Inferring Unmet Demand from Taxi Probe Data

Afian Anwar*, Amedeo Odoni† Daniela Rus*
* Computer Science and Artificial Intelligence Lab / Massachusetts Institute of Technology, USA
† Operations Research Center / Massachusetts Institute of Technology, USA

***Abstract*—Matching taxi supply with demand is one of the biggest challenges faced by taxi fleet operators today. One of the reasons why this problem is so hard to solve is because there are no readily available methods to infer unmet taxi demand from data. An algorithm that reliably does so would be of enormous value to fleet operators because it could be used to dispatch available taxis to areas where passenger demand greatly exceeds supply. In this paper, we formally define unmet taxi demand and develop a heuristic algorithm to quantify it. We explain how our method improves on traditional approaches and present the theoretical details which underpin our algorithm. Finally, we develop a smartphone application that uses our algorithm together with a live taxi data feed to provide real time recommendations to participating drivers and efficiently route taxis to where they are needed most.**

I. Introduction

In this paper we describe a heuristic algorithm for inferring unmet passenger demand for taxis ("unmet demand") from data. Matching taxi supply with demand is one of the biggest challenges faced by taxi fleets worldwide. In Taiwan, it is estimated that taxis in the capital Taipei spend more than a third of their driving time empty looking for customers [1] while in Singapore, it is not uncommon for taxis to spend hours in queue waiting for a fare at the airport [2]. Despite this, there is a significant imbalance of supply and demand in both these cities particularly during peak periods, often resulting in unacceptably high waiting times for passengers [3]. One possible explanation is that taxi drivers as a group do not know where and when to find passengers and so spend time unproductively cruising empty or waiting in queue [4]. To help these taxi drivers find passengers more efficiently and identify areas where taxis are needed, we have developed an algorithm that uses real time taxi probe data to infer areas of high unmet demand. Our algorithm is able to do this without any information on passenger queue lengths or arrival rates.

We measure unmet taxi demand by the quantity U, which is the answer to the question: "How many more taxis are needed in an area to completely satisfy all taxi demand for a given period of time?" The relationship between unmet demand U, the number of observed passenger boardings B and demand D is governed by the following equation:

$$D - B = U \geq 0 \tag{1}$$

Note that by our definition, U is always a non-negative integer and that we make a distinction between D, the true (but unknown) demand for taxi service and B, the number of passenger boardings observed from data. For example, consider

Fig. 1. A taxi queue with high unmet taxi demand, U. Because passenger demand for taxis is much greater than supply, the observed demand B (the number of boardings) underestimates the actual demand for taxis, D.

a situation where you observe a hundred people at an isolated taxi stand and that within an hour, only five taxis passed by to pick up passengers (Figure 1). If you only counted boardings, you would mistakenly conclude that demand for taxis at this taxi stand was only five passengers per hour, when in fact it was much greater. The only time we can be absolutely confident of estimating U is when $U = 0$, implying that $D = B$ i.e. taxis were in such abundance that the number of observed boardings is indicative of actual demand.

We can also describe $U = D - B$ as the residual queue length of people at the taxi stand i.e. the number of people who needed a taxi less the number of people who eventually found one . Similarly, the the residual queue length of taxis is the supply of taxis S less the number of taxis that found a passenger B.

We want to infer U from data. Our approach borrows heavily from queueing theory, particularly the observation that at equilibrium, the quantity of service supplied will be greater or equal to the equilibrium quantity demanded by a certain amount of slack [5]. When there is little slack relative to demand, the residual queue length of people is much larger than the residual queue length of taxis. Since the number of people that are matched with a taxi, B, is the same as the number of taxis that are matched with people (also B), we can say that when unmet demand U is large:

$$U = D - B \gg S - B \implies D \gg S \implies \frac{D}{S} \gg 1 \quad (2)$$

$$\frac{U}{S-B} = \frac{D-B}{S-B} = \frac{D}{S-B} - \frac{B}{S-B} \quad (3)$$

The ratio $\frac{D}{S}$ is known as the *utilization* of a queueing system. Rather than calculate U directly, we observe from (2) and (3) that when U is large, $\frac{D}{S}$ is also large as is $\frac{D}{S-B}$, the ratio of demand to the residual queue length of taxis.

In Section V, we use this to approximate U by defining a quantity ρ, the ratio of taxi demand to excess supply per unit time and prove that it is a constant multiple of $\frac{D}{S-B}$. The rest of the paper is organized as follows. We describe the data we use for this study in Section III. Section IV introduces the problem setup, defines notation and states assumptions that allow us to use taxi probe data to estimate unmet demand. Section VI applies our heuristic using real world taxi data from Singapore. Finally in Section VII, we propose a service model and use our unmet demand algorithm to develop a smartphone application that uses a live data feed to provide real time recommendations to participating taxi drivers.

The main contributions of this paper are:

- a survey of current methods and strategies that take a data driven approach to taxi optimization
- a formal definition of unmet taxi demand and an algorithm to estimate it
- a rigorous analysis of the theoretical details which underpin our algorithm
- the development of an online recommendation engine and smartphone application that directs taxi drivers to areas of high unmet demand

II. Related Work

Measuring unmet demand is particularly important to transit agencies because this information is used to inform taxi licensing policy. Traditionally, this data has been collected using surveys conducted by human observers who would position themselves at taxi stands to measure passenger wait times and queue lengths. For example in the UK, unmet taxi demand surveys taken in Cornwall [6] and Dundee [7] were used to help policy makers decide if they should permit new taxi registrations while in Hong Kong, the Department of Transportation has been conducting annual taxi surveys since the 1980s with the aim of ensuring that the number of new taxi licenses issued keeps up with demand [8]. In response to public feedback that taxis were more difficult to find despite supply increasing by almost 50% since 2003 [9], the Singapore Land Transport Authority (LTA) commissioned monthly surveys of passenger wait times at selected taxi stands in the central business district [10]. More recently, the LTA embarked on a pilot program to install video cameras at seven taxi stands that would use image recognition to count the number of passengers in queue [11].

These methods are popular because they present an administratively simple way for agencies to obtain ground truth data but as noted in [12], they are are costly, time-consuming and limited to small sample sizes. In contrast, by using individual taxis as a distributed network of sensors, we are able to infer unmet demand at any arbitrary location with sufficient taxi activity without the use of expensive fixed equipment or human surveyors.

Historically, our problem of matching taxi supply and demand has been examined through the lens of combinatorial optimization, where it falls within the general class of dynamic pickup and delivery problems in which people or objects have to be collected and delivered in real time [13] [14]. Because this problem is NP-hard [15], heuristic solutions are needed. [16] used simulated annealing to maximize customer utility and minimize fleet operating costs while [17] designed a branch and cut algorithm to return a minimum-cost set of vehicle routes that satisfies all user requests under time constraints. Since these problems are typically solved using mathematical programming, they scale poorly [18], making them unsuitable for solving large fleet assignment problems in real time.

Another way to frame taxi supply and demand matching is to view it as a rebalancing problem in a networked, mobility on demand system where we model passenger locations as nodes and taxis as vehicles that drive autonomously from one delivery location to the next according to some global rebalancing policy. Such studies emphasize formally characterizing the fundamental performance limits of such systems and devising dynamic routing strategies with provable performance guarantees [19]. For example, [20] developed a provably optimal rebalancing policy that minimized the number of empty vehicle (rebalancing) trips while ensuring that the number of waiting customers remained bounded and [21] developed a systematic approach to size a fleet of shared, automated vehicles based on actual mobility patterns in a city.

These studies work well in the context of a robotic fleet of autonomous vehicles which collaborate to maximize aggregate quality of service, but they do not capture the dynamics of real world taxi fleets where drivers compete with one another to maximize individual utility and earning power.

More recently, the availability of data collected from onboard GPS devices has made it possible to analyze and understand the movements of taxi fleets at scale. This new body of research falls into two groups. The first focuses on using real world data to build theoretical models of personal mobility [22], [23], with a special emphasis on epidemiology [24]–[26], city planning [27] and the role that taxies can play in supporting public transportation [28], [29]. The second seeks to develop data driven decision support tools that enable taxi drivers to operate more efficiently. It is this second group that is most similar to our work. [2] used queueing theory to combine flight arrival with taxi supply data to predict passenger demand for taxis at different airport terminals in Singapore while [30] used k-means clustering to describe the spatiotemporal structure of the taxi demand on Jeju Island, South Korea. Other strategies include stable matching to optimally assign taxis to passengers [31], computing dynamic patrolling loops that minimize the

distance driven by taxi drivers to get to their next customer [32] and employing ensemble time series forecasting techniques to predict near term taxi demand [33].

In developing recommendation systems for taxi drivers, each of the above trained their demand predictions on *observed* demand - the number of people who board a taxi B, not the number of taxis needed D. As we explained in Section I, observed demand B is only equal to true demand D when the number of available taxis in an area is greater than or equal to the number of boardings. Advising taxi drivers to avoid a low B area may not be optimal if B is low precisely because taxi supply is limited. This is analogous to asking taxi drivers to give a taxi stand with many waiting passengers and very few taxis a miss. To the best of our knowledge, our approach of targeting unmet demand U, instead of observed demand B has no parallel in the literature.

III. Data

We use data from a large taxi fleet in Singapore continuously acquired over a three month period (June 2012 - August 2012) using telematics installed on each of the fleet's 16000 taxis. The 1 TB of data we used contains some 1.5 billion taxi records, where each record stores the taxi id, driver id, time stamp, speed, latitude, longitude and operational state (FREE - the taxi is available and looking for a passenger, POB - the taxi is busy with a passenger, and ONCALL - the taxi is on his way to fulfill a passenger request made through the booking system). These states are manually set by the taxi driver via the taxi's in-vehicle dispatch unit. Records are logged at discrete one minute intervals, which allows us to track the minute by minute position and state of each taxi over the the entire three month period.

IV. Problem Formulation

In this section we formulate the problem, define notation and state assumptions that allow us to infer unmet taxi demand from data. Recall that in Section I, we defined the unmet passenger demand for taxis U as the difference between actual demand D and observed taxi boardings B. Since it is not possible to calculate U directly, one simple idea that comes to mind is to pose unmet demand as the ratio of taxi demand to availability. This requires us to come up with a reasonable measures for both that are verifiable from data.

A. Assumptions

We receive a live data stream from each taxi in the form of a 4-tuple (*taxi_id, taxi_stand_id, time_period, state*).

- *taxi_id* allows us to uniquely identify each taxi
- *taxi_stand_id* refers to the virtual taxi stand that the taxi occupies. The bounds of each taxi_stand are clearly defined, so it is trivial to use the taxi's latitude and longitude to identify which taxi stand it is in.
- *time_period* is the discretized time interval in which the data was received

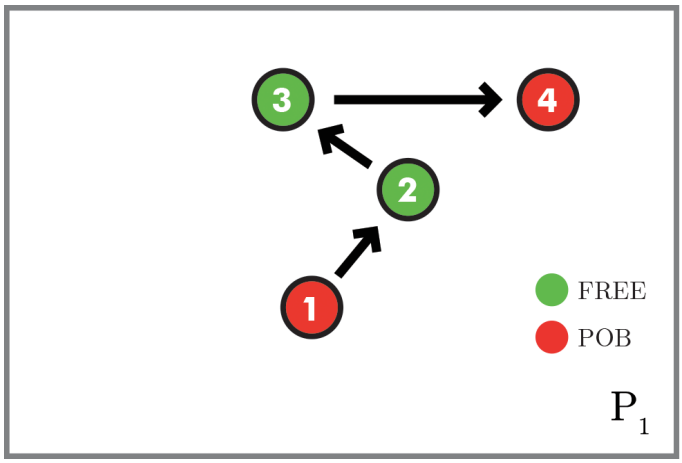

Fig. 2. A single taxi servicing an urban area contained within virtual taxi stand P_1.

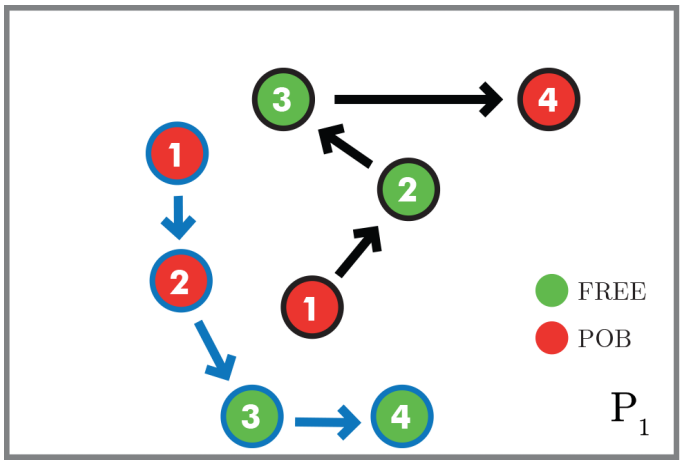

Fig. 3. Two taxis servicing an urban area contained within a single virtual taxi stand P_1.

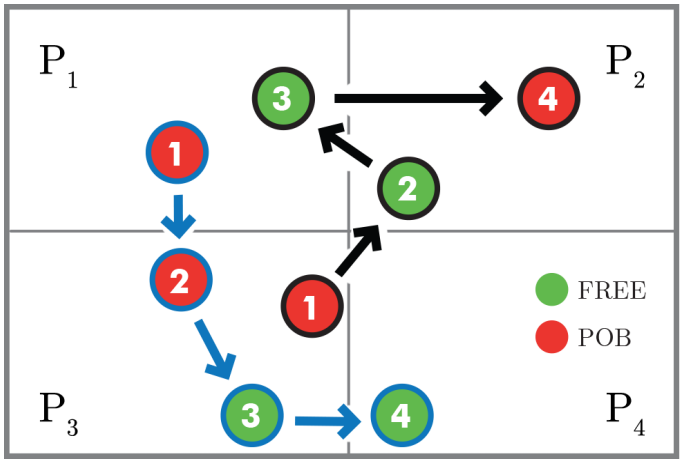

Fig. 4. Two taxis servicing an urban area contained within a four virtual taxi stands $p = \{1, 2, 3, 4\}$.

For simplicity, we make the assumption that a taxi's *state* can only take one of two values - FREE and POB. In practice, we treat the third state ONCALL identically to POB because the taxi is not available to accept passengers. We further assume that data is received at exact intervals and that a taxi cannot share multiple states within within each *time_period*. For example, if we set the *time_period* to be one minute long, then our system receives updates at one minute intervals and will never run into a situation where a taxi updates its status twice or more per minute. Lastly, we assume that exactly one passenger boards a taxi. This is assumption is critical because as explained in Section I, we define the residual queue length as the number of people (or taxis) less the number of boardings. If more than one person boards a taxi, this calculation cannot hold. In real world implementation we approximate the queue length of people by multiplying the number of boardings B by 1.1, the average number of passengers per taxi trip in Singapore [34].

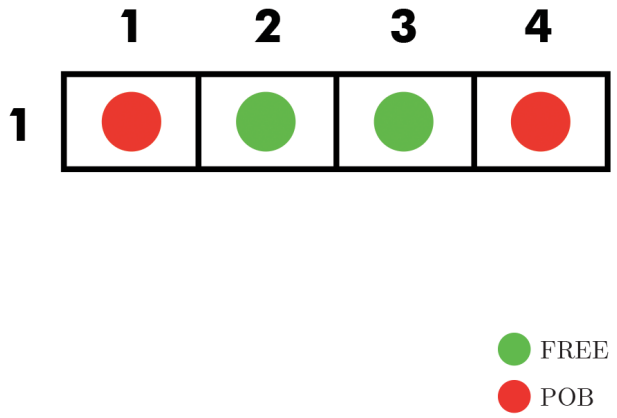

Fig. 5. A 1×4 matrix that completely summarizes the activity of a single taxi $m = 1$ over four discrete time periods $n = \{1, 2, 3, 4\}$.

B. Taxi Slack

Suppose we want to quantify taxi availability in a city at a point in time or over a short period of time e.g. 15 minutes. We introduce the concept of the free taxi minute, a metric that represents the state of the taxi when available (FREE) in one minute and show how this can be used to measure taxi availability. First suppose we partition our city into a single $p = 1$ virtual taxi stand P_1. Consider a taxi $m = 1$ (outlined in black), servicing the urban area completely contained in P_1 (Figure 2). The position and state of the taxi is sampled at four discrete time periods $n = \{1, 2, 3, 4\}$ (shown in white inside each circle). The state of the taxi can take one of two values - FREE (green) and POB (red).

We can completely describe the activity of this taxi with a 1×4 matrix as in Figure 3 where the $(m, n)^{th}$ entry represents the state of the m^{th} taxi at time n (Figure 5). Let us now add a second taxi $m = 2$ (blue) as shown in Figure 3. We can similarly summarize the activity of both taxis using a 2×4 matrix as in Figure 6.

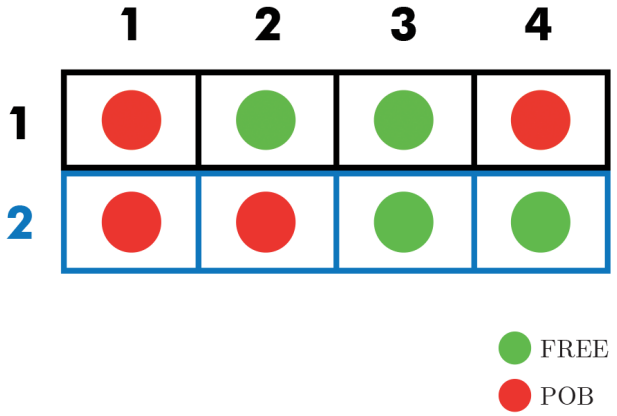

Fig. 6. A 2×4 matrix that completely summarizes the activity of two taxis $m = 2$ at four discrete time periods $n = \{1, 2, 3, 4\}$. By counting the number of FREE (green) taxis column wise, it is easy to see that taxi availability peaks at time period $n = 3$

$$\text{SLACK}(M, n', p') = \sum_{m \in M} \text{FREE}(m, n', p') \tag{4}$$

$$\text{SLACK}(M, T, p') = \sum_{t \in T} \sum_{m \in M} \text{FREE}(m, t, p') \tag{5}$$

Where

$$\text{FREE}(m, n, p) = \begin{cases} 1 & \text{if the } (m, n, p)^{th} \text{ entry is FREE} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

By summing the state of each taxi along each column we can reconstruct a chart of taxi activity. It is easy from Figure 6 to see that taxi availability (as measured by the number of FREE taxis) peaks at time period $n = 3$.

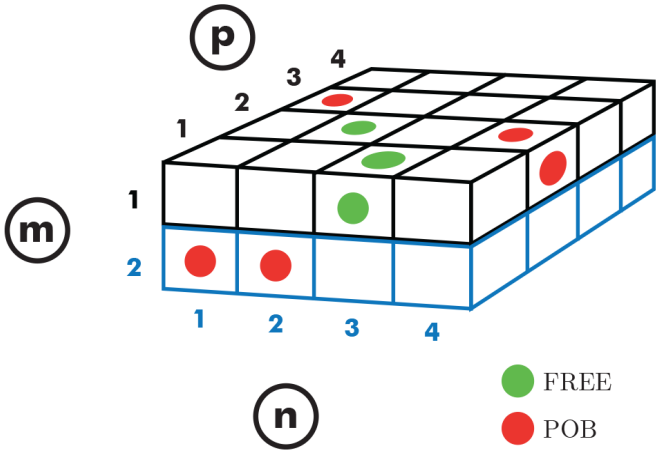

Fig. 7. A $2 \times 4 \times 4$ matrix that completely summarizes the activity of two taxis $m = 2$ in an urban area spanning $p = 4$ virtual taxi stands over four discrete time periods $n = \{1, 2, 3, 4\}$.

Finally, we subdivide the city into four virtual taxi stands $p = \{1, 2, 3, 4\}$ as in Figure 4 and use the $(m, n, p)^{th}$ entry of a $2 \times 4 \times 4$ matrix (Figure 7) to represent the state of taxi m, at time period n in grid p. Again, observe that for a given grid p, summing the state of each taxi column wise plots a graph of taxi activity over time. It is easy to see that without any loss of generality, we can extend our approach to model an environment with M taxis, N time periods and P virtual taxi stands using an $(M \times N \times P)$ matrix.

Using the above reasoning, we can now define a measure of taxi availability or *slack* for at any given virtual taxi stand $p = p'$ for any time period $n = n'$, by summing over taxis and counting the number of entries that are FREE (4).

Extending this analysis further, we quantify the availability SLACK(M, T, p'), or weighted number of free taxis at virtual taxi stand $p = p'$ over a *time window* T, by summing the number of free taxis at p' over time periods $t \in T$ (5).

Suppose we have 24 hours of data and choose $N = 1440$ (so each time interval n is exactly one minute long). If we define a "taxi minute" to be the number of minutes that a taxi spends in a particular state and applied (6) to each non null entry in the $(M \times N \times P)$ matrix, we get the fleet wide total taxi minutes that are spent FREE. Dividing this number by the total number of taxis gives us the average time that each taxi spent driving empty. This calculation is analogous to the man hour, which is the amount of work performed by the average worker in one hour. Returning to the example of the taxi stand with high unmet demand, by carefully defining the boundaries of p' to completely cover the taxi stand and its queueing area, we would find that a high unmet demand taxi stand will generate very few FREE taxi minutes because whenever a FREE taxi arrives at the taxi stand, it is immediately "consumed" and turned into a POB taxi. Conversely, a low unmet demand taxi stand with many taxis waiting in queue will generate a comparatively large number of FREE taxi minutes.

C. Taxi Demand

The most straightforward way to quantify taxi demand at a particular taxi stand is to use B, the number of observed passenger boardings. A change of state from FREE to POB indicates that a taxi picked up a passenger. We can use this fact together with the notation introduced in Section IV-B to formally define $B(M, n', p')$ the number of observed taxi boardings at a given taxi stand p' for a particular time period n' (7).

$$B(M, n', p') = \sum_{m \in M} \text{PICKUP}(m, n', p') \tag{7}$$

$$B(M, T, p') = \sum_{t \in T} \sum_{m \in M} \text{PICKUP}(m, t, p') \tag{8}$$

Where

$$\text{PICKUP}(m, n, p) = \begin{cases} 1 & \text{if the } (m, n, p)^{th} \text{ entry is POB and} \\ & \text{the } (m, n-1, p)^{th} \text{ entry is FREE} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Similarly, $B(M, T, p')$, the number of observed taxi boardings at a given taxi stand p' for time window T can be found by summing the number of pickups over time periods $t \in T$ (8).

D. Unmet Demand

Using the results from (5) and (8), we define a heuristic for unmet demand, $\rho(M, T, p')$, the unmet demand intensity at virtual taxi stand p' over time window T (10).

$$\rho(M, T, p') = \frac{B(M, T, p')}{\text{SLACK}(M, T, p')} \tag{10}$$

The unmet demand intensity ρ is simply the ratio of observed taxi demand to slack. This captures the imbalance in supply and demand and as we shall prove in the next section, ρ is a constant multiple of the ratio of demand to residual taxi queue length, $\frac{D}{S-B}$, of our virtual taxi stand.

V. Analysis

To motivate our analysis, let us consider the dynamics of a single taxi stand over τ discrete time periods i.e. $t = \{1, 2, ..., \tau\}$. Taxis and people arrive deterministically at the taxi stand with unknown rates λ_T and λ_P and are instantly serviced (when a passenger boards a taxi) with observed rate μ. For simplicity, we assume that there is an excess supply of taxis so $\mu \leq \lambda_P \leq \lambda_T$ and that at time $t = 1$, the taxi stand is completely empty. Our goal is to use this information to find $\frac{D}{S-B}$, the ratio of taxi demand D to residual taxi queue length $S - B$.

Lemma V.1. *The total passenger demand for taxis D over τ time periods is $\tau\lambda_P = \tau\mu$.*

Proof. The proof follows from the definition of arrival rate λ. By assuming deterministic arrivals, exactly λ_P people arrive at the taxi stand in each time period. Then after τ time periods, exactly $\tau\lambda_P$ people would have arrived. Because taxis in excess, each them would have been matched with a taxi with rate μ i.e. observed demand B is equal to actual demand D. The total demand for taxis D is then $\tau\mu$, which is equal to total passenger arrivals $\tau\lambda_P$. □

Lemma V.2. *The total amount of SLACK (free taxi minutes) generated over τ time periods is $\frac{1}{2}\tau^2(\lambda_T - \lambda_P)$.*

Proof. The number of free taxi minutes generated by a taxi in queue is equivalent to its waiting time. Taxis arrive with rate λ_T and are serviced at rate $\mu = \lambda_P$. Since $\lambda_P \leq \lambda_T$, the queue of taxis accumulates at rate $\lambda_T - \lambda_P$. After τ time periods, the expected queue length of taxis is $\tau(\lambda_T - \lambda_P)$ and the total delay experienced in queue is the area of a triangle of base τ and height $\tau(\lambda_T - \lambda_P)$, which is equal to $\frac{1}{2}\tau^2(\lambda_T - \lambda_P)$. □

Theorem V.1. *ρ is a constant factor approximation of the ratio of taxi demand D to residual taxi queue length $S - B$.*

Proof.

$$\begin{aligned}
\rho &= \frac{B}{\text{SLACK}} && \text{(from 10)} \\
&= \frac{\tau\mu}{\frac{1}{2}\tau^2(\lambda_T - \lambda_P)} && \text{(from V.2)} \\
&= \frac{\tau\lambda_P}{\frac{1}{2}\tau^2(\lambda_T - \lambda_P)} && \text{(from V.1)} \\
&= \frac{2\tau\lambda_P}{\tau^2(\lambda_T - \lambda_P)} \\
&= \frac{2}{\tau} \cdot \frac{\tau\lambda_P}{\tau\lambda_T - \tau\lambda_P} \\
&= \frac{2}{\tau} \cdot \frac{D}{S - B}
\end{aligned}$$

□

But why not find λ_T directly from data to calculate ρ? It turns out that while this is simple in the case of a single virtual taxi stand where you can check if a taxi has just entered the taxi stand's boundary, it is difficult to do for multiple taxi stands for many taxis, particularly when you have the added constraint in Section VII that ρ must be calculated in realtime.

VI. Discussion

It is important to note that our heuristic for unmet demand is arbitrary. Unlike a metric such as average wait time or queue length, unmet demand intensity ρ does not represent a real world quantity. A higher ρ simply means that there is more demand pressure and it is harder for a passenger to find an available taxi, ceteris paribus. This result can be viewed in two ways. By keeping T constant and comparing ρ across different taxi stands $\{p_1, p_2, ..., p_P\}$, we can visualize how unmet demand is spatially distributed at a particular point in time. Alternatively, by varying T, we can see how unmet demand at a specific taxi stand p' changes over time.

To illustrate this we conducted a simple experiment where we defined each virtual taxi stand as a square 2km by 2km grid overlaid on the the urban environment of Singapore. We discretized time by setting $N = 1440$, intervals $t = \{1, 2, ..., 1440\}$ to be one minute long and time windows $T = \{[1, 2, ..., 15], [16, 17, ..., 30], ..., [1426, 1427, ..., 1440]\}$ to be 15 minutes long. Using data from $M = 16000$ taxis, we calculated values for slack and boarding for $P = 1144$ virtual taxi stands.

A. Unmet Demand Comparison over Space

We normalized the ρ at each virtual taxi stand to obtain a heat map of unmet demand for each 15 minute time window (Figure 8). The heuristic correctly picks out unmet demand in high density, high income residential areas in the mornings and the central business district in the evenings.

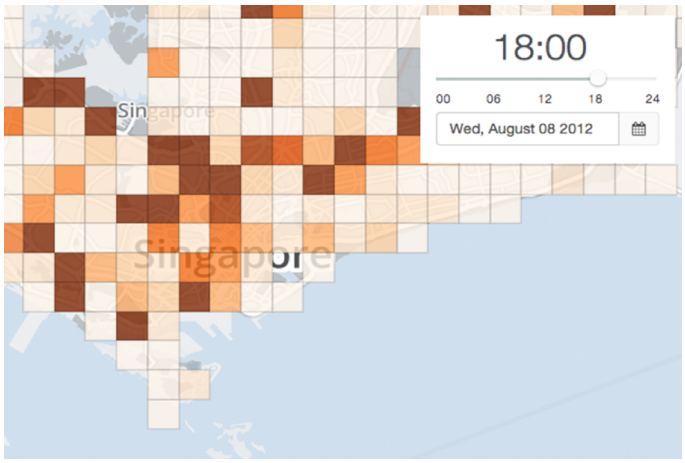

Fig. 8. A map of unmet taxi demand trained on data from 18:00 - 18:15 hrs on a Wednesday. As expected, the map shows hotspots of unmet demand forming in the central business district and other high density office districts as people leave work for home.

B. Unmet Demand Comparison over Time

Even though the grids defining our virtual taxi stands were arbitrarily chosen, several correspond to well defined neighborhoods in Singapore and are thus able to capture their characteristic taxi activity. Figure 9 shows how unmet demand varies at Orchard Road, a high end shopping and residential area on a Friday. In the mornings, unmet demand spikes briefly as high income residents (predominately expatriates) take taxis to work. Another peak can be seen during the evening rush hour as shoppers and office workers take taxis to return home.

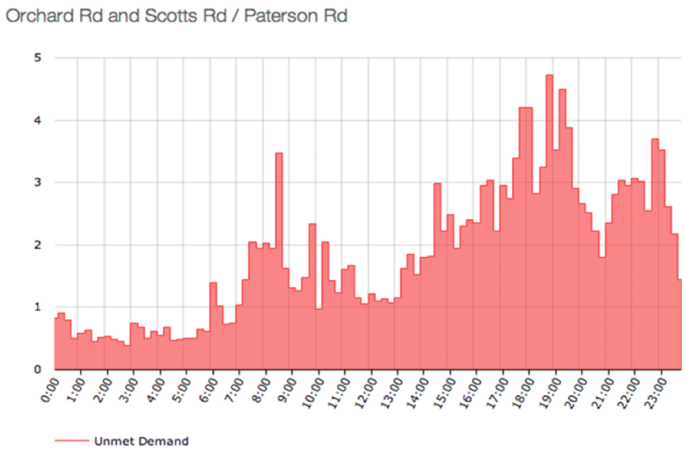

Fig. 9. Unmet demand at Orchard Road on a typical Friday. Two peaks are observed - one in the morning at about 8 am and another in the evening at 7 pm. This behavior is consistent with general taxi patterns in that neighborhood.

VII. Smartphone App and Recommendation Engine

In this section, we explain how the unmet demand heuristic described in Section IV-D can be used to build a recommendation engine and smartphone application that uses a real time stream of taxi probe data to direct participating drivers to areas

of high unmet demand. The recommendation engine connects to a live data stream (a time ordered sequence of data in the format specified in Section IV-A) and uses this data to calculate at instant t, the unmet demand at each virtual taxi stand during the interval $[t-15, t-1]$. Our smartphone application then ranks and displays these results to drivers in a convincing and accessible way so that they can easily find nearby hotspots of unmet demand.

Fig. 10. Unmet demand app showing (left) driving directions to the nearest unmet demand hotspot and (right) unmet demand at individual taxi stands

A. Recommendation Engine

To store and manipulate data received by the live stream, we implement the three dimensional $M \times N \times P$ data structure described in Section IV-B in a relational database. Each row in the database corresponds to a taxi minute record with column attributes *taxi_id* (M), *time_period* (N), *taxi_stand_id* (P) and an extra column for *state*. Each *taxi_stand_id* refers to a virtual taxi stand created by a grid mesh superimposed on a map of the city at three zoom levels - high (0.5 km x 0.5 km), medium (1.0 km x 1.0 km) and low (2.0 km x 2.0 km).

At one minute intervals, a background task retrieves records received in the last 15 minutes and applies (10) on each taxi stand to calculate its unmet demand intensity on a sliding 15 minute time window. Concretely, this means that at time t, the recommendation engine calculates $\rho(\{t-15, t-14, ..., t-1\}, p) \quad \forall p = \{1, 2, ..., 1444\}$. This result is stored in a separate high fidelity database as a 3-tuple (*taxi_stand_id, time_period, unmet_demand*) where it is exposed via API to our smartphone application.

B. Smartphone Application

Our smartphone application is designed to show taxi drivers the location of nearby unmet demand hotspots in a clear and visually intuitive way. At regular one minute intervals, the application sends a request to the recommendation engine, specifying the current time and location of the user. The engine uses this information to return a list of virtual taxi stands and their unmet demand intensity from the previous 15 minutes.

The top half of the app (Figure 10 left) displays a grid overlay of each virtual taxi stand, color coded by *normalized* unmet demand intensity (darker grids have a higher intensity). The bottom half of the app shows a listing of the top 3 taxi stands within a 15, 30 and 45 minute driving radius together with their unmet demand intensity normalized on a 1 - 10 scale. Instead of simply displaying the *taxi_stand_id*, we name these virtual taxi stands according to the general area that best describes the taxi stand's location. This allows users to easily compare *relative* unmet demand between different neighborhoods, so that they can quickly decide where to go. Clicking on a taxi stand's name brings up driving directions from the user's current location to that taxi stand.

At the lowest resolution, the heat map identifies areas of high unmet demand. As you zoom in, the map automatically subdivides into smaller grids and at the highest resolution, it correctly picks out individual taxi stands (Figure 10 right).

An interesting feature of our system is that it has a built in real time control policy that limits too many taxis from converging on a single taxi stand. As more taxis enter a high unmet demand area, they collectively generate large amounts of FREE taxi minutes. As demand clears, less observed boardings are generated. Both of these factors put downward pressure on the unmet demand intensity so the taxi stand no longer appears as a hot spot, and new taxis no longer have an incentive to go there.

VIII. Conclusions and Future Work

In this paper we examine the problem of inferring unmet passenger demand for taxis. We formalized the notion of unmet demand in our problem context and present a novel heuristic algorithm to estimate it without any information on passenger queue lengths or arrival rates. Along the way, we introduce the concept of the FREE taxi minute and show how it can be used to to quantify slack, the general availability of taxis.

Our heuristic, the unmet demand intensity ρ (10), is the ratio of observed taxi demand to slack. We show that ρ is a constant factor approximation of the ratio of taxi demand D to residual taxi queue length $S - B$ and test it using real world taxi data from Singapore with promising early results. The algorithm correctly captures unmet demand at different locations and times of day.

Finally, we develop a recommendation engine and smartphone application that balances fleet wide supply and demand by directing taxi drivers to nearby hotspots of unmet demand. We describe the architecture and technical details of this system and show how it uses a stream of real time taxi probe data to calculate citywide unmet demand in an online way. This information is displayed on a smartphone app that allows taxi drivers to easily identify nearby unmet demand hotspots.

This work is a first step towards a real time control system to match supply and demand in a city. Unlike competing

approaches, we do this by considering unmet demand, rather than just observed demand. In future work, we plan to test the performance of our system in a small scale fleet test in Singapore. We believe that the deployment of such a system in a taxi fleet will increase the productivity and utilization of the fleet by improving the distribution of the vacant vehicles throughout a city.

IX. Acknowledgements

Support for this research has been provided by the Singapore-MIT Alliance for Research and Technology (Future Urban Mobility Project) and SMART Innovation Center grants ING13057-ICT and EG11011. We would like to thank Li Hongyi for his valuable contributions towards the user interface and user experience design of the smartphone applications.

References

[1] T. C. Post, "Taxi drivers earn average of NT$21,305 per month," http://www.chinapost.com.tw/taiwan/national/national-news/2008/08/16/170293/Taxi-drivers.htm, 2008, [Online; accessed 13-April-2015].

[2] A. Anwar, M. Volkov, and D. Rus, "Changinow: A mobile application for efficient taxi allocation at airports," in *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*. IEEE, 2013, pp. 694–701.

[3] D. Santani, R. K. Balan, and C. J. Woodard, "Spatio-temporal efficiency in a taxi dispatch system," in *6th International Conference on Mobile Systems, Applications, and Services, MobiSys*, 2008.

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

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

[6] C. Council, "Unmet taxi demand survey 2014," https://www.cornwall.gov.uk/media/3625289/Carrick-Taxi-Zone-Survey.pdf, 2014, [Online; accessed 13-April-2015].

[7] D. Council, "Unmet taxi demand survey 2013," http://www.dundeecity.gov.uk/sites/default/files/FINAL\%20REPORT.pdf, 2013, [Online; accessed 13-April-2015].

[8] B. P. Loo, B. S. Leung, S. Wong, and H. Yang, "Taxi license premiums in hong kong: Can their fluctuations be explained by taxi as a mode of public transport?" *International Journal of Sustainable Transportation*, vol. 1, no. 4, pp. 249–266, 2007.

[9] S. T. Singapore, "A taxi in hand is worth two on the road," http://transport.asiaone.com/news/general/story/taxi-hand-worth-two-road, 2013, [Online; accessed 13-April-2015].

[10] S. L. T. Authority, "Average passenger waiting time (Orchard cluster)," http://www.lta.gov.sg/content/dam/ltaweb/corp/PublicTransport/files/PWT_Orchard.pdf, 2015, [Online; accessed 13-April-2015].

[11] L. T. Authority, "Taxi Information System," http://www.lta.gov.sg/content/dam/ltaweb/corp/PublicTransport/files/TIS.pdf, 2014, [Online; accessed 13-April-2015].

[12] R. G. Mishalani, Y. Ji, and M. R. McCord, "Effect of onboard survey sample size on estimation of transit bus route passenger origin-destination flow matrix using automatic passenger counter data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2246, no. 1, pp. 64–73, 2011.

[13] G. Berbeglia, J.-F. Cordeau, and G. Laporte, "Dynamic pickup and delivery problems," *European journal of operational research*, vol. 202, no. 1, pp. 8–15, 2010.

[14] J. Yang, P. Jaillet, and H. Mahmassani, "Real-time multivehicle truckload pickup and delivery problems," *Transportation Science*, vol. 38, no. 2, pp. 135–148, 2004.

[15] M. W. Savelsbergh, "Local search in routing problems with time windows," *Annals of Operations research*, vol. 4, no. 1, pp. 285–305, 1985.

[16] Y. Lin, W. Li, F. Qiu, and H. Xu, "Research on optimization of vehicle routing problem for ride-sharing taxi," *Procedia-Social and Behavioral Sciences*, vol. 43, pp. 494–502, 2012.

[17] J.-F. Cordeau, "A branch-and-cut algorithm for the dial-a-ride problem," *Operations Research*, vol. 54, no. 3, pp. 573–586, 2006.

[18] N. Megiddo, *On the complexity of linear programming*. IBM Thomas J. Watson Research Division, 1986.

[19] M. Pavone, K. Treleaven, and E. Frazzoli, "Fundamental performance limits and efficient polices for transportation-on-demand systems," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 5622–5629.

[20] M. Pavone, S. L. Smith, and E. F. D. Rus, "Load balancing for mobility-on-demand systems," 2011.

[21] K. Spieser, K. Treleaven, R. Zhang, E. Frazzoli, D. Morton, and M. Pavone, "Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in singapore," in *Road Vehicle Automation*. Springer, 2014, pp. 229–245.

[22] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[23] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini, "Statistical laws in urban mobility from microscopic gps data in the area of florence," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010, no. 05, p. P05001, 2010.

[24] L. Hufnagel, D. Brockmann, and T. Geisel, "Forecast and control of epidemics in a globalized world," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15 124–15 129, 2004.

[25] V. Belik, T. Geisel, and D. Brockmann, "Natural human mobility patterns and spatial spread of infectious diseases," *Physical Review X*, vol. 1, no. 1, p. 011001, 2011.

[26] L. Sun, K. W. Axhausen, D.-H. Lee, and M. Cebrian, "Efficient detection of contagious outbreaks in massive metropolitan encounter networks," *Scientific reports*, vol. 4, 2014.

[27] M. Batty, *The new science of cities*. Mit Press, 2013.

[28] D. Austin and P. C. Zegras, "Taxicabs as public transportation in boston, massachusetts," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2277, no. 1, pp. 65–74, 2012.

[29] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, and C. Ratti, "Quantifying the benefits of vehicle pooling with shareability networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 37, pp. 13 290–13 294, 2014.

[30] J. Lee, I. Shin, and G.-L. Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, vol. 1. IEEE, 2008, pp. 199–204.

[31] R. Bai, J. Li, J. A. Atkin, and G. Kendall, "A novel approach to independent taxi scheduling problem based on stable matching," *Journal of the Operational Research Society*, vol. 65, no. 10, pp. 1501–1510, 2013.

[32] M. Volkov *et al.*, "Deployment algorithms for multi-agent exploration and patrolling," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

[33] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 3, pp. 1393–1402, 2013.

[34] e27, "Gomyway launches to allow singaporeans to share taxi rides," https://www.dropbox.com/s/9xhpmbqkfuylqr0/20120823_gomyway_avg_taxi_ridership.pdf?dl=0, 2012, [Online; accessed 25-June-2015].