysis. A cell has two indicators with which we are concerned: expected duration and arrival rate. The expected duration refers to how much time, on average, is spent by the vehicles refueling in this cell. The arrival rate indicates how many drivers have visited this cell.

Our system is built on three kinds of data sources. The trajectory dataset was generated by over 30,000 taxis in Beijing during a period of nearly two months, from which taxi drivers' refueling events can be detected. The POI dataset contains hundreds of thousands of POIs in this city, where gas stations are one category of particular interest. The road network dataset covers about 150,000 road segments in the urban area, where each segment is described as a sequence of geospatial points as well as some other attributes (such as road level, the number of lanes, etc.).

SYSTEM OVERVIEW

Our system provides insight into the refueling behavior in the city. This behavior is captured by estimating each knowledge cell's expected duration and arrival rate. As a preliminary step, we first identify the refueling events in the trajectory data. Then, for knowledge cells with a sufficient number of refueling events, we model the expected duration as the average of the values of the contained REs. For knowledge cells that have few or even no REs, we propose a context aware collaborative filtering model to predict the expected durations from similar knowledge cells. With the expected durations estimated, we model each knowledge cell as a queue system in order to calculate the cell's arrival rate. Finally, with each cell's expected duration and arrival rate estimated, we can perform spatial and temporal analyses on a city scale. There are four main components in our system as shown in Figure 4.

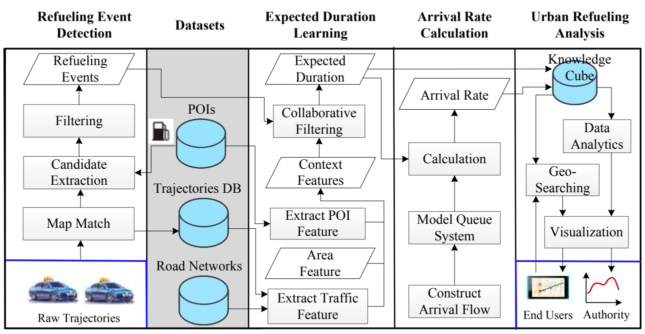


Figure 4. System overview

Refueling Event Detection. In this component, a large number of candidates are extracted from raw trajectories, and then a filtering algorithm is applied to obtain the final results.

Expected Duration Learning. For a cell containing sufficient detected REs, its expected duration is represented with the detected REs' average durations. Then, for cells with insufficient REs present, we train a collaborative filtering method to predict their expected durations. We also consider the stations' contextual features that would have an influence on drivers' refueling behavior and incorporate these features into our model.

Arrival Rate Calculation. We model each gas station as a queue system and make a statistical inference of its arrival rate depending on a cell's time spent.

Urban Refueling Analysis. Based on the detected REs and each cell's two indicators, we analyze taxi drivers' refueling activity as well as the entire city's refueling behavior.

REFUELING EVENT DETECTION

By mapping the geospatial movement of cars to the positions of gas stations, it seems that refueling events can be easily discovered. However, this direct approach has difficulties due to the noise of the GPS readings and it cannot support perfect matching. The GPS devices generally have an error of 10 meters and the position of a gas station is merely depicted as a single point (which is actually an area with hundreds of square meters), these two factors lead the direct approach to mistake drivers' other behavior for refueling events while pass up real refueling behavior. This section details the process of detecting refueling events from the taxis' raw trajectories under uncertainty. We first extract the refueling candidates and then use a supervised method to filter the errant candidates.

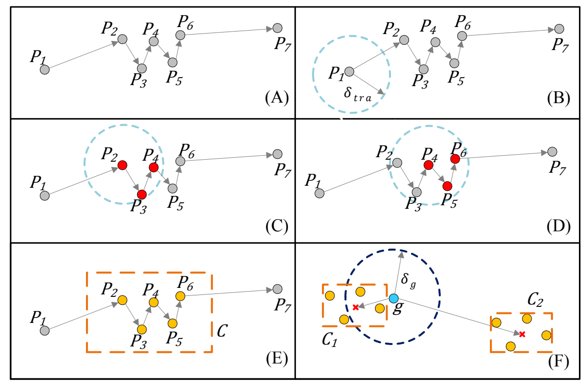


Figure 5. Candidate extraction

Algorithm 1: ParametersLearning

Input: Labeled Real Refueling Events Collection $\mathcal{R}$
Output: δ_{tra}^{best}, τ^{best}, δ_g^{best}

$err_{\mathcal{R}}^{min} = \infty$;
for δ_{tra} *in range*$(LowerBound_{\delta_{tra}}, UpperBound_{\delta_{tra}})$ **do**
  for τ *in range*$(LowerBound_{\tau}, UpperBound_{\tau})$ **do**
    for δ_g *in range*$(LowerBound_{\delta_g}, UpperBound_{\delta_g})$ **do**
      $\mathcal{C}$ = CandidatesExtracting(δ_{tra}, δ_{tra}, δ_g);
      $err_{\mathcal{R}} = 0$;
      for *each real refueling event* r *in* $\mathcal{R}$ **do**
        $err_{\mathcal{R}} = err_{\mathcal{R}} + Min_{c \in \mathcal{C}}\ TemporalDist(r, c)$;
      end
      if $err_{\mathcal{R}} < err_{\mathcal{R}}^{min}$ **then**
        $\delta_{tra}^{best} = \delta_{tra}$, $\tau^{best} = \tau$, $\delta_g^{best} = \delta_g$, $err_{\mathcal{R}}^{min} = err_{\mathcal{R}}$
      end
    end
  end
end
return δ_{tra}^{best}, τ^{best}, δ_g^{best} ;

Procedure 1: TemporalDist(r, c)

Input: real refueling event r, candidate c
Output: $dist$

if $r.taxi.equals(c.taxi)$ & r *have temporal interaction with* c **then**
  /* AT : arrival time, DT : departure time */
  $dist = |c.AT - r.AT| + |c.DT - r.DT|$
end
else
  $dist = \infty$
end
return $dist$;

Candidate Extraction

We extract refueling event candidates by considering mobility and geographic constraints.

For the mobility constraint, we ensure a refueling event candidate corresponds to a period of slow movement. As shown in Figure 5(A), given a trajectory $p_1 \to p_2 \to \cdots \to p_7$, we first check the distance between each point until the distance is larger than a threshold δ_{tra}. As shown in Figure 5(B), since $dist(p_1, p_2) > \delta_{tra}$, we move next and take p_2 as "pivot point". We find that $dist(p_2, p_3) < \delta_{tra}$ and $dist(p_2, p_4) < \delta_{tra}$ while $dist(p_2, p_5) > \delta_{tra}$, as shown in Figure 5(C). If the interval between $p_2.t$ and $p_4.t$ is smaller than τ, p_2, p_3, p_4 forms a cluster. Then, as shown in Figure 5(D), we fix p_4 as a "pivot point" to check on the later points. Finally, we take p_2, p_3, p_4, p_5, p_6 as a "stay point", which is shown in Figure 5(E).

For the geographic constraint, we check the distance between a stay point's center point and the nearest gas station. Then we take those stay points satisfying $dist(c, g) < \delta_g$ as refueling event candidates. As shown in Figure 5(F), C_1 is reserved while C_2 is discarded directly.

We manually labelled 200 real refueling events by plotting the raw trajectories in digital maps and used this dataset to learn the parameters ($\delta_{tra}, \tau, \delta_g$). The parameters were determined by traversing combinations of values, as shown in Algorithm 1. The temporal distance (Procedure 1) signifies how accurate a candidate can represent a real refueling event's arrival time and departure time. We ensure that each real refueling event corresponds to a candidate (they should have temporal overlap, if not, the distance is infinite) while still guaranteeing the temporal distances gathered from all real refueling events are minimized.

Filtering

The candidates extraction process finds clusters of points in close proximity to gas stations. However, a candidate could be generated by some other behavior. For example, for gas stations that are close to roads or intersections, the candidate might indicate a traffic jam or a car wait for signals at a traffic light. Some other POIs such as repair shops, car washes, or even parking lots, might be located close to gas stations and create false candidates. Figure 6 show a real refueling event compared with pseudo candidates. To filter these non-refueling events out of the candidate pool, we apply a supervised model, using the following features:

(a) Real RE (b) Traffic jam or waiting for signal (c) Parking place

Figure 6. Real RE w.r.t pseudo candidates

Spatial-Temporal features including: 1) *Encompassment*. A binary value indicating whether the gas station is contained in the candidate's minimum bounding box. 2) *Gas Station Distance*. The average distance between the candidate's points and the gas station. 3) *Distance To Road*. The average distance between the candidate's points and a matched road segment. 4) *Minimum Bounding Box Ratio*. The ratio between the minimum bounding box's width and height represented as $Min\left(\frac{width}{length}, \frac{length}{width}\right)$. 5) *Duration*. The temporal duration of a candidate.

POI features including: 1) *Neighbor Count*. The number of POIs in the gas station's neighborhood. 2) *Distance To POI*. The minimum average distance between a candidate's points and nearby POIs.

We use a manually labelled dataset to train a gradient tree boosting classifier[1], and then use the trained model to distinguish real refueling events from other behavior.

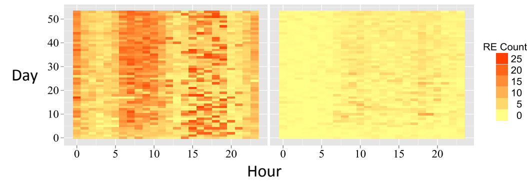


Figure 7. Detected REs' heatmap in two gas stations

EXPECTED DURATION LEARNING

A knowledge cell's expected duration is an indicator that shows the average time spent at a gas station during a certain period. Currently, we have discovered taxis' refueling events from the trajectory dataset. If there are enough REs incorporated, we can use their average durations to estimate this indicator. However, only a portion of the knowledge cells are filled with enough detected REs. Figure 7 shows two slices from the knowledge cube along the gas station dimension, and each small colored grid corresponds to a knowledge cell, where the color signifies the number of detected REs. Even though the left gas station is popular for taxis, during some periods it was still rare for taxis to arrive. The situation is even worse for the station on the right. To predict the remaining cells' expected duration, we apply a context aware collaborative filtering model to solve the data sparsity problem and then detail how to extract the gas station's contextual features to improve the performance.

Context-Aware Collaborative Filtering

Currently, for cells with enough detected REs, their expected durations are obtained and treated as observable data. Our concern is then to find the remaining cells' expected durations. The problem actually concerns collaborative filtering, where the timestamp of the hour is treated as the user, the gas station is treated as the item and the timestamp of the day can be treated as the temporal factor[2]. The scene in our system could be imagined that there are 24 users (each user relates to an hour), they give variant rates to different items (each item relates to a gas station) at different times (each time snapshot relates to a day). Naturally, the expected duration of a knowledge cell could be interpreted as a user rate on an item at a certain time snapshot. In our system, the user ratings are analogous to the expected duration of the knowledge cells. Matrix factorization is the state-of-the-art model used for collaborative filtering when dealing with user-item rating prediction. Tensor factorization is therefore applied to the high dimensional prediction problem[3]. We first discuss how to apply tensor factorization to predict knowledge cell's expected duration.

Formally, we denote the knowledge cube's expected duration as a sparse three-dimensional tensor, denoted by $Y \in R^{h \times g \times d}$, where h is the number of hours in a day, g the number of gas stations, and d the number of days. Whenever a knowledge cell C_{ijk} covers more than 2 detected REs, we regard Y_{ijk} as being observed and use the REs' (who fall in this cell) average duration to denote.

We apply High Order Singular Value Decomposition (HOSVD)[4] to factorize the three-dimensional tensor into three matrices $H \in R^{h\times d_H}, G \in R^{g\times d_G}$, $D \in R^{d\times d_D}$ and one central tensor $S \in R^{d_H\times d_G\times d_D}$. The three matrices are compact representations of the three attributes in subspaces, where d_H, d_G, d_D are dimensionality parameters to balance capability and generalization. The reconstructed value for cell C_{ijk} in the traditional tensor factorization[3] is given as

$$F_{ijk} = S \times_H H_{i*} \times_G G_{j*} \times_D D_{k*} \quad (1)$$

We denote the tensor matrix multiplication as $\times_U$, where the subscript denotes the direction, ie. $T = Y \times_U U$ is $T_{ljk} = \sum_{i=1}^{h} Y_{ijk} \times U_{ij}$. The entries of the ith row of matrix U is represented as U_{i*}.

Additionally, the single tensor factorization does not take full advantage of our data, since it only tries to find out the three attributes' latent connections in subspaces through what we have already observed, it does not consider other factors would also influence the observations. Another important signal, gas stations' contextual features, have not been considered. Actually, economists have found exogenous and endogenous factors (i.e. location, nearby traffic flow, its size, etc.) have a great effect on gas stations' competitive condition[5]. Thus these factors could influence the time spent in gas stations (if a gas station is popular and always busy, it will typically have a longer waiting time). An item's contextual features are often modeled in collaborative filtering to help reduce uncertainty issues[2]. Assume there are $c_1, c_2, \cdots, c_L$ features, where feature c_l has categorical values $1,2,\cdots z_l$ to refer to contextual conditions. By integrating the tensor factorization with the context features[6], the reconstructed value for cell C_{ijk} is redefined as

$$F_{ijk} = S \times_H H_{i*} \times_G G_{j*} \times_D D_{k*} + \sum_{l=1}^{L} B_{lc_l} \quad (2)$$

where B_{lc_l} is the parameter modeling how contextual feature c_l with condition would have an affect on the reconstructed value. This introduced contextual parameters guarantee the fact that stations with similar contextual features tend to have similar time spent (the part $\sum_{l=1}^{L} B_{lc_l}$ tends to be similar between similar stations).

In order to generate expected duration predictions, the model parameters should be learned using the observable data. We define the learning procedure as an optimization problem:

$$\min_{H,G,D,S,B} L(Y,F) + \Omega(H,G,D,S,B) \quad (3)$$

where $L(Y,F)$ is the loss function given as

$$L(Y,F) = \frac{1}{||S||_1} \sum_{i,j,k} Z_{ijk} \cdot \left(Y_{ijk} - F_{ijk}\right)^2 \quad (4)$$

where $Z \in \{0,1\}^{h\times g\times d}$ is a binary tensor with nonzero entries Z_{ijk} whenever Y_{ijk} is observed. Equation (4) indicates we consider the reconstructed accuracy for observed cells. $\Omega(H,G,D,S,B)$ is the regularization term to prevent overfitting, which is given as

$$\Omega(H,G,D,S,B) = \frac{1}{2}\lambda \times (||H||_{Frob}^2 + ||G||_{Frob}^2 + ||D||_{Frob}^2 + ||S||_{Frob}^2 + ||B||_{Frob}^2) \quad (5)$$

Equation (3) guarantees our model could reconstruct the observations as accurately as possible and meanwhile maintaining the capability of generalization. We use stochastic gradient descent[7] to solve this optimization problem.

Contextual Features Extraction

We consider three types of contextual features for gas stations, POIs, traffic flow and the size of the gas station.

POI feature F_P: We determine the POI feature according to a gas station's nearby POIs. For each category C of the POIs, to discover its correlation to the gas station, we use the metrics defined by Jensen et al. in [8], which is given as

$$J_c = \frac{\#co\_location(C,g)}{\#C} \quad (6)$$

where $\#co\_location(C,g)$ refers to the frequency of co-location for category C with the gas station, while $\#C$ indicates the individual frequency. The top 5 discovered POIs are {Service Zone At Motorway, Toll station, Factory, Vehicle Maintenance and Vehicle Service}. Aggregating nearby POIs, the POI feature of a gas station is given as

$$F_p(g_i) = \sum_C N(C, g_i) \cdot J_C \quad (7)$$

where $N(C, g_i)$ indicates the frequency of the category C standing by station g_i.

Traffic feature F_T: The traffic feature of a gas station depends on its nearby traffic flow and competitive conditions. By aggregating all the trajectory data for each road, we can estimate this road's traffic flow. We determine how a road's traffic flow influences nearby gas stations based on the Huff Probability Model[9], which is given as

$$TF(r \to g_i) = TF_r \cdot \frac{\frac{1}{dist(g_i,r)}}{\sum_{g_j} \frac{1}{dist(g_j,r)}} \quad (8)$$

Finally, the traffic feature of a gas station is given as

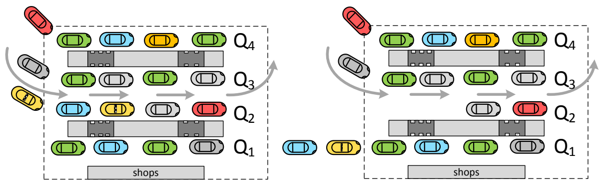
$$F_T(g_i) = \sum_r TF(r \to g_i) \quad (9)$$

Area feature F_A: The area feature of a gas station reflects its passenger capability and therefore it influences the time spent of this station. We manually labelled the gas stations' areas in satellite maps.

Ultimately, because context aware collaborative filtering needs categorical variables, we divided each feature into five categories separately and used them as the gas stations' contextual features (the three features correspond to c_1, c_2, c_3 separately).

ARRIVAL RATE CALCULATION

A knowledge cell's expected duration indicates the time spent there. We also want to know how many vehicles have visited the cell, from which we can estimate the energy consumption. However, our dataset only covers about 30,000 taxicabs, which is only a small portion of the total number of vehicles in this city. To solve the sparsity problem inside a gas station, we estimate the total arrival rate by modeling each gas station as a queue system.



(a) Optimal (b) Suboptimal

Figure 8. Optimal w.r.t. Suboptimal inside a gas station

Queue System

A gas station diagram is shown in Figure 8(a). There are several queues and each queue could simultaneously serve several vehicles. To reduce the complexity of the system, we make some simplifications. First, we ignore transfers from one queue to another queue and assume each vehicle is fixed to a certain queue. This assumption guarantees each queue can be treated as an independent queue system. Moreover, we make the assumption that drivers will always choose the shortest queue to join. In Figure 8(b), Q_1 is much longer than Q_2 and we believe these drivers would not prefer to this suboptimal option, and such a case would not happen typically in reality. Therefore, this assumption ensures each queue will share the same waiting time on the whole.

Assume there are Q queues in the gas station. We know a knowledge cell corresponds to a gas station during a certain period. In this cell, the vehicles' arrival flow for each queue Q_i is described as a homogeneous Poisson process $N(t, \lambda_i)$, which indicates the number of vehicles in the period $[0, t]$ is a Poisson distribution with parameter $\lambda_i \cdot t$[10]. The unit of t is hours, the same as the period of the cell. Thus, λ_i is the number of vehicles that had joined this queue in this cell, and the overall arrival rate of this cell is given as $\lambda = \sum_{i=1}^{Q} \lambda_i$.

Calculation

In the queue system, given customers' arrival stochastic process and servers' service time distribution, the equilibrium indicators such as waiting time, system time, etc., can be obtained[10]. We assume all refueling equipment is undifferentiated and its service time satisfies exponential distribution $Exp(\mu)$. For the ith queue Q_i, we assume it has c_i servers, so that this queue can be treated as a $M/M/c_i$ system. Its average arrival rate is λ_i and its average service time is $\frac{1}{\mu}$. The equilibrium indicators can be computed as follows[10]:

$$W_s = \frac{\frac{\lambda_i^{c_i}}{\mu^{c_i+1}(c_i-1)!}}{\left(1-\frac{\lambda_i}{c_i\mu}\right)^2} \cdot \left[\sum_{k=0}^{c_i-1}\frac{1}{k!}\left(\frac{\lambda_i}{\mu}\right)^k + \frac{1}{c_i!-\frac{\lambda_i(c_i-1)!}{\mu}}\left(\frac{\lambda_i}{\mu}\right)^{c_i}\right]^{-1} + \frac{1}{\mu} \quad (10)$$

W_s is the equilibrium system time (including both the waiting time and service time), which means at the equilibrium state, when a vehicle joins this queue, the length of time the vehicle is expected to stay. Since we believe drivers are rational, each queue's equilibrium system time is the same and we use each cell's predicted expected duration to represent W_s. We see that W_s only depends on λ_i, μ and c_i. Given μ, c_i and W_s, we solve the equation to get parameter λ_i, where the equation can be solved by a numerical algorithm, such as the Newton Raphson method. Finally, this cell's arrival rate λ is gathered by each queue's corresponding λ_i.

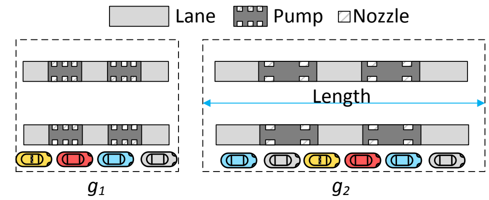


Figure 9. Layout of gas stations

Parameters Determination

We assume the shortest duration of all detected REs corresponds to service time (there are some cases taxis can refuel directly). We select the top 1000 shortest durations and use their average value to estimate $\frac{1}{\mu}$. We then need to determine Q (number of queues) and c_i (number of servers in a queue) at each gas station, which is dependent on the gas station's area, the arrangement of pumps and how many nozzles, as shown in Figure 9. A pump has several nozzles and nozzle plays the role of server. As mentioned before, we measure the stations' lengths in satellite maps. We also go through the street view maps to observe the number of lanes N_l, the number of nozzles along a queue N_{n_i}. We see a pump can serve both sides simultaneously and therefore Q is equal to $2 \times N_l$. It is a little tricky to determine c_i. The figure shows gas station g_1 has 6 nozzles along a queue, however due to the length limitation, it can only serves 4 vehicles simultaneously. The situation is contrary for g_2. Thus we set $c_i = \min(N_{n_i}, \frac{length\ of\ gas\ station}{length\ of\ a\ vehicle})$. In reality, the length of a normal automobile is about 4.5 meters, therefore we set it at 5 meters in view of the gap between vehicles.

EXPERIMENT

In this section, we first describe the datasets and then evaluate the performance of refueling event detection, the expected duration learning and the arrival rate calculation.

Data Description

Road Network We evaluated our methods using the road network of Beijing, which contains 106,579 road nodes and 141,380 road segments.

Taxi Trajectories The dataset covers the GPS trajectories from 2012, which were collected by about 30,000 taxicabs located in Beijing during the period of Oct and Nov. The details are presented in Table 1.

POIs There are a total of 369,668 POIs with 602 kinds of categories. 1221 gas stations are located in this city, of which 689 gas stations are located in the areas covered by our road network while the others are not. In our system, we only concentrate on the former.

Human-Labelled Dataset We employ four human labelled datasets for training and evaluation as follows:

1) HLD-1 We manually labelled 250 refueling events by plotting the taxis' raw trajectories on digital maps. 200 of

them were used to learn the parameters in candidate extraction and the remaining were used to validate the performance.

2) HLD-2 We manually labelled 2,000 candidates (True\False) by plotting the extracted candidates on digital maps.

3) HLD-3 This dataset covers 33 trajectories collected by two authors, and each trajectory contains a recorded refueling event (arrival time, departure time, selected gas station), which is used to evaluate the performance of refueling events detection.

4) HLD-4 To evaluate whether the expected duration learning component and the arrival rate calculation component work well in reality, we chose two gas stations on which to perform a field study. We recorded the vehicles' arrival and departure times (there were many vehicles and we could not record all their information, so we just selectively recorded some cases) and also how many vehicles had refueled there in that period. This field study lasted from Oct.17 to Nov.15 in 2012, ranging from 5:00pm to 6:00pm each time. Totally, 14 days of records were collected (each station had 7 days' worth of records).

Raw Trajectories	Total Taxi Count	32,476
	Duration	54 day
	Ave Distance By Day	226.76 km
	Ave Sampling Interval	1.02 minute
Detected REs	Total Count	638,645
	Average Temporal Interval	1.84 day
	Average Distance Interval	378.61 km
	Average Duration	10.53 minute
	Minimal Duration	3.74 minute
	Maximal Duration	42.72 minute

Table 1. Trajectory dataset w.r.t. detected REs

Experiments for Refueling Event Detection

In this subsection, we evaluate the effectiveness of candidate extraction and the filtering model separately.

Temporal Distance (minute)	HLD-1		HLD-3	
	Mean	Std.	Mean	Std.
$\lvert r.AT - c.AT \rvert$	1.07	0.41	0.52	0.27
$\lvert r.DT - c.DT \rvert$	1.25	0.53	0.71	0.22
$\lvert r.AT - c.AT \rvert + \lvert r.DT - c.DT \rvert$	2.32	0.46	1.23	0.24

Table 2. Temporal distance between candidate and real RE

Results of Candidate Extraction

We used 200 instances in HLD-1 to learn the parameters and evaluated the performance both on the remaining 50 instances in HLD-1 and the authors' collected dataset HLD-3. As shown in Table 2, we computed the temporal distance (AT corresponds to arrival time and DT corresponds to departure time) between the labelled refueling time and the nearest candidates discovered. The performance was better in HLD-3 because the GPS devices used by the authors have a lower sampling interval (the sampling interval is about 5 second while the taxis' GPS sampling interval is about 1minute).

Results of Filtering

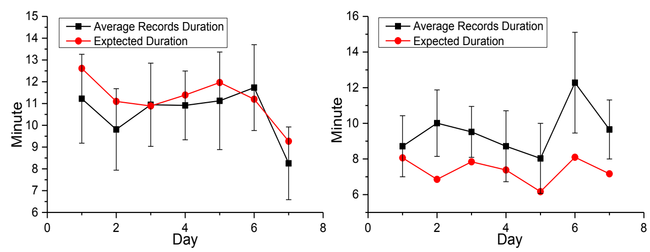
The precision and recall w.r.t. features we used for the classifier are presented in Table 3. We applied a 10-fold cross validation method on dataset HLD-2. The performance on HLD-3 was still better than HLD-2, because there was less noise in the candidates. Compared to private car owners, taxi drivers visit gas stations' nearby POIs more frequently, such as vehicle repair shops or parking areas, which can generate pseudo candidates. What's more, we found that temporal feature plays an important role in both datasets. In any case, the precision and recall were both higher than 90%, which is accurate enough for the next step. After applying the method to all the candidates, the description of detected refueling events is presented in Table 1. The average temporal interval shows that a taxi would almost drive to refuel about every two days, similar to the indication of the average distance interval. The average duration shows taxi drivers' average time spent is 10.53 minutes. The minimal duration implies a vehicle will take at least 3.74 minutes to finish refueling behavior, while the maximal duration indicates long waiting time.

	Features	Precision	Recall
HLD-2	Non-Filtering	0.464	1.0
	Spatial	0.623	0.73
	Spatial+Temporal	0.891	0.862
	Spatial+Temporal+POIs	0.915	0.907
HLD-3	Non-Filtering	0.825	1.0
	Spatial	0.875	0.848
	Spatial+Temporal	0.941	0.969
	Spatial+Temporal+POIs	0.941	0.969

Table 3. Results of filtering

	D_1	D_2	D_3	D_4	D_5	D_6	D_7
g_1	7	6	5	5	6	6	4
g_2	0	1	0	0	0	0	2

Table 4. Number of detected REs in each cell



(a) g_1 (b) g_2

Figure 10. Records' duration w.r.t. expected duration

Experiments for Expected Duration Learning

There are a total number of 892,944 cells (24 hours × 689 gas stations × 54 days) in the knowledge cube, and each cell incorporates 0.715 refueling events on average, which indicates amount of cells were lack of enough detected REs to estimate the expected duration. Table 4 details how many detected REs are covered during the period of our case study at these two gas stations. As we see, g_1 is more attractive to taxis and these cells incorporate enough detected REs, while taxis rarely patronize g_2. Therefore, for each cell in g_1, its expected duration is represented by the detected REs' average duration, and the results are shown in Figure 10(a) and compared with the results of the recorded vehicles' duration in g_1 in the field study. The standard deviation of records is about 2 minutes, which shows that during an hour, the refueling time spent is almost stable.

The results of Expected Duration Learning

There are four baselines we used for the comparison:

Average Filling: 1) *AWH* (Average within Hour). For a knowledge cell without sufficient detected REs, AWH finds all the other knowledge cells with the same hour timestamp, and uses their average expected durations to estimate this cell's expected duration. 2) *AWD* (Average within Day). Similar to AWH, AWD uses the average value within the same day. 3) *AWG* (Average within Gas Station). Analogous to the previous two methods, AWG uses the average value within the same gas station.

SVM. It uses the contextual features of the gas stations, as well as the timestamp of the hour and the timestamp of the day as temporal features to train a supervised model using SVM regression.

We selected the cells that incorporate more than 2 detected REs, and obtained 312,537 cells as observable data. We evaluated our model using 10-fold cross validation to the observable data and the results are presented in Table 5, where MeanErr signifies the average offset between the observable value and the predicted value for all testing data in the 10 fold cross validation. The unit of MeanErr is minute and similar to Std. The table shows that the contextual features of the gas stations play an important role in improving the performance. The SVM model performed even worse than AWH, perhaps because the data tensor is quite sparse and a supervised model is not fit for this situation. The results indicate time spent error estimated in a cell could be limited within about 2 minutes on average.

	MeanErr	Std
AAH	3.03	0.97
AAD	3.74	1.29
AAG	3.11	1.12
SVM	3.18	1.26
TF	2.66	0.83
TF + F_P	2.49	1.02
TF + F_P + F_T	2.27	0.86
TF + F_P + F_T + F_A	1.98	0.84

Table 5. Results of collaborative learning w.r.t. baselines

Additionally, to evaluate the performance of collaborative filtering, we compared the predicted value with gas station g_2's records in Figure 10(b). It seems our model prefers to give a lower value.

	N_L	N_{n_i}	$Length$	Q	c_i
g_1	3	4	27.2 m	6	4
g_2	2	4	18.7 m	4	3

Table 6. Description of two gas stations

Experiments for Arrival Rate Calculation

In this subsection, we discuss the experiment with the calculation of the knowledge cells' arrival rate.

Table 6 details the records of two gas stations as well as their determined queue-model parameters. These two gas stations have an identical number of nozzles in each queue, denoted as N_{n_i}. Similarly, each queue's number of servers is denoted as c_i.

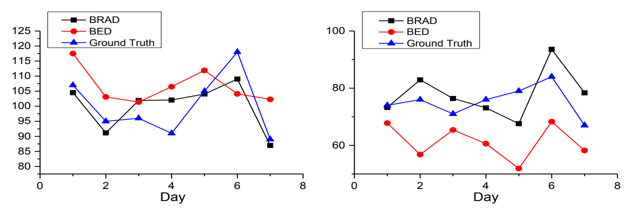
For the service time parameter μ, we selected the top 1000 shortest durations among all the detected refueling events and finally obtained $\tilde{\mu} = 4.06$ minutes.

We compared the following methods with the ground truth (the recorded total vehicles' visits of two gas stations' in each day):

BRAD (Based on Recorded Average Duration). This method uses the selectively recorded vehicles' average duration to estimate equilibrium system time W_s.

BED (Based on Expected Duration). This method makes use of each cell's expected duration to estimate W_s.

The results are shown in Figure 11. The figure shows BRAD approximates to the ground truth, which illustrates the effectiveness of our queue system model. In addition, the figure indicates BRAD is more accurate than BED, because BED is dependent on the results of refueling event detection and expected duration learning, the errors accumulated in these two parts exert an influence on arrival rate's results. However, for both gas stations, we found that the gap between BED and the ground truth was acceptable.



(a) g_1 (b) g_2

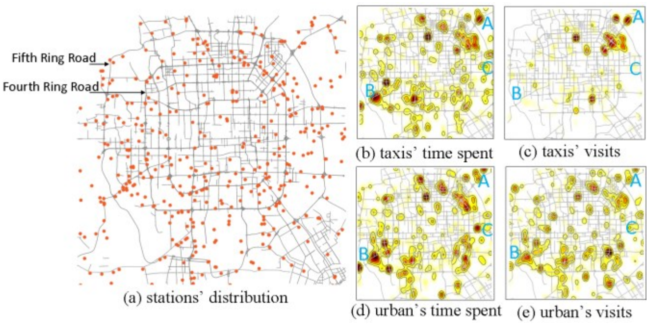
Figure 11. Results of arrival rate

URBAN REFUELING ANALYSIS

We obtained taxis drivers' (in our dataset) refueling events as well as two indicators for each knowledge cell, expected duration and arrival rate. This knowledge reveals taxis drivers' refueling behavior and at the same time presents the whole city's refueling behavior from spatial and temporal prospective.

Geographic View

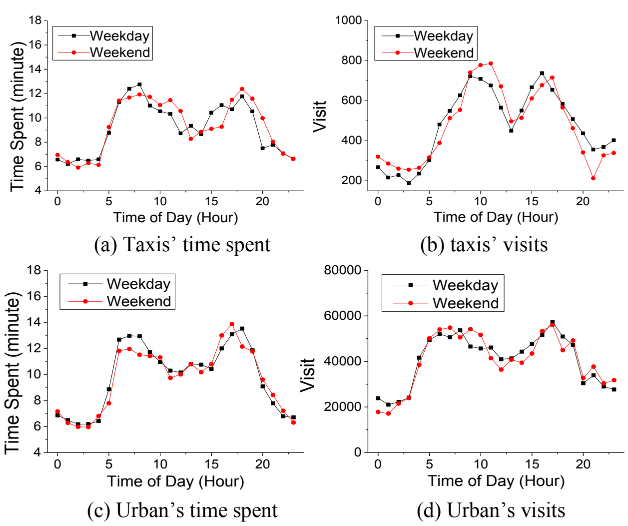
Figure 12(a) pictures how gas stations are scattered in this city. The gray lines depict the city's road network. The figure shows that a large portion of stations are located between the fourth ring road and the fifth ring road, while fewer stations are distributed in the central part of Beijing. Figure 12(b) presents the spatial distribution of taxi drivers' time spent while Figure 12(c) shows the distribution of their visits. Redder color refers to longer time spent or more visits. We see that most of the areas taxi drivers frequently visited were also endowed with longer time spent. On the other side, taxis drivers rarely patronized stations in area B, however long waiting time was still required, which implies that there were many other vehicles refueling thereby. According to our survey, we found area B was near at the entrance of a major highway and thus many private vehicles refueled there. Taxi drivers frequently refueled on the southeast part of the fifth ring road, or several other small-scale hotspots scattered in the south and north. Actually these hot areas are transportation hub and have a dominant advantage to attract taxis. For instance, the hot area A is near the highway directly to airport, where many taxis travel and they will refuel at nearby gas stations with high probability.



(a) stations' distribution (b) taxis' time spent (c) taxis' visits (d) urban's time spent (e) urban's visits

Figure 12. Refueling behavior's spatial distribution

For the entire city's refueling behavior, we aggregated the knowledge cells corresponding to the same gas station together, then used these cells' average expected duration to denote this station's time spent and used averaged arrival rate to denote this station's visits. This city's refueling time spent and visits are spatially distributed in Figure 12(d) and Figure 12(e). The figures shows that longer time spent tends to indicate more visits, however, some exceptions exist such as area C. We found there are many small-size stations in that area, the fact drivers had to wait longer is mainly due to these stations' limited capacity. Compared with Figure 12(a), we found that although a large amount of stations have been built in area B, the long time spent suggests new stations still should be planned nearby.



(a) Taxis' time spent (b) taxis' visits

(c) Urban's time spent (d) Urban's visits

Figure 13. Refueling behavior's temporal distribution

Temporal View

We aggregated cells corresponding to the same timestamp of the hour together and denote this city's time spent and visits using the average value. Additionally, weekdays and weekends were separated. Figure 13(a) and Figure 13(c) separately show how taxi drivers' and the city's refueling time varied during a day. During the rush hours (7am, 8am, 6pm, 7pm), many private vehicles came to refuel, and more waiting time was needed. On the other side, the figures show in weekends, a little less waiting time was needed than weekdays at about 7am and 8am, while a little more time was needed at about 9am and 10am. This phenomenon accords with office workers' habits, they often choose to refuel on the way to work in the morning and they do not need to week up early on weekends, and therefore there were fewer customers early in the weekends' morning. Figure 13(b) shows taxi drivers' refueling climax was at about 10am, which indicates they chose to stagger the busy period at about 8:00am. The two peaks in Figure 13(d) indicates higher petrol consumption during these periods and they also warns people to avoid refueling at that time.

DISCUSSION

We discuss the generalization of our methods as well as the limitation of the system in this section.

Our work is currently only dependent on taxis' trajectories, however, other vehicles can be seamlessly incorporated into this system. During the refueling detection phase, as shown in the experiments, we see that the result for private cars even outperform that for taxis, because taxi drivers tend to generate more other behavior nearby gas stations. When we obtain the two indicators in the cell, we actually only rely on the detected REs' time spent, which is the result of refueling event detection and is independent on whether this vehicle is a taxi or not. Our taxi trajectories can be regarded as a sampling of the whole trajectories generated by all vehicles in this city.

On the other side, taxi drivers' might care more about price, which will lead them to some special gas stations and aggravate the sparsity issues of other stations in our current system, the potential different refueling regularity between taxi drivers and normal drivers might degrade the accuracy in practice. We also use taxis' refueling time solely to estimate the parameter of refueling time distribution, which might bring some bias (some other vehicles' refueling time will be usually larger than taxis, such as trucks). In addition, the drivers' behavior in the gas stations' queue system is ideally assumed and the reality is usually more complex than we can capture. For instance, when the lane in a gas station is narrow, a car who has finished refueling might be blocked by the car in front, this special case is difficult incorporated into our system. The contextual features here is also confined and we intend to take more factors into consideration in future work, such as price, brand, etc..

RELATED WORK

Customer Refueling Behavior Analysis

The refueling issues mainly focus on understanding customers' refueling regularity to help make decision. Kelly et al. [11] interviewed 259 drivers in southern California, they analyzed these interviewers' refueling behavior, the results were used to help to select appropriate optimal facility location models. Li et al. [12] used a smart phone application to build a driving behavior monitoring and analysis system especially for hybrid vehicles. Compared with their interesting and influential work which primarily aimed at individuals, our system steps further on macro-scale analysis through large-scale datasets.

Gas Station Analysis and Planning

Gas stations problems mostly focuses on facility location problem or economic factors. For instance, Chan et al. [13] proposed an econometric model to analyze both the geographic locations of gasoline retailers in Singapore, as well as price competition between these retailers. [5,14] examined how product design, prices and locational characteristics influenced price competition in retail gasoline markets. These works concentrated on analyzing stations' self-characteristics, while our work tries to discover stations' petrol consumption through passive human sensing, in a more intuitive way to understand stations' operating status.

Urban Computing

With the popularity of diverse sensors, exploring the rule of urban is a burgeoning and attractive area in computer science. The term "Urban Computing", has emerged to concentrates on the integration of computing, sensing, and actuation technologies into everyday urban settings and lifestyles. Recent years, amount of interesting work based on spatial temporal analysis, have been proposed to explore the status of the city[15,16]. Leontiadis et al. [17] performed a case study that evaluated whether a decentralized intelligent transportation system can help drivers to minimize trip times. In [18], a strategy is provided to find efficient driving directions based on taxis drivers' knowledge. Yuan et al. [19] presents a recommender system for both taxi drivers and passengers based on passengers' mobility patterns and taxis' drivers' picking-up/dropping-off behavior. Our work concentrates on catching a glimpse of urban transportation's energy consumption, which is a closely concerned topic covered in urban computing area.

CONCLUSION

In this paper, we propose a framework for discovering urban refueling behavior using taxis trajectories, POIs and road network. Depending on taxis' detected refueling behavior and the estimated result of gas stations, we analyze urban refueling behavior from both the spatial and temporal perspectives. The discovered refueling regularity could benefit a variety of application. In the mindset of customer, the gas stations' wait time could be used to recommend the least time-consumption choice. For governmental department, they could rethink whether current layout of stations is reasonable, whether some stations are excessively dense in an area while other areas might lack of this infrastructure. In the business perspective, the investors could analyze drivers' refueling behavior to help choose location that is most promising to attract customers. We evaluate our system with large scale dataset, including two-month taxis trajectories in 2012, together with POIs and road network in Beijing, as well as several human collected datasets.

We will further study how to give real-time inference of gas station status. At the same time, we will collect more gas stations' detail information (such as price, payment type, brand) to enhance the performance of collaborative filtering and queue system.

REFERENCES

1. J. H. Friedman, "Stochastic gradient boosting," in Computational Statistics & Data Analysis, vol. 1, no. 3, 2002, pp. 367–378.
2. G. Adomavicius and A. Tuzhilin, Recommender Systems Handbook. Boston, MA: Springer US, 2011, pp. 217–253.
3. A. Karatzoglou, X. Amatriain, and N. Oliver, "Multiverse Recommendation : N-dimensional Tensor Factorization for Context-aware Collaborative Filtering," pp. 79–86.
4. L. De Lathauwer, B. De Moor, and J. Vandewalle, "A Multilinear Singular Value Decomposition," SIAM Journal on Matrix Analysis and Applications, vol. 21, no. 4, pp. 1253–1278, Jan. 2000.
5. G. Iyer and P. B. Seetharaman, "Quality and location in retail gasoline markets," 2005.
6. L. Baltrunas, B. Ludwig, and F. Ricci, "Matrix Factorization Techniques for Context Aware," In Proceedings of the fifth ACM conference on Recommender systems, pp. 301–304.
7. T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in Twenty-first international conference on Machine learning - ICML '04, 2004, p. 116.
8. P. Jensen, "A network-based prediction of retail stores commercial categories and optimal locations," pp. 1–5, 2008.
9. D. L. Huff, "Defining and Estimating a Trading Area," Journal of Marketing, vol. Vol. 28, N, pp. 34–38.
10. L. Kleinrock and J. Wiley, "Queueing Systems," IEEE Transactions on Communications, vol. Volume 1:, pp. 178–179, 1977.
11. S. Kelley, M. Kuby, G. Sciences, and U. Planning, "On the Way or Around the Corner ? Observed Refueling Choices of Alternative Fuel Vehicle Drivers in Southern California," no. 1, p. 1025313, 2006.
12. K. Li, M. Lu, F. Lu, Q. Lv, L. Shang, and D. Maksimovic, "Personalized Driving Behavior Monitoring and Analysis for Emerging Hybrid Vehicles," in Pervasive Computing, 2012, p. pp 1–19.
13. T. Y. Chan, An Econometric Model of Location and Pricing in the Gasoline Market, no. August 2004. 2006, pp. 1–37.
14. G. Iyer and P. B. Seetharaman, "Too close to be similar: Product and price competition in retail gasoline markets," Quantitative Marketing and Economics, vol. 6, no. 3, pp. 205–234, May 2008.
15. Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11, 2011, p. 89.
16. T. Kindberg, M. Chalmers, and E. Paulos, "Urban Computing," 2007.
17. I. Leontiadis, G. Marfia, D. Mack, G. Pau, C. Mascolo, and M. Gerla, "On the Effectiveness of an Opportunistic Traffic Management System for Vehicular Networks," IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 4, pp. 1537–1548, Dec. 2011.
18. J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-Drive : Driving Directions Based on Taxi Trajectories." In Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems. ACM, 99–108, 2010
19. N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-Finder: A Recommender System for Finding Passengers and Vacant Taxis," IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2012.